Starting a Sentence in Dutch

A corpus study of subject- and object-fronting

Nederlandse Organisatie voor Wetenschappelijk Onderzoek

The work presented here was carried out as part of the project *Conflicts in Interpretation*, in the framework of the Netherlands Organization for Scientific Research NWO Cognition Programme, grant number 051-02-071, principal investigators Petra Hendriks, Helen de Hoop and Henriëtte de Swart.

rijksuniversiteit groningen

The work presented here was carried out under the auspices of the School of Behavioural and Cognitive Neuroscience and the Center for Language and Cognition Groningen of the Faculty of Arts of the University of Groningen. Additional financial support has come from the Stichting Nicolaas Muleriusfonds.

Groningen Dissertations in Linguistics 66
ISSN 0928-0030

©2008, G.J. Bouma
Document prepared with LaTeX 2ε and typeset by pdfTeX
Printed by Optima Grafische Communicatie

RIJKSUNIVERSITEIT GRONINGEN

Starting a Sentence in Dutch

A corpus study of subject- and object-fronting

Proefschrift

ter verkrijging van het doctoraat in de
Letteren
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
donderdag 21 februari 2008
om 16.15 uur

door

Gerlof Johannes Bouma
geboren op 20 juni 1979
te Bedum

Promotores:
Prof.dr. P. Hendriks
Prof.dr. J. Hoeksema
Prof.dr. H.E. de Swart

Beoordelingscommissie:
Dr. D.I. Beaver
Prof.dr. E. Engdahl
Prof.dr.ir. J. Nerbonne

Acknowledgments

I am first and foremost indebted to my direct supervisor, or should I say: *Doktormutter*, Petra Hendriks. Petra, your continuing patience – even when progress was perhaps not as it should – and the occasional well-needed impatience, your advice on writing and matters linguistic, and the overall freedom you have given me to follow my own path have been essential in the last five years. Moreover, I am extremely pleased that, last minute, we were able to strike you from the list of *co-promotores*, and add you as a *promotor*. As an aside, this move has put you at the top of the list of names on the facing page – a position you deserve not only because of alphabetical order. I am very grateful as well to my *promotores*-from-the-start, Jack Hoeksema and Henriëtte de Swart, for contributing with much more than just their *ius promovendi*. I greatly appreciate the fact that I could always come to you for discussion, advice, and comments on papers; and, more recently, the elaborate feedback on dissertation drafts. Henriëtte, thank you for sending me back to my writing desk when you did. In the line-up of supervisors, I should also mention my almost-namesake Gosse Bouma, who has in many ways acted as a supervisor, and who has, amongst other things, given valuable comments on one of the final drafts of this thesis.

I thank my fellow members of the NWO Cognition project *Conflicts in Interpretation* Petra Hendriks, Helen de Hoop, Irene Krämer, Henriëtte de Swart and Joost Zwarts. I may not have always enjoyed the *vergaderingen*, but I certainly did enjoy the collaborations that were part of the project, and the many drinks and dinners in Groningen, Nijmegen and Utrecht.

At the Faculty of Arts at the University of Groningen, there are simply too many people to thank, and I shall only be able to mention a few of them by name. I would like to thank my office mates over the years, (if I recall correctly) Robbert, Tamás, Francisco, Holger, Jennifer, Tim, Jens, and Jelena, for being good company, and for all the chats, coffee, and so on. I thank John Nerbonne and the members of the Department of Alfa-informatica for adopting me into their great group and for providing a pleasant and stimulating working environment. I thank all my CLCG colleagues in the different departments for allowing me to learn about all the different aspects of linguistics. I would also like to thank the

secretaries on the 4th floor and Wyke and Anna for putting up with all of us, even during the most stressful of times. Special thanks go to Leonoor van der Beek and Holger Hopp. I took great pleasure in our past collaborations, and even greater pleasure in the time spent together not working.

I thank the Department of Swedish at Gothenburg University, in particular Elisabet Engdahl, Maia Andreasson, and Benjamin Lyngfelt, for welcoming me as a guest in November 2004. I thank everybody at the Department of Linguistics at Stanford University for an inspiring and lovely time in the spring of 2005. I would like to especially thank David Beaver for making it possible for me to come to Stanford in the first place and for introducing me to the Existentials project. It is safe to say that the topic and approach of this dissertation would not have been the same without this visit.

In 2007, I finished up the final drafts of this dissertation whilst living in Oslo, where I appreciated the contacts with the members of *Tekstlaboratoriet*, the ILN and the ILOS at the University of Oslo. I am thankful to the ILOS for providing me with office space over the summer. In November, I moved to Potsdam, where my current colleagues at the Department of Linguistics at the University of Potsdam have made me feel certain that I will enjoy working amongst and with them in the years to come.

The acknowledgments chapter in any dissertation is probably the chapter family, friends, and colleagues of the author will want to read first. In order not to spoil the surprise for any of these inquiring minds, I did not ask anyone else to proofread these pages before printing... This need not worry the gentle reader, however, since the same cannot be said about the rest of the book. I am grateful to Holger, Jaap, Jennifer, Laura, and Neal for proofreading the drafts for grammar, style, and spelling: You have greatly increased the quality of the book. Any remaining errors are of course my own and have been introduced after proofreading.

I am very happy and honoured to have my dear friend Frank and my good colleague and primary Southern Dutch informant Tim as my *paranimf*s. Thank you both for acting as my witnesses when I defend my dissertation.

Finally, I would like to thank my parents, brother and sister, and my fiancée. You may think you did not contribute much to this dissertation, and I may not be able to clearly tell you in which ways you did, but I am quite certain it would not have been written without you. *Lieve Oane, mem, Jaap, Jonne, en de steeds maar groeiende aanhang, liefste Kajsa, zonder jullie niets.*

gjb – Potsdam, 13 January 2008

Dit boek draag ik op aan mijn ouders.

Contents

1 Introduction · 13

2 Preverbal Behaviour · 19
2.1 Topological fields · 20
2.2 Vorfeld occupants · 23
2.2.1 Vorfeld subjects · 23
2.2.2 Topicalization · 25
2.2.3 Preposition stranding · 28
2.2.4 Non-arguments in the Vorfeld · 30
2.3 Subjecthood and Vorfeld pronouns · 31
2.4 Violations of V2 · 35
2.4.1 No elements in the Vorfeld · 36
2.4.2 Left dislocation and hanging topics · 37
2.4.3 Multiple elements in the Vorfeld · 39
2.4.4 Two left brackets · 41
2.4.5 Summary · 42
2.5 Topicalization and information structure · 42
2.5.1 Focus topicalization · 43
2.5.2 Topic topicalization · 44
2.6 Word order trends · 49
2.6.1 Canonical argument order · 50
2.6.2 Definiteness · 53
2.6.3 Grammatical complexity · 57
2.7 Conclusion · 60

3 Methods, Techniques & Material · 63
3.1 About the *Corpus Gesproken Nederlands* · 63
3.2 Syntactic annotation in the CGN · 65

3.2.1 Dependencies and phrases · 66
3.2.2 Multiple Dependencies · 70
3.3 Finding the Vorfeld in CGN · 73
3.4 Implementation · 76
3.5 Statistical methods · 80
3.6 Summary · 85

4 A Corpus Study of the Vorfeld · 87
4.1 Some first statistics · 88
4.1.1 Data selection · 90
4.1.2 Subjects and objects in a sentence · 93
4.2 Arguments · 95
4.2.1 Corpus results · 95
4.2.2 Summary · 102
4.3 Definiteness · 103
4.3.1 Operationalizing definiteness · 103
4.3.2 Corpus results · 106
4.3.3 Pronouns in the Vorfeld · 118
4.3.4 Summary · 120
4.4 Grammatical complexity · 122
4.4.1 Corpus results · 122
4.4.2 Locating the complexity effect · 127
4.4.3 Two tentative proposals for theoretical consequences · 130
4.4.4 Two types of Nachfeld occupation · 134
4.4.5 Summary · 136
4.5 Grammatical function, definiteness and complexity · 136
4.5.1 Model definition · 138
4.5.2 Modelling results · 140
4.5.3 Summary · 142
4.6 The presence of negation, and other modifiers · 144
4.7 Conclusion · 150

5 Word Order Freezing · 153
5.1 Introduction · 154
5.1.1 Word order freezing in Dutch · 155
5.1.2 Approaches to word order freezing · 157
5.1.3 Relation with corpus study · 161
5.2 A brief introduction to Optimality Theory · 163
5.3 A bidirectional account of word order freezing · 170
5.3.1 Word order freezing in Hindi · 170

5.3.2 Analysis · 172
5.4 Against a bidirectional account? · 177
5.4.1 Problems with a bidirectional account of freezing · 178
5.4.2 Extending unidirectional OT to model freezing · 182
5.4.3 Problems for unidirectional production models · 191
5.4.4 Summary · 195
5.5 A bidirectional account of freezing, revisited · 197
5.5.1 Ambiguity and optionality in bidirectional OT · 197
5.5.2 Wh-questions · 201
5.5.3 Information structure in frozen sentences · 207
5.5.4 Focus fronting · 216
5.5.5 Animacy, world knowledge and gender agreement · 220
5.5.6 Summary · 223
5.6 Combining variation, production, and comprehension · 224
5.6.1 Variation in strong bidirectional OT · 225
5.6.2 Symmetric bidirectionality · 230
5.6.3 Asymmetric bidirectionality · 233
5.7 Conclusion · 237

6 A Corpus Investigation into Word Order Freezing · 241
6.1 Preliminaries · 242
6.2 Relative definiteness · 245
6.2.1 Pronominal subjects and object placement · 249
6.2.2 Relative definiteness and object placement · 251
6.3 Relative animacy and object placement · 256
6.4 Negative evidence for word order freezing? · 263
6.5 Conclusions · 265

7 Conclusions · 267
7.1 Summary of main findings · 268
7.2 Directions for future work · 272

A List of Abbreviations · 275
A.1 Syntactic categories · 275
A.2 Dependencies · 276

B Examples of Vorfeld Occupants in CGN · 279

Bibliography · 285

Chapter 1

Introduction

A commercial that was broadcast on Dutch television in the past featured a song with the following lines:

(1) a. Koning, keizer, admiraal,
King emperor admiral
b. Popla kennen ze allemaal!
Popla know they all
'It doesn't matter whether they're king, emperor or admiral, they're all familiar with Popla.'

The direct object of (1b), *Popla*, a brand of toilet paper, directly precedes the finite verb. This position has traditionally been referred to as the *Vorfeld*.

Dutch is a verb-second, verb-final language: In a declarative main clause the finite verb has to occur in second position, and any non-finite verbs are placed towards the end of the sentence. Otherwise, Dutch allows for a moderate amount of word order variation. One of the liberties a speaker of Dutch has is the choice of a Vorfeld occupant. For instance, the writers of the aforementioned commercial could have used (1b′) to express the proposition that everybody is familiar with their brand of toilet paper.

(1) b′. Ze kennen Popla allemaal!
they know Popla all
'They're all familiar with Popla.'

In (1b′), it is the subject *ze* that sits in the Vorfeld. The central research question in this dissertation is what determines the choice for a Vorfeld occupant.

There are relatively few constraints on which constituent may occupy the Vorfeld. Sentences (1b) and (1b′) only demonstrate two of the many possibilities. However, I will

not be concerned with which Vorfeld occupants are *grammatically* possible. My primary interest in this dissertation is which syntactic, discourse semantic, and communicative factors influence which constituent a speaker puts in the Vorfeld. To get more insight into this issue, I will study Vorfeld occupation by subjects, direct objects and indirect objects in naturally occurring spoken Dutch. The range of investigated constructions is thus as illustrated in (2):

(2) a. Ik heb jou dat verteld.
I have you that told
b. Jou heb ik dat verteld.
you I that
c. Dat heb ik jou verteld.
that I you
'I told you that.'

In (2a), the subject occupies the Vorfeld, in (2b), the indirect object does, and in (2c), the direct object. I will use a combination of theoretical modeling and corpus investigation to pin down some of the factors that make a speaker choose (2a), (2b), or (2c).

One type of influence on Vorfeld occupation that we may find comes from known word order tendencies. These tendencies include the tendency to realize subjects earlier on in the sentence than objects, pronouns earlier on than full NPs, and definite or given material earlier on than indefinite or new material. These tendencies have been extensively studied and are fairly well established for the Dutch, and also German, postverbal domain – that is, the domain to the right of the finite verb. The part that these tendencies play in selecting a Vorfeld occupant is not well studied, however. Word order variation with respect to the Vorfeld is less restricted than postverbal word order variation, and as a result harder to investigate. Let us take the tendency to realize subjects early in a sentence as an example. This can not be a categorical constraint on Vorfeld occupation since (2b) and (2c) are grammatical. In (2b) the indirect object is realized before the subject and in (2c) the direct object is realized before the subject. However, intuitively, the subject initial sentence (2a) is the least marked of the three. We might hypothesize that the tendency to realize subjects early on in the sentence affects Vorfeld occupation, and that the unmarked status of (2a) is a result of this. One of the hypotheses to be investigated in this dissertation is the hypothesis that tendencies such as those mentioned above are *global tendencies*. If they are global tendencies, they are not restricted to the postverbal domain, but influence Vorfeld occupation, too.

The hypothesized global word order tendencies need not be the only determinants in the choice of a Vorfeld occupant. Compare the alternative answers to the question in (3A). Capitals indicate main stress.

(3) A Wie heb je het hof gemaakt?
'Who did you court?'
B Ik heb Grace KELLY het hof gemaakt.
I have Grace Kelly courted
B′ Grace KELLY heb ik het hof gemaakt.
Grace Kelly I
'I courted Grace Kelly.'

Intuitively, the difference between (3B) and (3B′) is that *Grace Kelly* receives extra attention when it is put in the Vorfeld: Grace Kelly is presented as an extra newsworthy or unexpected object of courtship. Apparently, the Vorfeld, as a left-peripheral position, has the special function of hosting important material. We shall see that the effects of the nature of the Vorfeld as a position for important material – informal as this concept may be – can be observed separately from the global word order tendencies illustrated above.

The effects that global word order tendencies and the Vorfeld as a position for important material have on the choice of a Vorfeld occupant have in common that a constituent property is linked to Vorfeld occupation. In the case of the global word order tendencies, we connect being a subject, being a pronoun, or being definite to appearing in the Vorfeld. In the case of the identification as the Vorfeld as a place for important material, we connect being newsworthy, highlighted, or unexpected to Vorfeld occupation. The general topic of this dissertation is what determines Vorfeld occupation. In light of this general topic, the link between constituent properties and Vorfeld occupation raises the first subquestion to be investigated in this dissertation: How do properties of a constituent influence the chance that it is selected as Vorfeld occupant? A large scale corpus study of spoken Dutch is at the heart of answering the first subquestion.

Let us return to the hypothetical task of a speaker of Dutch selecting a Vorfeld occupant. After answering the first subquestion, we know more about what makes a constituent an attractive Vorfeld occupant for the speaker. However, this ignores the purpose of producing an utterance, which is that some intended meaning is transmitted to a hearer. It may be that selecting a certain Vorfeld occupant, even if it is attractive to the speaker, jeopardizes this goal. Consider the example in (4).

(4) Fitz zag Ella
Fitz saw Ella
'Fitz saw Ella.' *Or, but dispreferred:* 'Ella saw Fitz.'

The sentence in (4) can in principle be an SVO sentence or an OVS sentence, depending on whether the utterer of (4) selected the subject or the object as the Vorfeld occupant. However, presented like this, the SVO interpretation is strongest, and the OVS interpretation is suppressed. On its own, this is not surprising since there is nothing about *Fitz* or

Ella to tell us which is the subject and which the object. The global tendency to have subjects early in the sentence makes SVO the strongest reading. So, a hearer interpreting (4) will be inclined to interpret the sentence as meaning that Fitz did the seeing, and Ella was seen. This observation about hearer behaviour may have consequences for the freedom of word order variation of the speaker. If the speaker intends to communicate that Ella did the seeing and Fitz was seen, the word order in (4) is a poor choice: The intended meaning would be the one that is dispreferred by a hearer. The chances of successful communication would be better if the speaker puts the subject in the Vorfeld.

The second half of the dissertation approaches the question of what determines Vorfeld occupation from the communication perspective illustrated above. The subquestion to be answered in this part of the dissertation is: How does the chance of communicative success influence the choice of a Vorfeld occupant? Building on the results from the first half of the thesis, I will show in the second half of the thesis that we can formulate a theoretical model of the interaction between speaker and hearer preferences, and that the effect of hearer preferences on speaker choices can be observed in the corpus of spoken Dutch as a statistical tendency in the use of non-canonical word order.

The dissertation is structured as follows. **Chapter 2: Preverbal Behaviour** introduces the topological model of the clause, used in traditional German and Dutch grammars. Formal and functional properties of the left-periphery of the clause are discussed, especially in relation to the Vorfeld. On the basis of this discussion, the range of constructions that will be investigated in the rest of the dissertation is delimited. I will also formulate expectations for the corpus results on the basis of existing knowledge about word order variation in other languages and in other domains in the Dutch clause. The factors to be investigated in the corpus are grammatical function, definiteness and grammatical complexity.

In **Chapter 3: Methods, Techniques & Material**, I will lay out the tools that are used in the corpus investigations, and describe the spoken Dutch corpus *Corpus Gesproken Nederlands* (CGN), that is used throughout the dissertation. A definition of the Vorfeld will be given in terms of the CGN, so that Vorfeld material can automatically be retrieved. The chapter also provides an informal introduction in *logistic regression*, one of the statistical modeling techniques that will be used in the analysis of the corpus data.

Chapter 4: A Corpus Study of the Vorfeld is the first of two empirical chapters in this dissertation. For each of the factors grammatical function, definiteness, and grammatical complexity it is investigated whether they contribute in the choice for a Vorfeld occupant. We will see that not all factors are of the same nature, and that not all factors influence Vorfeld occupation directly. Some of the factors are directly tied to the Vorfeld, some influence word order on a global level, and some only indirectly affect Vorfeld occupation by targeting other positions in the sentence.

Chapter 5: Word Order Freezing is a theoretical chapter. In this chapter, I will extend a formalization of the interaction between speaker's and hearer's preferences using so-called *bidirectional Optimality Theory*. A wide range of data on word order freezing can be elegantly captured in the resulting model. I will also contrast the bidirectional approach with other Optimality-theoretic accounts of word order freezing, and show that bidirectionality is required for a satisfactory account.

The shorter chapter that follows, **Chapter 6: A Corpus Investigation into Word Order Freezing**, is the second empirical chapter of the dissertation. In this chapter, I investigate whether we can observe the predictions that are made by the bidirectional model developed in the previous chapter as quantitative trends in a corpus. We will see that this is the case. The corpus investigations in this chapter thus provide evidence for word order freezing in spoken Dutch discourse. They also provide support for a bidirectional model of word order, and further our understanding of Vorfeld occupation as a whole.

Finally, I will summarize the findings on the determinants of word order in **Chapter 7: Conclusions**. Some of the many directions for further research will be elaborated upon, and I will speculate on some of the theoretical consequences of the findings in this dissertation.

Chapter 2

Preverbal Behaviour

This dissertation deals with word order variation in the left-peripheral, directly preverbal domain. The position that is investigated is known as the *Vorfeld* ('prefield', German) – a term that comes from a traditional model of grammatical description. In this chapter, I will give an overview of word order variation in the preverbal domain, introduce necessary terminology, and delimit the range of investigated constructions. At the end of the chapter, I will present predictions about Vorfeld occupation that will be tested against a corpus in Chapter 4.

I will begin by introducing the descriptive model of the Dutch clause in Section 2.1. The four sections that follow will then give an overview of variation in Vorfeld occupation. In Section 2.2, I will show data in which the Vorfeld is occupied by constituents of different grammatical functions and of different syntactic categories. Section 2.3 elaborates on the special relation between the Vorfeld and pronouns. In general, the Vorfeld contains exactly one constituent, however, in Section 2.4, I will discuss data with more than one preverbal element. The information structural properties of the Vorfeld are discussed in Section 2.5.

After the formal and functional overview of material in the preverbal domain, Section 2.6 discusses existing results on word order variation of Dutch, German and English. On the hypothesis that the word order trends discussed are caused by global trends in word order, I will formulate expectations for Vorfeld variation that mirror these global trends. Other predictions regarding Vorfeld occupation will come from the functional properties of the Vorfeld discussed earlier in the chapter.

2.1 Topological fields

Dutch allows for a fair amount of word order variation throughout the sentence, provided a couple of core positions are taken by certain elements. Consider the twelve variations on a theme in (1). Each example is a declarative main clause expressing the same propositional content: Tom Boonen (subject) would have beaten Jens Voigt (object) if the two cyclists would have sprinted against each other.[1]

(1) a. Boonen zou Voigt in de sprint geklopt hebben.
Boonen would Voigt in the sprint beaten have
'Boonen would have beaten Voigt in the sprint.'
b. Boonen zou Voigt in de sprint hebben geklopt.
c. Boonen zou in de sprint Voigt geklopt hebben.
d. Boonen zou in de sprint Voigt hebben geklopt.
e. In de sprint zou Boonen Voigt geklopt hebben.
f. In de sprint zou Boonen Voigt hebben geklopt.
g. Boonen zou Voigt geklopt hebben in de sprint.
h. Boonen zou Voigt hebben geklopt in de sprint.
i. Voigt zou Boonen in de sprint geklopt hebben.
j. Voigt zou Boonen in de sprint hebben geklopt.
k. Voigt zou Boonen geklopt hebben in de sprint.
l. Voigt zou Boonen hebben geklopt in de sprint.

Across these twelve sentences, there is variation in the place and order of the arguments, the position of the PP, and the order of the non-finite verbs. The constant skeleton in all of these sentences is formed by the finite verb *zou* in second position, and the non-finite verbs *hebben* & *geklopt*, which are clustered towards the end of the sentence. Dutch can be considered to be a verb-second, verb-final language. The other constituents in the sentences in (1) occur in the three 'fields' that are to the far left of, in between, and to the far right of these verbal positions.

It is common to describe the clauses and clause types for German and Dutch in terms of *topological fields*, characterizations of which can already be found in 19th century German grammars (Höhle, ms). Figure 2.1 shows topological fields for the description of Dutch used in the reference grammar Haeseryn et al. (1997).

The template in Figure 2.1 gives the names of the topological fields, as well as some typical inhabitants. For main declarative clauses like (1) – ignoring the fields *lead* and *tail* for a moment – the *Vorfeld* spans from the left edge of the clause until the *left bracket*,

[1] This interpretation is available for (1i)–(1l), with the direct object *Voigt* in initial position. Native speakers may find they need to pronounce *Voigt* with stress, and deaccent *Boonen* to get the reading. Also, it might be worth pointing out that although Jens Voigt is an excellent cyclist, he is no match for specialist Tom Boonen when it comes to sprinting.

Figure 2.1: Traditional clause template and typical field occupants

lead	*Vorfeld*	left bracket	*Mittelfeld*	right bracket	*Nachfeld*	tail
XP*	(XP)	V_{fin} |C	XP*	V*	(PP|CP)*	XP*

and is occupied by exactly one constituent.[2] The *left bracket* contains the finite verb in main clauses. The *Mittelfeld* is found between the two brackets and may contain any amount of material. In the *right bracket*, or verbal cluster, the remaining verbs in the clause are found and possibly a small amount of non-verbal material. Dutch allows for some variation in the ordering of verbal material in the verb cluster. Finally, extraposed material, such as CP complements, or PP complements and adjuncts occupy the *Nachfeld*.

The fields *lead* and *tail* are not a common part of clause topology, and may be considered as not belonging to the clause proper. They are reserved for material that is more loosely associated with the clause such as hanging topics, afterthoughts and vocatives. An example sentence with material in the tail is (2). The Nachfeld-tail border is marked with a '|'.

(2) Ze had een mooie naam, | mijn gids.
she had a pretty name my guide
'My guide had a pretty name.'

The topological field template is not only used to describe main declarative clauses, but it is also used to describe subordinate clauses and interrogative clauses. Examples of different clause types in the template can be found in Table 2.1, p22. Clauses generally follow the V2 generalization. However, in the case of a polar interrogative (example e in Table 2.1), the Vorfeld is empty, resulting in a V1 clause. In a subordinate clause (f), all verbs are in the verb cluster, and the complementizer is assumed to be in the left bracket.

Topological field templates for clauses are found in descriptive grammars of Dutch such as (Haeseryn et al., 1997) and German (Kunkel-Razum and Münzberg, 2005). In many theoretical and empirical linguistic studies, topology is used in a purely descriptive manner, although some researchers have given topological fields theoretical status, too (Kathol, 2000).

As a descriptive model, topology is very useful, but it is not perfect. There are cases in which it is not easy to decide how a given sentence should fit into the template. The identification of topological fields relies on V2 to separate the Vorfeld from the rest, and (other) verbs clustering towards the end of the sentence to form a Mittelfeld and a Nachfeld. If there is, say, more than one constituent before the finite verb, it is not clear how to assign these to topological fields: The constituents could be forced into

[2]But see Section 2.4 for exceptions to this generalization, that is, declarative main clauses with more or less than one constituent in the Vorfeld.

Table 2.1: Examples of different constructions in the clause template

Vorfeld	left bracket	*Mittelfeld*	right bracket	*Nachfeld*
a. *subject initial, declarative main clause (V2):*				
Boonen	zou	Voigt in de sprint	geklopt hebben.	
Boonen	would	Voigt in the sprint	beaten have	
b. *preposed PP, declarative main clause (V2):*				
In de sprint	zou	Boonen Voigt	geklopt hebben.	
in the sprint	would	Boonen Voigt	beaten have	
c. *object initial, declarative main clause (V2):*				
Voigt	zou	Boonen	geklopt hebben	in de sprint.
Voigt	would	Boonen	beaten have	in the sprint
d. *constituent question, main clause (V2):*				
Wie	zou	Boonen	geklopt hebben	in de sprint?
Who	would	Boonen	beaten have	in the sprint
e. *polar interrogative, main clause (V1):*				
	Zou	Boonen Voigt in de sprint	geklopt hebben?	
	would	Boonen Voigt in the sprint	beaten have	
f. *subordinate clause (V-final):*				
	dat	Boonen Voigt	geklopt zou hebben	in de sprint.
	that	Boonen Voigt	beaten would have	in the sprint

the Vorfeld (breaking V2), they could be analyzed as one constituent (respecting V2), or one constituent could be assigned to the Vorfeld, and the others to the lead (possibly respecting V2). In this dissertation, I will confine myself to sentences with exactly one preverbal constituent, that is, to one constituent in the Vorfeld, though brief discussion and examples of 'multiple Vorfeld occupancy' can be found in Section 2.4.

Another case in which it is unclear how to apply the topological field template is when there is no material in the verbal cluster to separate the Mittelfeld from the Nachfeld. In (3a) the Mittelfeld and Nachfeld are separated by the main verb, and in (3b) a verb particle marks the right bracket, but in (3c) there is nothing – apart from the analogy with (3a) – that tells us where the Mittelfeld ends and the Nachfeld begins. In the examples, field borders are indicated by '|'.

(3) a. Ella | had | Fitz | ontmoet | bij Gerald thuis.
Ella had Fitz met at Gerald's

b. Ella | kwam | Fitz | tegen | bij Gerald thuis.
Ella met Fitz VPART at Gerald's

c. Ella | ontmoette | Fitz bij Gerald thuis.
Ella met Fitz at Gerald's
'Ella met Fitz at Gerald's.'

Since this dissertation is concerned with the Vorfeld, the distinction between Mittelfeld and Nachfeld is not always relevant. In certain situations, to avoid making false or unfounded claims about the position of the elements, I will therefore use the terms *preverbal* (left of the left bracket) and *postverbal* (right of the left bracket) to refer more generally to the position of material in the clause. In the investigation of the influence of grammatical complexity on Vorfeld occupation, presented in Section 4.4, it will turn out that the distinction between Mittelfeld and Nachfeld is relevant. In that section, I therefore define further fields or positions distinguished by whether we can recognize the Mittelfeld from the Nachfeld or not.

The topological approach to clause description can also be found in the literature on Scandinavian languages, for instance in the reference grammars Teleman, Hellberg, and Andersson (1999, Swedish), and Faarlund, Lie, and Vannebo (1997, Norwegian). These languages are V2 languages like Dutch and German, which means that the Scandinavian Vorfeld is very similar to the Vorfeld in Dutch and German. However, since the Scandinavian languages are not verb final, Mittelfeld and Nachfeld in descriptions of those languages do not correspond well to the Dutch and German fields.

Now that we have the topological field model for Dutch in place, I will concentrate on word order variation that targets the Vorfeld. I will use the topological field template throughout this dissertation as a descriptive device.

2.2 Vorfeld occupants

In this section, I will present data to give a feel for the kind of material, in terms of grammatical function and syntactic category, that can appear in the Vorfeld. I will concentrate on subjects and objects in the Vorfeld, because they are the focus of this dissertation. However, I will briefly mention predicate fronting and Vorfeld adjuncts, too.

2.2.1 Vorfeld subjects

In the course of this chapter, we shall see that the subject is the unmarked Vorfeld occupant. In (4), the subject in the Vorfeld is alternatively a pronoun, a proper name or a definite full NP.

(4) Hij / Landis / De voormalig gele trui-drager ligt bijna 7 minuten voor
he Landis the former yellow jersey bearer leads almost 7 minutes VPART
'He / Landis / The former bearer of the yellow jersey leads by almost 7 minutes.'

Quantificational and/or indefinite subjects can also appear in the Vorfeld, as shown in (5).

(5) Iedereen / Iemand / Niemand ligt / Twee renners liggen 7 minuten voor
Everybody Nobody Someone lies Two riders lie 7 minutes ahead
'Everybody / Nobody / Someone leads / Two riders lead by 7 minutes.'

Even though indefinite subjects are allowed in the Vorfeld, there is a strong tendency to use an existential construction (EC) when the subject is indefinite. In that case, the Vorfeld is typically occupied by an expletive subject *er* 'there', and the logical subject appears postverbally (6a). Dutch, unlike English and the mainland Scandinavian languages, allows for transitive ECs.

(6) a. *(Er) heeft iemand bijna 7 minuten voorsprong.
EXPL has someone almost 7 minutes lead
'Someone is leading by almost 7 minutes.'
b. Gek genoeg had (er) iemand bijna 7 minuten voorsprong.
funnily enough had EXPL someone almost seven minutes lead
'Funnily enough, someone was leading by almost 7 minutes.'
c. ...weil (*es) niemand gearbeitet hat.
because EXPL nobody worked has
'...because nobody was working.' (German, Meinunger, 2007, example 7a)

The EC in (6b) shows that the use of an expletive subject is not restricted to the Vorfeld in Dutch. This contrasts with German, where expletive *es* in an EC is restricted to the Vorfeld. On the other hand, Dutch does not allow for universally quantified or definite logical subjects in ECs (7a), which German readily allows (7b).

(7) a. *Er komt de dood. Dutch
EXPL comes the death
b. Es kommt der Tod.
EXPL comes the death
'Death is coming.' (German)

Clausal subjects are allowed to be in the Vorfeld (8a), although there is a tendency for them to appear extraposed in the Nachfeld (8b). In case of extraposition, an expletive or 'preliminary' subject *het* 'it' appears earlier in the sentence.

(8) a. Dat de Harmonie 's avonds sluit is misschien maar goed.
that the Harmonie at night closes is maybe PART good

b. Het is misschien maar goed dat de Harmonie 's avonds sluit.
EXPL is maybe PART good that the Harmonie closes at night
'Maybe it is a good thing that the Harmonie-building closes at night.'

Clausal subjects in the Mittelfeld are only marginally possible in Dutch (9a). The extraposed variant (9b) is strongly preferred.

(9) a. ?Waarom is dat de Harmonie sluit misschien maar goed?
why is that the Harmonie closes maybe PART good
b. Waarom is het misschien maar goed dat de Harmonie sluit?
why is it maybe PART good that the Harmonie closes
'Why would it be a good thing that the Harmonie-building closes?'

The expletive pronouns *er* (6a) and *het* (8b) are examples of so called reduced pronouns in the Vorfeld. Dutch has a paradigm of *reduced* (or: *weak*) pronouns that are more restricted in their distribution than *full* (or: *strong*) pronouns. One restriction on reduced pronouns is that a reduced pronoun object is not allowed to appear in the Vorfeld. I will discuss the relation between Vorfeld occupation and reduced pronouns in more detail in Section 2.3. For reference, the two paradigms are given in that section, in Table 2.2, p32. For now, however, it is worth pointing out that *er* and *het* demonstrate that reduced pronoun subjects are allowed in the Vorfeld. This is an indication of the default status of subjects as Vorfeld occupants.

2.2.2 Topicalization

Other arguments besides subjects can also appear in the Vorfeld. Direct objects, indirect objects and oblique/prepositional complements can all be fronted. I will refer to placing a non-subject in the Vorfeld as *topicalization* – the terms fronting and Vorfeld occupation can apply to any constituent. Topicalization should be understood in a purely formal sense. Although a topicalized constituent may be a discourse topic or a sentence topic, it is not clear that topicalization is restricted to topics, or that it makes constituents topics. The information structural properties of topicalization will be discussed in Section 2.5.

As a result of V2, topicalization forces the subject to appear in the postverbal domain. Typically the subject in a sentence with a topicalized constituent appears directly to right of the finite verb, although other configurations are possible. In (10), we can see examples of definite NPs and pronouns in the Vorfeld in several non-subject functions.

(10) a. *direct object:*
De koning van Frankrijk / Die / Hem kom ik vaak tegen op straat.
The king of France DEM him meet I regularly VPART on street
'I regularly meet the king of France / him in the street.'

b. *indirect object (experiencer in transitive):*
De koning van Frankrijk / Die / Hem bevalt het uitstekend in Oslo.
The king of France DEM him pleases it splendidly in Oslo
'The king of France / he likes it a lot in Oslo.'

c. *indirect object (recipient in ditransitive):*
De koning van Frankrijk / Die / Hem geef ik geen geld.
The king of France / DEM / him give I no money
'I do not give the king of France / him any money.'

d. *oblique complement:*
Tegen de koning van Frankrijk / hem zeg ik "U"
To the king of France him say I "U"
'I address the king of France / him with "U".'

Examples (10a-c) show that full personal pronouns (in this case *hem*) and demonstrative pronouns (*die*) can be topicalized.[3] The reduced variant of *hem* is *'m*. The reduced pronoun *'m* cannot be topicalized, as is shown in the counterparts of (10a-c) with reduced pronouns, given in (11).

(11) a. *'m Kom ik vaak tegen op straat.
b. *'m Bevalt 't uitstekend in Oslo.
c. *'m Geef ik geen geld.

Topicalization is not restricted to definite NPs. Example (12a), contains a topicalized indefinite indirect object demonstrates. Example (12b) shows an indefinite (bare nominal) direct object in the Vorfeld.

(12) a. Een paar jongens heb ik een klap verkocht.
A couple guys have I a blow sold
'I punched a couple of the guys.'

b. Raketsla heeft kapitein Picard alle dagen op het menu staan.
Garden rocket has captain Picard all days on the menu stand
'Captain Picard eats garden rocket every day.'

A brief detour to topicalized predicates and verbs

In this dissertation, I investigate subject and object fronting. However, I wish to briefly mention here that predicates (whether they are adjectival or nominal), can be topicalized. Examples are given in (13).

[3] *Demonstrative* refers to form alone. Demonstrative pronouns are frequently used as anaphoric pronouns in Dutch, and are not restricted to deixis.

(13) a. Het / Dat is mooi
it that is nice
'It / That is nice.'
b. Mooi is dat / ??het!
nice is that / it
'Nice...' (ironic)
c. Mooi is het niet
nice is it not
'It is not nice.' (understatement)

These examples are somewhat idiomatic, but not all constructions involving Vorfeld predicates are. The reason to bring up topicalized predicates is the role that negation plays in facilitating these types of topicalization. For instance, the negation in (13c) is needed to make the variant with reduced pronominal subject *het* 'it' acceptable.

Birner and Ward (1998) ascribe the (nearly) obligatory presence of the negation for English counterparts of (13c) to a rhetoric device they call *proposition denial*. An example of propositional denial can be found in (14).

(14) The international hordes now streaming in from the west and south have, in contrast, no-nonsense ideas about what they want: a chance to work hard and make money. Laid back they are not. (Birner and Ward's 71b, p68)

Birner and Ward also note that fronted predicate APs or NPs are often found in an explicit contrast that involves a negation, as in (15):

(15) Pretty they aren't. But a sweet golden grapefruit taste they have. (their 43b, p48)

Topicalizing non-finite verbs appears to be facilitated by negation, too. The example in (16a) even features an explicit contrast in polarity. Example (16b) shows that topicalizing the non-finite verb is marginal without a negation.

(16) a. Geregend heeft het niet, wel gedauwd.
Rained has it not, AFF dewed
'It hasn't rained, but it has dewed.' (Kruisinga, 1938, 4d, p66)
b. ??Geregend heeft het.

I will not be able offer an explanation of why negation has a positive effect on acceptability of topicalization. However, in Section 4.6 I do investigate whether the facilitating effect of negation is found with object topicalization, too. We will see that direct object topicalization is more frequent in the presence of certain sentence adverbs, than when these adverbs are not present. As far as I am aware, this is a hitherto unknown fact about direct object fronting.

2.2.3 Preposition stranding

Standard Dutch allows for preposition stranding provided the object of the preposition is one of the R-pronouns *daar*, *waar*, *hier*, *d'r* en *er* ('there', 'where', 'here' and two reduced forms of 'there'), or one of the locatives *overal* ('everywhere') and *(n)ergens* ('no-/somewhere'). When they are the object of a preposition, the R-pronouns, formally identical to locative adverbs, refer as regular pronouns do. Many speakers will even accept them with a human referent. The locatives *overal* and *(n)ergens* have denotations like quantified NPs, respectively *alles* 'everything' and *(n)iets* 'no-/something', when they are the object of a preposition.

In contrast to their regular counterparts, R-pronouns obligatorily precede their preposition.[4] It is quite common for other material to intervene between the R-object and its preposition. In other words, the object may be extracted from the PP, and the preposition is stranded in its canonical position. Preposition stranding occurs with Mittelfeld scrambling as well as with topicalization to the Vorfeld (Bech, 1952; Van Riemsdijk, 1978; Haeseryn et al., 1997).

In (17), we see a regular personal pronoun as the object of a PP. The pronoun can appear in situ (17a), but not alone in the Vorfeld with the preposition stranded (17b). Fronting the complete PP is fine, as in (17c). The preposition and its object are in boldface.[5]

(17) a. Ik kan goed **met hem** praten
I can well with him talk

b. ??**Hem** kan ik goed **mee** praten
Him can I well with talk.

c. **Met hem** kan ik goed praten.
With him can I well talk
'For me he is a good person to talk to.'

When the object is an R-pronoun, for instance *daar*, it appears left of the preposition (18). The object may even appear alone in the Vorfeld (18c). Like before, fronting the full PP is no problem (18d).

(18) a. *Ik kan goed **met daar** praten.
I can goed with there talk.

b. Ik kan **daar** goed **mee** praten
I can there well with talk

[4]In fact, Van Riemsdijk (1978) talks of *post*position stranding.

[5]The examples show two forms of the preposition translated as 'with'. The prepositional form *met* is used when the object of the PP follows the preposition. The adverbial or independent form *mee* shows up in preposition stranding, as a postposition, or as a verb particle.

c. **Daar** kan ik goed **mee** praten
there can I well with talk

d. **Daarmee** kan ik goed praten.
There with can I well talk
'For me he is a good person to talk to.'

With this behaviour, the Dutch standard language resides between English and most Scandinavian languages on one side, and German on the other. English and the Scandinavian languages allow full fledged preposition stranding and Standard German allows none. Some German dialects do however show the Dutch type of preposition stranding (Fleischer, 2002).

Colloquial Dutch and certain dialects are more flexible, however, and may allow preposition stranding with other NPs, too (De Vries, 1911; Jansen, 1981;[6] Haeseryn et al., 1997). The examples in (19) feature topicalized full NPs with preposition stranding.

(19) a. nou **dubbel glas** stap je niet zomaar **doorheen** hoor.
PRT double glass step you not just like that through.DIR PART
'You know, you don't just walk through double glazing.' (Nl-a 676:321)[7]

b. [playing Scrabble:]
de C kun je makkelijk iets **mee.**
the C can you easily something with
"'C" is easy to use' (Nl-a 491:117)

The constraint on reduced pronouns in the Vorfeld can be observed for topicalization of objects of prepositions, too. Since they are not subjects, these objects cannot appear in the Vorfeld in reduced fom *er* or *d'r* (spoken Dutch). As a result, we have the contrasts in (20a) and (20b).

(20) a. Ik kan hier / er / d'r niets mee
I can here there.RED there.RED nothing with

b. Hier / *d'r / *er kan ik niets mee
here there.RED there.RED can I nothing with
'This / It is useless' (lit.: 'I cannot do anything with this/it').

Preposition stranding is not restricted to oblique objects. Adverbial PPs and PP indirect objects (by means of dative alternation) allow it as well.

[6]Jansen (1981) presents an emperical study of several phenomena in spoken Dutch, including Vorfeld occupation. Part of his book can thus be seen as an early precursor of the current work.

[7]Sentences that are taken from the *Corpus Gesproken Nederlands*, the spoken Dutch corpus used in this dissertation, are always marked with a region code, a component code, and a sentence number. 'Nl-a' refers to: recorded in the Netherlands, *component a*. An overview of the corpus is given in Section 3.1. The region code 'Vl' is used for Flanders.

2.2.4 Non-arguments in the Vorfeld

The Vorfeld may be occupied by adjuncts. The variety in this group is great, and I will not attempt a survey of all possibilities. Not all adjuncts front with equal ease. Intuitively, 'speaker oriented' sentence adverbs are good Vorfeld occupants, where manner adverbs, for instance, are not. Some examples are given in (21). Examples (21a) and (21b) involve a complex and a simple temporal adverbial, respectively. In example (21c) a conjunction occupies the Vorfeld. The Vorfeld in (21d) contains coordinated predicative adjuncts.

(21) a. Net als ik naar huis ga, loopt ze met me mee.
just when I to house go walks she with mee VPART
'Just when I am about to go home, she follows me.'

b. Gisteren ging ik naar de cinema.
yesterday went I to the cinema
'Yesterday, I went to the cinema.'

c. Daarom zijn de mensen zo moe.
Therefore are the people so tired
'That is why people are so tired.'

d. Dronken, dol en dwaas, beet ik in mijn bier.
drunk crazy and foolish bit I in my beer
'Drunk, crazy and foolish, I swigged at my beer.'

It is unclear whether fronting adverbials and fronting arguments are really alike. Many researchers have observed that fronting adverbials is not as restricted as fronting arguments. As said before, this dissertation focuses on subjects and objects in the Vorfeld, so I will not consider this issue much further. However, it is worth pointing out that the ban on non-subject reduced pronouns applies to certain adverbials, too (22).

(22) a. Ik heb in de disco / daar / er mijn koekie verloren.
I have in the disco there there.RED my cookie lost

b. In de disco / daar / *er heb ik mijn koekie verloren.
in the disco there there.RED have I my cookie lost
'I lost my cookie in the disco / there.'

In (22), the fronted adverbial was a locative PP or a locative adverb. Similar data can be given for at least causative and instrumental PPs, and extraction out of such PPs.

Under the header 'non-arguments in the Vorfeld', one should also mention *split topicalization* and *extraction out of NP*. In split topicalization (23a), it seems as if the head word of an argument NP (here: *onweersbuien*) is topicalized alone, leaving other material of the NP (here: *eentje*) behind. The parts of the split-NP need not be morphological identical with their counterparts in an in-situ realization (23b), however (Van Hoof, 1997, Salverda, 2000; for German: Kuthy, 2002, Féry, 2006).

(23) a. **Onweersbuien** hebben we maar **eentje** gehad.
thunderstorms have we only one had
'We have only experienced one thunderstorm.' (Salverda, 2000)
b. We hebben maar **een**(*tje) **onweersbui**(*en) gehad
c. Van Chomsky heb ik een boek gelezen.
by Chomsky have I a book read.
'I've read a book by Chomsky'

Despite the name 'extraction out of NP', it is a matter of debate whether the Vorfeld material in (23c) is an argument of the verb *gelezen*, or whether it is part of the NP headed by *boek*. Properties of the NP as well as properties of the verb influence the acceptability of the PP-topicalization (Bouma, 2004, and references therein). I will not be able to investigate these constructions in the context of this dissertation.

We have seen a number of examples that illustrated what kind of material can appear in the Vorfeld. One aspect of Vorfeld occupation that keeps returning in the discussion is the special position that subjects have with respect to the possibility of having a reduced pronoun in the Vorfeld. The next section will look into this issue in more detail.

2.3 Subjecthood and Vorfeld pronouns

The subject restriction on weak pronouns in the Vorfeld – and a similar restriction in German regarding *es* 'it' – has received a lot of attention in the literature. Some reduced and full personal pronouns have already figured in examples in this chapter. Table 2.2, p32, gives a more complete overview of the full and reduced personal pronominal paradigms (based on Haeseryn et al., 1997; Van Eynde, 1999).[8] The possessive pronouns show a similar full-reduced variation, but these are not relevant in the current discussion.

The reduced paradigm in Table 2.2 is reduced in several ways. First, most reduced forms can be considered to be phonological reductions of the strong forms. Secondly, but related to the first point, the full pronouns can be realized with prosodic prominence, but need not be, whereas reduced pronouns can never receive such prominence. Finally, the reduced paradigm shows fewer semantic distinctions, using one form for subject and object in over half of the cases, whereas the full forms in the table show this distinction throughout, except for *jullie*.[9] Compared to the full forms, the reduced forms are restricted in their distribution. We have seen this in relation to Vorfeld occupation. Another example of the restricted distribution is that reduced pronouns cannot generally be conjoined or modified (Cardinaletti and Starke, 1996; Van Eynde, 1999). On the other hand, unstressed full pronouns are marginal with inanimate referents, and ungrammatical with an inanimate referent when stressed. The obligatorily unstressed reduced pronouns show no such

Table 2.2: Full and reduced paradigms for personal pronouns in Dutch.

		Person singular			Person plural		
Function	Form	1st	2nd	3rd	1st	2nd	3rd
subject	full	ik	jij	hij / zij	wij	jullie	zij
	reduced	’k	je	-ie / ze / het / ’t	we	-	ze
non-subject	full	mij	jou	hem / haar	ons	jullie	hun / hen
	reduced	me	je	’m / (d)’r / het / ’t	-	-	ze
(demonstrative)		-	-	die / dat	-	-	die

Note: See footnote 8 for additional remarks.

restriction. In Table 2.2, I have also included the demonstrative forms *dat* and *die*. They behave a lot like the full personal pronouns, except for the fact that they have no animacy restrictions.

Observations about the non-topicalizability of *het* ‘it.RED’ in Dutch can already be found in Kruisinga (1938). More recently, the asymmetry has been used by theoretical syntacticians to argue in favour of an asymmetry in syntactic structure between canonical (that is, subject initial) and topicalized sentences in German and Dutch. Canonical word order sentences would be IPs, and have their subjects in SpecIP. Topicalized sentences would be CPs, with SpecCP containing the topicalized element. The asymmetry in topicalizability of reduced and full pronouns could then be modeled by restricting reduced pronouns to certain positions. This approach can be contrasted with approaches that assume that main clauses are always CPs. There are other issues besides reduced pronoun fronting involved in this debate, and I refer the interested reader to Zwart (1997), Gärtner and Steinbach (2003) and Van Craenenbroeck and Haegeman (2007) for comparison and references. These theoretical syntactic issues will not occupy us further here. The term Vorfeld allows us to refer to the position of interest without having to decide whether it is SpecIP or SpecCP.

[8]Some remarks on the paradigms in Table 2.2 are in order. The third person singular cells show different forms for gender (MASC, FEM, NEUTER), but the non-reduced paradigm lacks a neuter pronoun (like German *es*, which is also assumed to be weak/reduced). The form *-ie* (‘he’, reduced) behaves like a real enclitic (Weerman, 1989) – it requires attachment to a finite verb or a complementizer on its left. Furthermore, the table is biased towards (written) Northern Standard Dutch. Some omitted forms that are encountered in the Corpus Gesproken Nederlands are: *-’m* (‘he’, enclitic, Southern Dutch), *gij, ge, -de* and *u* (‘you’, full, reduced and enclitic subject, full non-subject, Southern Dutch); *ze* (‘her’, non-standard, reduced, both Northern and Southern); and *gijlen/gulle*, *wijlen* and *zulle* (‘you.PL’, ‘we’ and ‘they’, full, Southern).

[9]The exception *jullie* might be explained by its synthetic nature. One probable origin is that it formed out of second person pronoun *je* and noun *lui* (‘people’, cf. Eng. *you people/all/ones/guys*) (Philippa, Debrabandere, and Quak, 2003-9; Van der Sijs, 2004).

The data in (24) illustrates the by now familiar point that reduced pronouns cannot be topicalized. Example (24a) shows that reduced subject pronouns can occur in the Vorfeld, whereas (24b) shows that a reduced object pronoun cannot.

(24) a. Ze heeft Jan gekust.
she.RED has Jan kissed
'She kissed Jan.' (Zwart, 1997, p35)

b. *'r Heeft Jan gekust.
her.RED has Jan kissed
'Jan kissed her.' (ibid)

In contrast to reduced pronoun objects, full pronoun objects can occur in the Vorfeld. Example (25a) is to be read with the typical 'hat pattern' – a rising accent on *haar*, and the falling, nuclear accent on *gekust*, with either a low or a high level tone over *heeft Jan*. In example (25b), nuclear accent falls on the initial constituent *haar* and the verb and postverbal domain are deaccented. Slashes indicate rises and falls, capitals indicate nuclear stress. See Section 2.5 for a discussion of the information structural differences between (25a) and (25b).

(25) a. /Haar/ heeft Jan \geKUST\.
her.FULL has Jan kissed

b. \HAAR\ heeft Jan gekust.
her.FULL has Jan kissed
'Jan kissed her.'

On the basis of observations like the ones above, and similar data from German and Yiddish, Travis (1984) offers the constraint in (26).

(26) *Restriction on Topicalization:*
Unstressed pronouns may not topicalize.

The data in (24) and (25) seems to support this generalization. However, the question is what is meant by *unstressed*. Gärtner and Steinbach (2003, citing Lenerz, 1994) offer (27a) and (27b) as counterexamples to (26). Nuclear stress is on the finite verb. The fronted element, even though it appears to have some prosodic prominence, does not contain nearly as much stress as needed to make the topicalization in (25a) acceptable. And yet, example (27c) shows that a reduced pronoun is not allowed, even with neighbouring nuclear stress.

(27) a. Dich KENN ich doch
you know I PRT German

b. Jou KEN ik toch
you.FULL know I PRT Dutch

c. *Je KEN ik toch.
you.RED know I PRT
'(But) I KNOW you.'

We might be tempted to recast the constraint in (26) in terms of reduced pronouns. Still, according to Gärtner and Steinbach (2003), this runs into problems in two ways. First, examples of reduced non-subject pronouns in the Vorfeld *do* exist in the literature. Secondly, Gärtner and Steinbach claim that reduced pronoun *subjects* also show a tendency to avoid the Vorfeld, but that this tendency is not nearly as strong as the one observed for reduced pronoun objects.

Apropos of the first point, Gärtner and Steinbach present examples of reduced pronoun objects in the Vorfeld from colloquial German and German dialects. Weerman (1989) presents a Dutch example (28).

(28) 't Hebben we 'm gisteren nog verteld
it.RED have we him.RED yesterday PART told
'We told him that yesterday.' (Weerman's judgement)

The problem with a sentence like (28) is that I find it only marginally acceptable – an intuition shared by other native speakers of Dutch when asked. However, I do agree that it is not nearly as bad as (24b). Gärtner and Steinbach argue that we should look for examples of topicalized reduced pronouns in colloquial and/or dialectal data. In Chapter 4 I will discuss whether the corpus that I used to investigate Vorfeld occupation contains instances of topicalized reduced pronouns.[10]

Gärtner and Steinbach's second claim – that subject reduced pronouns also avoid the Vorfeld, but to a lesser extent – is something that is hard to establish on the basis of intuition data. However, frequency data would be ideal to evaluate such a claim. In Section 4.3.3, I investigate whether reduced pronoun subjects appear in the Vorfeld less often than full pronoun subjects do. Indeed, we will see that weak pronoun subjects are less likely to appear in the Vorfeld than full pronoun subjects. Gärtner and Steinbach propose that there is a phonological reason for the dissociation between the Vorfeld and reduced

[10]There is a class of examples with reduced pronoun *er* in the Vorfeld, when this *er* is, for instance, the object of a preposition (i). The preposition and its object are in boldface.

(i) **Er** zit een ondeugdelijk slot **op**
there.RED is a bad lock on
'The lock that is on it is no good.'

Examples like (i) have been analyzed as *er* actually being a expletive subject in a existential construction from early on (Bech, 1952; Bouma, 2000). The examples do therefore not necessarily form counterexamples to the reduced object pronoun restriction. Purported German counterexamples in Meinunger (2007), with Vorfeld object *es* can be analyzed in the same way. They all involve a pronoun that is homophonous with the expletive subject, and meet the requirements for the German existential constructions.

pronouns; however in the course of this dissertation, we will see two further possible explanations. The first alternative explanation refers to information structure: Weak pronouns avoid the Vorfeld because the information structural properties of the Vorfeld are incompatible with personal pronouns in general, and reduced personal pronouns in particular (Sections 2.5 and 2.6). Section 4.3.3 offers a comparison of the phonological account and the information structural account with the help of corpus data. The second alternative explanation is of a very different kind. In Chapters 5 and 6, I argue that the ease with which grammatical function assignment can be recovered from a string has an influence on freedom of word order. The effect this has on the placement of personal pronoun subjects will be spelled out and investigated in Section 6.2.1.

Matters surrounding the Vorfeld pronouns are further complicated by the availability of demonstrative pronouns in third person (also in Table 2.2, p32). Examples are given in (29).

(29) a. Die heeft hij geKUST
DEM has he kissed
'He kissed her/him/it/them.'

b. Dat WEET ik niet
DEM.N know I not
'I don't know.'

The demonstratives in (29) can be anaphoric. The demonstrative in (29a) may have a human referent. The mechanisms behind the choice between the demonstrative and the personal forms are still not very well understood, but it seems that demonstratives can be used to refer to a discourse salient referent which is not at that point the established or continuing discourse topic (Comrie, 2000; for German: Bosch, Katz, and Umbach, 2007). A typical examples is a demonstrative pronoun that refers to an entity that was recently (re-)introduced (and thereby salient, but not topical), or to abstract objects such as events, propositions, etcetera. Topicalized demonstrative pronouns do not have to be stressed. In fact, the demonstrative in (29b) need hardly receive any prosodic prominence, and can be reduced to [də] or even [d].

I will return to the issue of pronouns in the Vorfeld in Section 2.5, when I discuss the information structural properties of topicalization in Dutch. Investigating the distribution of pronominal elements in the Vorfeld fits in naturally with the investigation of the corpus predictions discussed in Section 2.6.

2.4 Violations of V2

Until now, I have assumed that there is exactly one preverbal constituent: the Vorfeld occupant. This assumption is justified when we consider a sentence where a constituent is

topicalized, but the subject does not move to the Mittelfeld. The result is ungrammatical, as illustrated in (30).

(30) *Die ik heb niet gezoend
DEM I have not kissed
'I haven't kissed him/her.'

In (30), there are two preverbal constituents, *die* and *ik*, that cannot be analyzed as one. As a result, the finite verb is not in second position. The V2 generalization is a simplification of the facts, however. There are numerous exceptions to V2 in declarative main clauses – at least at a superficial level. The finite verb may occur directly at the left edge (V1), or there may be more than one constituent before it (V3, etcetera). The discussion of non-V2 examples that is to follow serves to indicate the boundaries of the constructions that I will investigate in this dissertation.

2.4.1 No elements in the Vorfeld

Under certain circumstances, a Vorfeld element may be dropped. The result is a declarative sentence with no Vorfeld constituent. Subjects and direct objects can be dropped, as well as the objects of oblique arguments that have left their preposition behind (31). Object drop is more common than subject drop (Jansen, 1981). As the data in (32) illustrates, dropping indirect objects is less acceptable than direct objects or subjects (Thrift, 2003).[11]

(31) a. *subject*
(ik) Was inmiddels getrouwd.
I was by then married
'By that time I had married.'

b. *direct object*
(dat) Wil ik wel geloven.
that want I PRT believe
'I can believe that.'

c. *object of a oblique argument*
(daar) Heb ik veel aan gehad.
there have I much on had
'It's been useful to me.' (all from Jansen, 1981)

[11] Thrift (2003) bases her claims on data collected with a questionnaire. Thrift also considers demonstrative pronoun indirect objects ungrammatical in Dutch. She connects this alleged ungrammaticality of demonstrative pronoun indirect objects with the reduced *dropfähigkeit* of indirect objects. In my opinion, demonstrative indirect objects are fine, but the results in Section 4.3 will show that they are very infrequent.

(32) a. *experiencer indirect object*
?(Die / Haar) Bevalt het prima.
DEM her pleases it fine
'She likes it.'

b. *recipient indirect object*
?(Die / Haar) Geef ik dit cadeau.
DEM her give I this present
'I'll give this present to her.'

The sentences in (31) and (32) are recognizable as topic drop sentences, because they miss an argument. However, adjuncts can sometimes be dropped, too, which results in a sentence without a 'hole'. Such a sentence (33) will look exactly like any other V1 construction, for instance like a polar interrogative or an the antecedent clause in a conditional.

(33) (Toen) Ben ik in dienst gegaan.
then am I in service went
'Then I went into military service.' (from Jansen, 1981)

There is a interesting paradox concerning object pronouns in the Vorfeld and topic drop. Pronominal objects that appear in the Vorfeld cannot be reduced, as we have seen. They have to be stressed personal pronouns, or demonstrative pronouns. However, in many of those cases dropping the pronoun completely is unproblematic. The sentences in (31b) and (32) are cases in point: They would be ungrammatical with a reduced personal pronoun.

The differences between subjects, direct objects and indirect objects, and the paradoxical reduced pronouns and topic drop, mean that topic drop would be a very interesting subject for a corpus study. However, the large scale studies that I present in this dissertation are not suitable for this. The size of the corpus is such that manual investigation is too time consuming. Automatically establishing when something is dropped or missing is already slightly problematic. Automatically figuring out what has been dropped would appear to be nearly impossible. Hence, I do not look at topic drop in this dissertation.

2.4.2 Left dislocation and hanging topics

There are two common constructions in Dutch in which the verb does not appear to be in second position, but in third. It is not clear to which extent the extra elements actually appear in the clause, or outside of it (in terms of the topological field template: in the *lead*). The two constructions are (contrastive) left dislocation (CLD) and hanging topic (HT). An example of CLD is given in (34) and in (35) is an example of HT. Capitals indicate nuclear accent, '||' indicates a not further specified prosodic discontinuity.

(34) De kroket die had ze al OPgegeten.
the croquette DEM had she already eaten
'She had already eaten the croquette.'

(35) De kroKET, || zij had 'm al OPgegeten.
the croquette she had it already eaten
'She had already eaten the croquette.'

Superficially, the two constructions differ only in the place of the resumptive pronoun (preverbal in CLD, unrestricted in HT), and the type of the pronoun that corefers with the initial constituent (demonstrative in CLD, unrestricted in HT). However, it has long been observed that the two constructions are in fact not very similar. Apart from different discourse functions (Frey, 2005, for German, which has basically the same distinction as Dutch), the two constructions differ in 'connectedness'. The left dislocated constituent in CLD (34) behaves like part of the clause in terms of intonation and binding, while the initial element in HT (35) does not (Van Riemsdijk and Zwarts, 1997; Zaenen, 1997; German: Altmann, 1981; Grohmann, 2003; Frey, 2005).[12] The tighter connection of contrastive-left-dislocated material with the clause is also demonstrated in the contrast in (36). When there are two constituents 'before' the Vorfeld, it is the left-most (that is, outermost) one that is the hanging topic.

(36) a. Dat boek$_2$, || mijn moeder$_1$ die$_1$ zou ik 't$_2$ niet eens durven laten zien.
that book my mother DEM would I it not even dare to show
'I would not even dare to show that book to my mother.' (Zaenen, 1997)
b. *Mijn moeder$_1$, || dat boek$_2$ die$_1$ zou ik 't$_2$ niet eens durven laten zien.

The ungrammaticality of (36b) is due to the fact that the resumptive pronouns do not agree under the interpretation that is forced by the order of the preverbal elements. The outermost element has to be the hanging topic, the innermost element the contrastive-left-dislocated constituent. However, the demonstrative preverbal pronoun agrees with the outermost element, and the personal postverbal pronoun agrees with the innermost.

An important question in the context of this dissertation is whether we can safely ignore these constructions in our corpus investigation. Although it may be clear that HT does not really resemble topicalization, the same is not true for CLD. As a matter of fact, topicalization has been analyzed as a combination of CLD and topic drop (for instance Zwart, 1997; also Odijk, 1998, for clausal complements).[13]

[12]For many other languages two constructions have been observed that are similar and different in much the same ways as Dutch and German CLD and HT for Dutch. See Grohmann (2003), for a comparison with Romance data.

[13]Hanging topics can also be combined with topic drop, but the construction is clearly distinguishable from a topicalized sentence by prosody and other measures of connectedness, and possibly by a hole in the argument structure.

However, CLD is only possible in a subset of the cases that topicalization is allowed in. For instance, contrastive left dislocation is not allowed with certain quantified NPs, reflexive pronouns or foci.

(37) a. Geen enkele film van Godard (*die) heeft hij gezien.
No single film by Godard DEM has he seen.
'He hasn't seen a single film by Godard.' (Zaenen, 1997)

b. Zichzelf, (??die) respecTEERT hij niet.
himself, DEM respects he not
'He doesn't respect himself.' (Zwart, 1997)

c. Wie zou je wel 's willen ontmoeten?
'Who would you like to meet?'
ELvis / ??ELvis die / *Elvis DIE zou ik wel 's willen ontmoeten
Elvis Elvis DEM Elvis DEM would I PART want meet
'I would love to meet ELVIS sometime.'[14]

It seems that contrastive left dislocation, although perhaps formally related or similar to topicalization, does not have the same function as topicalization has (Frey, 2005, for a much more thorough analysis but with the same conclusion for German). I therefore deem it safe not to include contrastive left dislocation in the corpus study in this dissertation. However, the comparison of the two constructions would be an interesting issue for future research (see Snider and Zaenen, 2006, on English topicalization and left dislocation).

2.4.3 Multiple elements in the Vorfeld

Other cases of multiple constituents in the Vorfeld cannot as easily be analyzed as containing clause external constituents. Whether these cases should actually be considered to be instances of multiple Vorfeld occupation or whether the multiple constituents should be analyzed as one at some level is up for debate.

In the syntactic annotation of the corpus that is used in this dissertation (see Chapter 3), preverbal focus particles are not attached to a fronted NP, but rather to the sentence as modifiers. As a result, these constructions contain two constituents in the Vorfeld, resulting in V3. The finite verb is in boldface, the two preverbal constituents are shown within brackets.

[14]Note that I am assuming a left dislocation prosody here. Pronouncing the answer as a hanging topic, with a break between the initial constituent and the Vorfeld, and with a nuclear accent on both the initial constituent and the demonstrative as in (i), is unproblematic.

(i) ELvis. || DIE zou ik wel eens willen ontmoeten.

(38) [zelfs] [Thomas' Summa Theologica] **was** in Parijs ooit verbrand.
even Thomas' Summa Theologica was in Paris once burned
'Even Thomas' Summa Theologica would have been burned in Paris at some point.'
(Nl-o 801475:11)

An analysis of focus particle attachment in German with the same consequence is given in Büring and Hartman (2001). Bouma, Hendriks, and Hoeksema (2007) argue in favour of an analysis that treats a focus particle and the XP that contains the associated focus as one constituent, partly on the basis of wishing to retain the V2 generalization.

However, there are other particles that may appear before the finite verb together with another constituent for which a one constituent analysis is not as likely. Example (39a) contains the discourse particle *dus* before the Vorfeld. Interestingly, *dus* may occur as a single Vorfeld constituent as well (39b), and can surface in several positions throughout the sentence (39c).

(39) a. [dus] [inzake preventie] **is** er niets geweest.
so concerning prevention has expl nothing been
'So nothing was done to prevent it.' (Vl-j 600065:60)
b. Dus **is** er niets geweest inzake preventie
'As a result nothing was done to prevent it.'
c. Inzake preventie (dus) **is** er (dus) niets geweest (dus).

It is unclear whether all examples in (39) have the same interpretation. Sentence (39b), where *dus* occurs in the Vorfeld alone, only allows for a causal reading. Examples (39a) and (39c) also have readings where *dus* is speaker oriented. Meinunger (2004) suggests that in German V3 surfaces in cases like (39a) to force a 'speech act reading' of adverbs like *ehrlich* 'honestly' (German) vs a sentence internal predicate reading. This contrast between V2 and V3 will briefly be discussed in the context of interpretation constraints on word order variation in Section 5.7.

Combinations of arguments and something else in the Vorfeld are rare. The sentence in (40) has received an annotation in the corpus used in this dissertation in which the subject and a PP that modifies the sentence are preverbal. This annotation is (presumably) motivated by the preferred interpretations of the sentence: The PP does not specify the location of the kitchenette, but it specifies where the kitchenette has the mentioned price.

(40) [$_{NP}$ een kitchenette] [$_{PP}$ in den Hubo] **kost** zevenduizend frank en nen
a kitchenette in the Hubo costs seven thousand franc and a
douche ook.
shower too
'A kitchenette costs BEF 7000 at Hubo and a shower, too.'
(*Not:* 'A kitchenette located at Hubo is BEF 7000') (Vl-d 900058:271)

However, changing the order of the preverbal NP and PP is not possible, which suggests that the PP is part of the NP after all. More common is a combination of modifiers, for instance locative and temporal. They function as complex frame-setting adverbials, indicating both time and place, or manner and time, etcetera, of the event at once. In these cases the order of the constituents could be reversed.

(41) a. [in de Brandpunt] [met uitgaan] **hadden** ze 'm dus ook al.
in de Brandpunt with going out had they him.RED PRT PRT PRT.
'So, in De Brandpunt, when we/they were going out, they already had it.'
(Nl-a 389:258)

b. [straks] [op Klara] **is** er Lut Van Der Eycken en ook ...
soon on Klara is EXPL Lut Van Der Eycken and too
'In a few moments you can listen to LVDE, and ..., too, on Klara.'
(Vl-f 600840:50)

Multiple Vorfeld occupancy of the type illustrated in (40) and (41) is also observed in German. Müller (2005) presents a wide range of German data, and proposes a HPSG analysis. In his analysis, the Vorfeld is occupied by one verbal constituent with an unpronounced head, so that the multiplicity is only apparent and V2 is respected.

2.4.4 Two left brackets

In spoken Dutch, it is not uncommon to find sentences in which the finite verb and possibly some other material is repeated after a part of the Mittelfeld. Effectively, the sentence contains two left brackets, meaning that the topological field template given in Figure 2.1 will not fit as neatly anymore. Haeseryn et al. (1997) describe the construction as involving two overlapping templates.

The construction is called a *herhalingsconstructie* 'repeating construction', but also *spiegel-* 'mirror-' (Huesken, 2001, unseen; Van der Wouden et al., 2002) or *overloopconstructie* 'overflow construction' (Haeseryn et al., 1997). In the example the finite verbs are in boldface.[15]

(42) we **zijn** min of meer **zijn** we 't ermee eens
we are more or less are we it.RED with it agreed
'We kind of agree with it.' (Nl-n 61:26)

Note that it is typically, but not necessarily, the case that the two finite verbs are the same. Informal inspection of corpus examples also shows examples in which there are changes

[15] To give you an idea of the frequency of this construction, I note that in the Corpus Gesproken Nederlands, $\sim$ 1.2% (943/77316) of SMAINs (roughly: declarative main clauses) contains more than one finite head verb. This means that the construction is as frequent as, for instance, the presence of an indirect/dative object in a sentence.

in agreement *gaan... ga* (plural... second singular 'go'), aspect *is... wordt* ('is' & 'becomes'), modality *kan... gaat* ('can go'... 'goes'), and tense *heet... heette* ('is called'... 'was called'), or which the verbs are near synonyms. An in depth analysis of the construction can be found in Huesken (2001, unseen).

2.4.5 Summary

We have seen a range of examples where the V2 generalization does not clearly hold, and only additional assumptions about the syntactic structure can decide whether V2 is respected or not. These examples were discussed to show where the boundaries of the investigation in this dissertation lie. As said in the introduction, I constrain the data in the corpus investigation to sentences that contain exactly one preverbal constituent. This constituent is be referred to as the Vorfeld occupant. A definition of Vorfeld occupant in terms of the annotation that is available in the corpus that is used will be given Section 3.3. Topic drop, contrastive left dislocation, hanging topics, multiple Vorfeld occupants and mirror constructions are not be investigated.

This choice is partly made on methodological grounds. In this dissertation I investigate properties of the Vorfeld occupant such as grammatical function, definiteness and grammatical complexity (see Section 2.6). Defining these properties on *the* Vorfeld occupant, when in fact there are multiple or none, is tricky. Furthermore, we saw that some constructions differ in functional and formal properties to the extent that they warrant a separate investigation, that is, topic drop, left dislocation, hanging topics and mirror constructions.

2.5 Topicalization and information structure

I have used the term topicalization to refer to the placement of non-subject arguments in the Vorfeld. What distinguishes topicalization from canonical subject initial word order, is that topicalization is restricted with respect to information structure. A subject in the Vorfeld is not restricted. In this section I will discuss the information structure restrictions on Vorfeld objects. Following Gundel (1974), we can distinguish two main types of topicalization: focus topicalization and topic topicalization. I will discuss these two in turn. We shall see that it is difficult to give a clear unifying principle of all examples of topicalization. I will invoke the concept of informational *importance* to explain the observations (Givón, 1988; Gundel, 1988). Although this concept is vague, we will be able to draw some plausible parallels between it and pronominal form. Informational importance can thus offer an explanation of the special relation between reduced pronouns and the Vorfeld.

2.5.1 Focus topicalization

The part of the sentence that is in *focus* contains material that is 'informative, newsy, dominant, or contrary-to-expectation' (Vallduví and Engdahl, 1996, p462). I will refer to non-focussed material as *background* material. We may induce a focus/background division using a constituent question (43), or an explicit contrast (44).

(43) A Wie heb je gisteravond gekust?
'Who did you kiss last night?'
B Ik heb gisteravond [$_{\text{focus}}$ Grace \KELLY\] gekust.
I have last night Grace Kelly kissed.
'I kissed Grace Kelly last night.'

(44) A Je hebt Gene Kelly gekust!?
'You kissed Gene Kelly!?'
B Nee, Ik heb [$_{\text{focus}}$ \GRACE\ Kelly] gekust!
No, I have Grace Kelly kissed
'No, I kissed Grace Kelly!'

In Dutch, focus is associated with prosodic prominence. In the examples in (44), the nuclear accent falls within the focussed constituent.

Focus topicalization puts the focussed constituent in the Vorfeld. The following sentences are responses to (43A) and (44A), respectively. As a result of focus topicalization, the direct objects are in the Vorfeld.

(43) B″ [$_{\text{focus}}$ Grace \KELLY\] heb ik gisteravond gekust.

(44) B″ Nee, [$_{\text{focus}}$ \GRACE\ Kelly] heb ik gekust!

Because of focus topicalization, nuclear accent resides in the Vorfeld. The rest of the sentence, the finite verb and the postverbal domain, is deaccented. The example of a topicalized stressed personal pronoun (25b), in Section 2.3, is a case of focus topicalization.

Jansen (1981) observes that focus topicalization in Dutch is rare. In a corpus he collected through interviews with native speakers, he found only a few cases of focus topicalization. Amongst these is the example in (45). The focus brackets were added by me.

(45) A What did you make in handicraft class?
B [$_{\text{focus}}$ zo'n Engelse theemuts] heb ik onder andere eens gemaakt
such an English tea pot cover have I amongst other once made
'Amongst other things I made one of those English tea pot covers.'

In spite of its rareness, the example in (45B) feels natural. It is in my ears completely equivalent in meaning to a canonical, subject-initial realization (45B′).

(45) B′ Ik heb onder andere al eens [$_{\text{focus}}$ zo'n Engelse theemuts] gemaakt.

Recall from Section 2.3 that reduced pronouns in Dutch cannot receive prosodic prominence. They can therefore not be focussed (46B). And, since focus on a reduced pronoun is not possible, focus topicalization is not compatible with a reduced pronoun either (46B′).

(46) A Who did you kiss?
B *Ik heb [$_{\text{focus}}$ 'r] gekust.
I have her.RED kissed
'I kissed her.'
B′ *[$_{\text{focus}}$ 'r] Heb ik gekust.

Thus, focus topicalization will not lead to a reduced pronoun in the Vorfeld.

2.5.2 Topic topicalization

In *topic topicalization*, a non-subject constituent moves to the Vorfeld, just like in focus topicalization. However, in topic topicalization, there is a focus after the Vorfeld. Nuclear stress falls on a syllable inside the focussed material. Therefore, the finite verb and postverbal domain are not deaccented in a topic topicalization. What kind of material appears in the Vorfeld in a topic topicalization may differ. We may distinguish the cases with contrastive material in the Vorfeld, and cases with non-contrastive material in the Vorfeld. Let me start with the former.

In *contrastive topic topicalization*, the material in the Vorfeld is prosodically prominent. An example is given in (47). There are now two prominent accents in the sentence: an accent in the contrastive topic in the Vorfeld, and nuclear accent in the postverbal focus. The initial accent receives a marked rise, possibly preceded by a slight dip, and the nuclear accent is a fall. The result is the so called *hat pattern*. See Jacobs (2001), Braun and Ladd (2003), and Féry (2006) for different characterizations of the rising accent in German contrastive topics. See 't Hart (1998) for a general description of the Dutch hat pattern.

(47) [$_{\text{cont. topic}}$ De /klare/] krijgt [$_{\text{focus}}$ \HJALLIS\].
The clear one gets Hjallis
'Hjallis gets the Dutch gin.' (OVS)

The topicalized stressed pronoun in (25a), Section 2.3, is also an example of contrastive topic topicalization.

There is a clear interpretation effect involved in using the contrastive topic construction. The utterer of (47) is either going to say something about, or is implying something about, alternatives of the Vorfeld constituent (in this case: other drinks). This inherent contrast means that contrastive topic is similar to focus as described above. Indeed, it has been

argued that contrastive topics are foci (Van Hoof, 2003), that they are semantically related to foci (Büring, 2003), or that they contain a focus (Steedman, 2000).

Roberts (1996), and later Büring (2003), argue that contrastive topic constructions like (47) should be understood as strategies to answer complex questions. For (47) we can imagine a complex question 'who is drinking what?'. One way to answer this question is to start listing the drinks, and ask for each drink who is drinking it. This scenario is made explicit in (48).

(48) In a bar, the waiter brings drinks to a table, and wonders: who is having what?

Who is having beer?	Het biertje gaat naar Gillis. the beer goes to Gillis	(SVO)
Who is having Dutch gin?	De klare krijgt Hjallis. the Dutch gin gets Hjallis	(OVS)
Who is having red wine?	De rooie wijn is voor Gunnis. the red wine is for Gunnis	(SVO)

In each case there are prominent accents on both the initial element (the contrastive topic) and the postverbal element (the focus). This is independent of whether the elements are objects or subjects. The choice for a specific answering strategy is not fixed by the question. In the case of the complex question answered in (48), one might as well have started with the drinkers and revealed for each drinker what they are drinking.

Like (ordinary) focus, contrastive topic is associated with prosodic prominence. Therefore, like focus, contrastive topic cannot fall on a reduced pronoun, independent of whether the contrastive topic is put into the Vorfeld or not. Contrastive topic topicalization will not lead to a non-subject reduced personal pronoun in the Vorfeld.

Not all cases of topic topicalization are contrastive. I will now turn to examples of non-subject Vorfeld occupants that do not involve contrast. These cases are very hard to characterize, and I will only be able to give an intuitive characterization. An example of a topic topicalization that does not involve contrast of the topicalized element is given in (49). Note that both sentences contain direct speech, and that the second sentence is a self-quotation (indicated in the translation).

(49) A hij zegt ja je kent Margarita toch wel
'He says: "surely you know Margarita?"'
B ja die ken ik wel maar
yes DEM know I AFF but
'(I said:) "Sure I know her, but…"' (Nl-a 389:77–78)

Sentence (49B) starts with the demonstrative *die*, referring to Margarita. Nuclear accent in (49b) falls on *ken*, which makes (49B) an example of topic topicalization. However,

the referent of the demonstrative pronoun is not contrasted with anything. There is no indication that the speaker will go on to talk about other people besides Margarita. Sentence (49B) is not a contrastive topic topicalization. A prosodic difference between contrastive topic topicalizations and the topic topicalization in (49) is that the topicalized *die* does not have to receive much prosodic prominence in (49).

The demonstrative pronoun in the Vorfeld feels very typical of a construction like (49). Still, in the context of (49A), it would have been possible to refer to Margarita in (49B) with a definite full NP, a proper name, or a full personal pronoun (49B′).[16] A reduced personal pronoun is ungrammatical. When the direct object is not topicalized and realized postverbally, a reduced personal pronoun is available as a referential form, too (49B′).

(49) B′ Dat meisje / Margarita / Haar / *’r ken ik wel.
that girl Margarita her.FULL her.RED know I AFF but
B″ Ik ken dat meisje / Margarita / haar / ’r wel.

This presents us with a problem, though. In the case of contrastive topic topicalization, and in the case of focus topicalization, we could explain the fact that a reduced personal pronoun could not be topicalized by appealing to the fact that the context did not allow a reduced personal pronoun to be used at all. However, this will not explain the data in (49B′), since (49B″) shows that a reduced pronoun can be used in the context. On a more general note, we can ask ourselves whether this is any connection between examples like (49B) and contrastive topic topicalization, apart from the fact that nuclear focus does not fall on the topicalized constituent.

In their study of non-canonical word order in English, Birner and Ward (1998) argue that being related to a contextually available referent through so called *partially-ordered-set* relations is a necessary condition for topic topicalization. Birner and Ward thus propose a weakened givenness requirement. According to them, a given NP may always be topicalized. An (indefinite) NP realizing a new referent may be topic-topicalized as long as the referent can be related to a given referent – where *related to a referent* means: to be a subset or part of it, to be of the same type as it, or to come before or after it in some given order, etcetera. This criterion is met by both contrastive topic topicalization, and by the examples of Vorfeld demonstratives in (49B). In the contrastive topic examples above, each contrastive topic (a drink) came from a contextually available set (the drinks on the waiter’s tray). In the case of (49B), the demonstrative coreferred with Margarita, who was mentioned in the sentence before. However, this relaxed givenness requirement does not explain why the constituent referring to Margarita in (49B) and (49B′) cannot be a reduced pronoun. The status of the referent as given does not change, only the way the referent is referred to does. As a result, being related to a contextually provided referent cannot be more than a necessary condition.[17]

[16] In my ears the version with *haar* ‘her.FULL’ sounds stilted.

Other researchers have proposed that appearing at the left periphery is indicative of being an *aboutness*-topic, and especially of being a changed – or *shifted* – aboutness-topic (Gundel, 1974; Reinhart, 1982; Givón, 1983; Gundel, 1988; Jacobs, 2001; Frey, 2006; Frascarelli and Hinterhölzl, 2007, amongst others).[18] An aboutness-topic is the entity about which a sentence can be understood to convey information. We speak of a shifted aboutness-topic when the aboutness-topic of a sentence is not that of the preceding sentence. The problem with aboutness is that it is hard to pin down. For instance the sentence (49B) could intuitively be understood as supplying information about the speaker (*ik* 'I'), as well as about *Margarita* (see McNally, 1998, for a critique of aboutness-topics; see Ward, 1988, Prince, 1998 for a discussion of the applicability of aboutness-topic tests). I will not try to answer the question of whether the concept of aboutness-topic is useful for the analysis of word order in Dutch. This will have to remain a question for further research.[19]

However, even without fully understanding the concept of aboutness-topic, we can see an interesting parallel between shifted aboutness-topics on one hand, and focus and contrastive topic on the other: Shifted aboutness-topics are not new or contrastive as such, but they are new in their role of aboutness-topic. On the assumption that aboutness-topics have a tendency to remain the same over a stretch of discourse, a shifted aboutness-topic is less predictable than a continuing one. Gundel (1988) lets the cross-linguistically observed tendency to put foci, contrastive topics and shifted topics at the front of a sentence follow from the principle stated in (50).

(50) *First-things-first*:
Provide the most important information first.

Gundel considers information that is new, unpredictable, contrastive or emphasized to be important. A very similar principle is proposed by Givón (1988) as *attend first to the most urgent task*, citing low predictability as a cause of urgency. Also see Herring (1990) for a comparative overview and discussion.

[17]In fact, Birner and Ward present data that suggests this, too. The contrast in (i) shows that a fronted constituent is subject to additional restrictions (the demonstrative determiner), even when the its referent is given.

(i) I have a recurring dream in which. . . I can't remember what I say. I usually wake up crying.
  a. This/#The dream I've had maybe three, four times.
  b. I've had this/the dream maybe three, four times. (their 284, p226)

They do not offer an explanation for this additional restriction.

[18]The first four cited works are of a general and/or cross-linguistic nature. The second three works are about German word order or partly about German word order. The works may show considerable differences in the details, but they share that the left periphery is related to (shifted) aboutness-topics. More references on the subject can be found in the cited works.

[19]I hope that the results presented in the rest of this thesis can be of help in this research.

I propose that we consider the Vorfeld, as a left-peripheral position in the Dutch clause, a place for important material. Topicalized constituents have to be in some sense important. In the case of focus topicalization and contrastive topic topicalization, the fronted material is important because it is new or contrastive. However, the importance of the fronted material in (49B), repeated below, still needs to be established.

(49) B die ken ik wel
DEM know I AFF
'I know her.'

Recall that the fronted material is not focussed, nor contrastive. What would justify calling it important, then? I think that the behaviour of demonstrative pronouns in discourse is key. Personal pronouns pick up referents that have typically been a) mentioned continuously in preceding discourse, b) realized as subjects and c) pronominalized before. Personal pronouns realize referents of which it is highly predictable that they will be mentioned again (for English Arnold, 1998; see also Arnold, 2006, for an overview and implications of this perspective on pronominal use). Demonstrative pronouns however, pick up referents that are either recently introduced, have not been repeatedly mentioned, not realized as a subject or not pronominalized before (for Dutch: Comrie, 2000, Kaiser and Trueswell, 2004; for German: Bosch, Katz, and Umbach, 2007). Compared to personal pronouns, demonstrative pronouns realize referents of which is it less predictable that they would be mentioned again. According to Gundel (1988) and Givón (1988), this difference in predictability translates into a difference in importance. The lower predictability of mention of a referent of a demonstrative pronoun means it qualifies as a Vorfeld occupant. This predictability should be understood in general terms, and is not completely a case of the context, though. For instance, that Margarita would be mentioned again in (49B) is rather likely. As (49B″) illustrated, the speaker of (49B) could have chosen to use a reduced personal pronoun. However, by *presenting* the referent as so highly predictable, and therefore unimportant, one loses the option of realizing it in the Vorfeld.

Non-subject material in the Vorfeld is characterized by its (relative) importance. Focus topicalization and topic topicalization involve material that is new, contrastive or relatively unpredictable. The relation between important material and the left periphery is made by Gundel's first-things-first principle, based upon cross-linguistic investigation of word order.

Even though I think it is appealing at an intuitive level, the characterization of the Vorfeld as containing important material leaves open many questions that will have to be answered in future research. The notion of importance is very vague, and for it to be used as a robust linguistic concept, one should try to find ways to measure importance. Givón (1988) gives several corpus based measures of predictability, and thereby of

importance. Future work should try to extend this work on larger corpora and with modern statistical methods (see Snider, 2005 for some results for English). By assuming that the lack of ability to realize important material is what separates personal pronouns, and especially reduced personal pronouns, from all other forms of NPs, we explain why they should behave differently when it comes to topicalization. Even if I am not able to measure importance at this point in the corpus that I will be using, I will be able to see whether the corpus data supports a split between personal pronouns and everything else.

The appeal to the first-things-first principle does make a further prediction that I have not mentioned yet. I have focussed on topicalized constituents, and I have tried to argue that they have to be important in order to appear in the Vorfeld. However, Gundel's statement of the first-things-first principle does not mention topicalization or non-canonical word order anywhere. In contrast, the principle is meant to be a universally applying (but violable) principle. Now, if the Dutch Vorfeld is preferably a place for important material, subjects should also be sensitive to this. Non-important subjects should be repelled from the Vorfeld, too. In particular, reduced pronoun subjects should show a tendency of avoiding the Vorfeld. This tendency cannot be as strong as with objects, because, obviously, reduced pronoun subjects are grammatical in the Vorfeld. However, we may see it as a statistical trend in the corpus. Thus the assumption that the Vorfeld is a place for important material makes the same prediction that Gärtner and Steinbach (2003) made on prosodic grounds. Section 4.3.3 compares the two approaches on the basis of corpus results.

2.6 Word order trends

In the preceding four sections, we have seen many sides of the preverbal domain in Dutch, and especially of material in the Vorfeld. We have seen what kind of material can occupy the Vorfeld, in terms of form and function. I have also contrasted clauses with exactly one Vorfeld occupant with clauses that deviate from this, in order to show which constructions will be investigated in the dissertation. The chapter thus far has been of a descriptive nature, although we have come across the 'accidental' question to be investigated quantitatively. In the brief discussion of the relation between negation predicate topicalization, I announced that the relation between sentence adverbs and direct object topicalization is investigated in Section 4.6. The discussion of the special status of pronouns in the Vorfeld brought up a claim put forward by Gärtner and Steinbach (2003), which is that reduced pronouns show a tendency to avoid the Vorfeld across the board. This prediction is particularly interesting because it can be made on the basis of the discussion of the information structural properties of the Vorfeld, too.

In this section, I will be more systematic in generating expectations or hypotheses to investigate quantitatively. The aim of this dissertation is to gain insight into the choice

of a Vorfeld occupant. In particular, we want to learn more about fronting of subjects, direct objects and indirect objects. The first step towards understanding this choice is to investigate which properties of potential Vorfeld occupants influence whether they end up in the Vorfeld or not. Put differently, we seek to answer the question: How do constituent properties relate to the chance that this constituent is fronted?

Before we can begin to generate expectations for the corpus research, we must decide what kind of properties we can expect to play a role in the choice of a Vorfeld occupant. We have already seen one property that (at least non-subject) constituents must have if they want to appear in the Vorfeld; they have to be important. Importance is not a concept we can directly use in a corpus study. However, I proposed that we could link NP form to importance, and use this to investigate whether the Vorfeld indeed prefers important material over unimportant material.

Importance is a property that links a constituent directly to the Vorfeld. However, other properties may cause a constituent to appear in the Vorfeld because these properties are associated with early realization in the sentence in general. In that case, Vorfeld occupation by a constituent is a result of the fact that the Vorfeld 'happens' to be the earliest possible position. Similarly, some properties may be associated with late realization in the sentence. A constituent with such a property may be prevented from appearing in the Vorfeld. Properties that promote early or late realization in the whole sentence form global word order trends. The general hypothesis that there are global word order trends that also affect Vorfeld occupation is a rich source of concretely testable hypotheses about trends in fronting. There is a large literature on word order in the Mittelfeld of German and Dutch, and on word order in the postverbal domain of English. The results from this literature can be used to generate predictions about the relation between certain constituent properties and Vorfeld occupation.

In the rest of the section, I will look at three factors in postverbal word order. On the basis of each, I will formulate concrete questions that can be answered quantitatively. In each case, the concrete questions are the result of applying the global word order hypothesis to the respective factor. The three factors are *canonical argument order*, *definiteness*, and *grammatical complexity*. These factors have been extensively discussed in the literature on Dutch, German, and English. They also have the advantage that they can be fairly robustly measured or annotated, which facilitates large scale corpus analysis.

2.6.1 Canonical argument order

Results about the canonical or unmarked order of subject, indirect object, and direct object depend on what is considered to be canonical or unmarked. Lenerz (1977) defines markedness of German Mittelfeld argument order in terms of distribution. Of two word orders, the one that appears only under special circumstances, whereas the other can be

used freely, is the more marked word order. The special circumstances may be related to information structure, definiteness, pronominality, etcetera. Let me paraphrase Lenerz' conception of markedness as follows: On the assumption that we have identified the factors favouring a certain word order that are not related to grammatical function, canonical word order is the allowed word order that is not favoured by any of those factors (see also Dryer, 1995). That is, canonical argument order is the word order that does not have an 'excuse', the order that cannot be derived from other principles.

The prevailing opinion in the literature on German is that the unmarked order in the Mittelfeld is subject before object (S$\prec$O). However, when we look at the order of indirect and direct object there are different claims. Lenerz (1977) and Uszkoreit (1987) conclude that IO$\prec$DO is the unmarked order. In contrast, Müller (1999) takes DO$\prec$IO to be the *underlying* order in German, on the basis of word order data and binding data. Müller is however careful to distinguish underlying from unmarked word order, describing the latter in terms of typicality or acceptability of a surface word order.[20] Other researchers have proposed that different verbs carry different unmarked word orders for their arguments (Haider, 1993; also see Müller, 1999 for discussion and references). Support for claims that different verbs prefer different argument orders can be found in corpus based work (Kurz, 2000a; Kempen and Harbusch, 2004; Heylen, 2005). Some ditransitive verbs were found to prefer DO$\prec$IO, and others IO$\prec$DO. Dative-selecting transitives frequently allow O$\prec$S (breaking the S$\prec$O generalization).

In the Dutch Mittelfeld, S$\prec$O is not only default, but even near-categorical. For instance, German readily allows pronominal objects to precede non-pronominal subjects. However, this is rare in Dutch. Important exceptions are formed by dative experiencers (51a) and certain reflexive arguments (51b), which may precede their subjects (Den Besten, 1985; Shannon, 2000; Haeseryn et al., 1997).

(51) a. Eigenlijk is me de accomodatie niet zo goed bevallen.
Really is me the accomodation not so well pleased
'I was not happy with the accomodation, to be honest.'

b. Bij dergelijke zinnen doen zich de volgende gevallen voor: ...
With such sentences happen REFL the following cases VPART
'Amongst such sentences we can find the following cases: ...'

(both from Haeseryn et al., 1997)

The reference grammar Haeseryn et al. (1997) takes the default order of two nominal objects in the Mittelfeld to be IO$\prec$DO throughout, as do Jansen (1981) and Van der Beek (2005).

[20] In his paper Müller formulates a number of constraints on linear order that account for the difference between the underlying and surface word orders. We may therefore interpret Müller's results as meaning that DO$\prec$IO is canonical word order in German in our sense, since it is the word order without any excuses.

How should these findings about the Dutch and German Mittelfeld be transferred to the Vorfeld, if we the assume that canonical argument order is a global word order trend? Obviously, S$\prec$O in Dutch is not near-categorical when the Vorfeld is considered. We have already seen many examples in this chapter where one of the objects preceded the subject by occupying the Vorfeld. The directly observable relation between canonical argument order and Vorfeld occupation is not the same as the relation between canonical argument order and word order in the Mittelfeld. This does not mean that we have to conclude that canonical argument order is confined to the Mittelfeld, however. Canonical argument order was the argument order that surfaced when other influences on word order have been controlled for. These other influences may be different for the Mittelfeld than for the Vorfeld. The preference that the Vorfeld has for important material is an example of such an influence. Canonical argument order S$\prec$O for Dutch predicts that *ceteris paribus* subjects have a stronger tendency to appear in the Vorfeld than objects do. This would also explain why the subject is information structurally unrestricted as Vorfeld occupants, and objects are not. As a first indication of whether we can expect to find this tendency, we can look at Jansen's (1981) corpus study of Vorfeld occupation. Jansen indeed reports that subjects front more often than other arguments.

Applying the same reasoning to IO$\prec$DO, we predict that indirect objects show a stronger tendency to appear in the Vorfeld than direct objects do. There are results in the literature that suggest that this prediction will not be borne out. Thrift (2003) concludes from questionnaire data that putting an indirect object in the Vorfeld is only marginally possible in Dutch. Lamers (2001) presents experimental data comparing Vorfeld occupation and word order in the Mittelfeld. A questionnaire study of the understandability of different word orders revealed no difference in understandability of subject or direct object in the Vorfeld. Indirect objects in the Vorfeld were however slightly harder to understand. In the Mittelfeld, subject and indirect object initial word orders were easiest to understand, and direct object initial word order hardest. The studies of Thrift and Lamers therefore suggest that the prediction that having an indirect object in the Vorfeld is less marked than having a direct object in the Vorfeld will not be borne out.

I have until now used qualitative terms like 'unmarked word order' and 'has a stronger tendency to front'. These terms can receive a quantitative interpretation which can be used in a corpus investigation. The interpretation of canonical word order that I started out with can be related in a natural way to statistical analysis. The idea that of two word orders, canonical argument order is the one that can be encountered in most circumstances has a straightforward statistical interpretation. It means that if we *control* for other influences on word order in our statistics, the unmarked or canonical of two possible word orders should be the most frequent one.[21]

Canonical argument order gives rise to the following prediction about Vorfeld occupation, that is tested against the corpus in Chapter 4: Elements that are higher on the

grammatical function scale in (52) front more often after we correct for the influence of other factors.

(52) subject $\prec$ indirect object $\prec$ direct object

In the statistical models used to investigate the data in Section 4.5, canonical argument order is a factor treated on a par with other factors on Vorfeld occupation. I will now turn to the discussion of those other factors: definiteness and complexity.

2.6.2 Definiteness

The second factor in word order that is investigated is definiteness. The tendency to realize definite material early in the Mittelfeld, and indefinite material late in the Mittelfeld, has been observed for Dutch as well as for German. In the discussion of these effects below, I will make two assumptions. First, I will treat definiteness as a formal property, purely on the basis of surface form. This does not mean that I deny the possibility that there are forces underlying definiteness effects on word order, such as givenness/anaphoricity, referentiality, or intonation. However, treating definiteness as NP form allows us to investigate definiteness in our corpus. The corpus does not have annotation for any of the properties mentioned, but automatically categorizing NPs after form is fully feasible. Secondly, I will distinguish three main levels of definiteness: pronoun, definite full NP, and indefinite full NP. As we will see below, these three levels of definiteness have different effects on word order: Pronouns have a strong tendency of being realized early in the Mittelfeld, and indefinite NPs have a strong tendency of appearing later in the Mittelfeld. The behaviour of definite full NPs falls somewhere in between.

In the Dutch and German Mittelfeld, arguments can take part in two types of word order variation: *scrambling over argument*, referring to the reordering of arguments with respect to canonical word order, and *scrambling over adverb*, which refers to positioning an argument left of an adverb in the Mittelfeld. Definiteness influences word order in both types of scrambling.

In Dutch, scrambling over adverb is generally impossible with indefinite full NPs (but see exceptions below), optional with definite full NPs, and obligatory with pronouns (Haeseryn et al., 1997; de Hoop, 2003). The examples in (53) illustrate this partition. The adverb is in boldface.

[21] Lenerz (1977) himself opposes the idea that one can derive canonical word order from corpus frequencies. However, the examples that he uses to argue his point are examples where the corpus linguist fails to control for variables that the linguist using intuition data does control for. In my opinion, the conclusion of this should not be that canonical word order and quantitative analysis are incompatible, but that a linguist using quantitative analysis should not rely on simple counting alone.

(53) a. We moeten het / de was / *een voorstel **maar** doen.
we should it the laundry a proposal PART do
'We should do it / do the laundry / make a proposal' (De Hoop's 9–11)
b. We moeten **maar** *het / de was / een voorstel doen.

Van der Does and de Hoop (1998) and de Hoop (2003) also demonstrate that deaccented, anaphoric definite NPs scramble optionally, although they might have a stronger tendency to scramble than non-anaphoric definite NPs. Accented pronouns are also free to appear right of the adverb. Specific indefinite full NPs form an exception to the ban on scrambling of indefinite full NPs. They are allowed (but not forced) to scramble left of an adverb.[22]

Scrambling over argument does not appear to be influenced by the definite/indefinite full NP distinction in Dutch. Pronouns, however, show a clear preference for early realization in the Mittelfeld. For instance, the basic order IO$\prec$DO in a Dutch ditransitive is not respected when the direct object is a pronoun, and the indirect object a full NP (Van der Beek, 2005). The indirect object is in boldface.

(54) a. Kees reikt **Koos** ??'m / de kaasschaaf aan.
Kees hands Koos it$_{\text{MASC.RED}}$ / the cheese knife over
'Kees hands over the cheese knife/it to Koos.'
b. Kees reikt 'm / *de kaasschaaf **Koos** aan.

Apparently, in the Mittelfeld, the order pronoun$\prec$full NP is preferred over the canonical argument order IO$\prec$DO.[23]

To summarize, we see that in the Dutch Mittelfeld, pronouns tend to appear early, and indefinite full NPs appear late. These are tendencies, and not categorical statements, as we have seen that there are exceptions to the generalization.

Similar trends as those illustrated above for Dutch have been found for the German Mittelfeld. The main difference between Dutch and German is that the definiteness effects are more pronounced in German, which is presumably related to the fact that word order in the German Mittelfeld is freer than in the Dutch Mittelfeld. Pronominal arguments have a strong tendency to move leftward in the German Mittelfeld, and may scramble over subjects, even when accusative. Scrambling over adverb behaves as in Dutch. In addition, scrambling over argument shows a sensitivity to the definite-indefinite full NP distinction that is not apparent in Dutch (see Lenerz, 1977, Uszkoreit, 1987, Müller, 1999, for the contrast pronoun-full NP and the contrast definite-indefinite full

[22]On some definiteness scales in the literature, for instance Aissen (1999), specific indefinite full NPs are ranked above non-specific indefinite NPs. This difference in ranking between specific and non-specific indefinites is motivated by independent distributional properties.

[23]I will gloss over many intricate details of Dutch and German word order here. One interesting detail that I am ignoring is that when both IO and DO are reduced pronouns, German and Dutch order these DO$\prec$IO (Haeseryn et al., 1997; Müller, 1999). This, and other similar facts, does not contradict that there is a tendency to order pronouns before full NPs, and definite NPs before indefinite NPs, however.

NP; Büring, 2001, on definite-indefinite full NP; and Kempen and Harbusch, 2004 for corpus results).

The strategy I have taken when I formulated the expectations for the relation between grammatical function and Vorfeld occupation would lead to the following prediction about the relation between definiteness and Vorfeld occupation: Controlling for other factors, elements on the scale in (55) appear more often in the Vorfeld than elements to their right.

(55) pronoun $\prec$ definite full NP $\prec$ indefinite full NP.

As with canonical argument order, Vorfeld occupation need not be fully parallel to Mittelfeld word order. For instance, placing a full NP direct object before a pronominal subject in the Mittelfeld is ungrammatical in Dutch (56a). However, (56b) shows that placing a full NP direct object in the Vorfeld, while the pronominal subject is in the Mittelfeld, is fine (see Weber and Müller, 2004, for similar observations about German).

(56) a. *'s nachts houdt de buren ze wakker.
at night keeps the neighbours she awake
'She keeps the neighbours awake at night.'
b. de buren houdt ze 's nachts wakker.

We have two different possible reasons to expect that reduced personal pronouns do not like to appear in the Vorfeld: Gärtner and Steinbach's (2003) prosodic account, and the proposed status of the Vorfeld as a position for important material (Section 2.5). This contradicts with the predictions of the scale in (55). The corpus study will tell us whether the preference of the Vorfeld for important material is stronger than the global effect that definiteness has on word order. In any case, demonstrative pronouns should show a very strong tendency to front, since they are both important material and of the highest definiteness level.

The proposal that the Vorfeld prefers important material was based on a universal word order principle put forward by Gundel (1988): *first-things-first*. Gundel observes a conflict that is similar to the conflict between the Vorfeld's preference for important material and the global effect of the definiteness scale on word order sketched above. Gundel (1988) proposes two possibly conflicting universal word order principles, given in (57).

(57) a. *Given-before-new:*
State what is given before what is new in relation to it.
b. *First-things-first:*
Provide the most important information first.

The principles conflict when the given material in a sentence is not important. The first-things-first principle then prefers that the new material be provided first, since new

material is always important. The given-before-new principle disfavours this word order.[24] The conflict between first-things-first and the definiteness hierarchy is likely to make the relation between NP form and Vorfeld occupation in the corpus quite complex. The results from Jansen (1981) may give us an first impression of this relation. Jansen finds that pronouns front more often than other NPs, and that definite NPs front more often than indefinite NPs. This is compatible with the definiteness scale. In addition, Jansen reports that demonstrative pronouns front more often than personal pronouns, a difference predicted by the first-things-first principle.

Before I conclude this section about definiteness and Vorfeld occupation, I have to mention a factor in the distribution of subjects that may be unrelated to whatever underlies the definiteness hierarchy given above. In Section 2.2.1, I mentioned the Dutch existential construction (EC). A sentence with an indefinite logical subject (*geen beeld* 'no statue' in 58) is preferably not realized as a canonical sentence (58a), but as an EC (58b). In the EC in (58b), the expletive subject *er* is in the Vorfeld. Alternatively, other material like a locative PP can be placed in the Vorfeld, in which case the expletive *er* is optional (58c). Mikkelsen (2002, for Danish) and De Hoop and Krämer (2006) propose that the use of an EC is to repel a 'poor' subject from its canonical position. The optionality of *er* in (58c) suggests that in Dutch, a PP may also serve this purpose. The logical subject is in boldface.

(58) a. ??**geen beeld** stond in de tuin
no statue stood in the garden
'There was no statue in the garden.'[25]
b. er stond **geen beeld** in de tuin.
EXPL stood no statue in the garden.
c. in de tuin stond (er) geen beeld.[26]

The fact that (part of the) indefinite logical subjects will appear in a construction that moves them out of the Vorfeld means that we have an additional reason to expect to see a definiteness effect for subjects, where indefinite subjects are less likely to appear in the Vorfeld than definite or pronominal subjects. Note that the logical subject of an EC is not

[24]It may be tempting to propose that the definiteness scale is a grammaticalized version of the given-before-new principle. Pronouns are most likely to realize given referents, and indefinites typically introduce new referents. However, by *given* and *new* Gundel (1988) means something more like aboutness-topic and comment, although she does recognize the correlation between aboutness-topichood and discourse givenness.

[25]The sentence is acceptable under the partitive reading: None of the statues was in the garden.

[26]It may be that this is a locative inversion, even though in a Dutch main clause they would be hard to distinguish from any other PP topicalization. The matter is irrelevant here, however, because all example (c) is supposed to show is that the expletive *er* is optional when a locative PP is in the Vorfeld. Whether (c) without *er* is an EC, a locative inversion, or something else is besides the point.

prohibited from appearing in the Vorfeld. Example (58d) shows that focus topicalization may be used to put the logical indefinite subject of an EC 'back' into the Vorfeld.

(58) d. **geen BEELD** stond er in de tuin.

However, we expect focus topicalization to be relatively rare (Section 2.5.1), so that the overall effect of EC on the statistics will be that indefinite subjects are less likely to be realized in the Vorfeld.

In summary, we expect to find the definiteness scale pronouns $\prec$ definite full NPs $\prec$ indefinite full NPs reflected in the Vorfeld statistics, with the possible exception of personal pronouns. This prediction can be explained as arising from the interaction between two tendencies: a) the tendency to realize definite/given material early on in the sentence, b) the tendency to use the Vorfeld for important, unpredictable or highlighted information. The existential construction in Dutch may be an independent factor to negatively influence Vorfeld occupation by indefinite subjects.

2.6.3 Grammatical complexity

The classic statement about the effect of grammatical complexity on word order is due to Behaghel (1909). One of the word order laws formulated by Behaghel is the *Gesetz der Wachsenden Glieder* ('Law of increasing constituent size'), which states that when possible, shorter constituents precede longer ones. A similar principle can be found in Haeseryn et al. (1997) for Dutch. Their *Complexity Principle* states that more complex, heavier or longer constituents have a tendency to appear towards the end of a sentence. For instance, the sentence in (59) is ungrammatical with a demonstrative pronoun. A direct object is not allowed to appear to the right of the adverbs in the Mittelfeld when it is a demonstrative pronoun. This word order is allowed, however, with the more complex demonstrative full NP.

(59) Ik heb gisteren eindelijk *dat / dat artikel gelezen.
I have yesterday finally that that article read
'I finally read that / that article yesterday.' (judgement from Haeseryn et al., 1997)

Note that the predictions of the complexity principle for (59) overlap with the effect of pronominality on word order that we discussed above. Without making explicit reductionist claims, Haeseryn et al. (1997) suggest that other word order facts could also be understood in this fashion. For instance the order object-before-oblique in Dutch (not discussed here) could be reduced to the complexity principle, because NPs are generally shorter than PPs. See also the discussion of Hawkins (1994), below.

Empirical work on English has shown the influence of grammatical complexity in a large variety of constructions. In each case, the conclusion was that there is a preference to put more complex constituents more towards the end of a phrase or clause. For instance, in the postverbal domain in English, this effect has been found in *heavy NP shift*, particle-verb constructions, order of PP sentence modifiers, and the dative alternation (Hawkins, 1994; Wasow, 2002; Arnold et al., 2004; Bresnan et al., 2007, to name but a few). Bresnan et al. present the examples in (60). The contrast in grammaticality of (60a) and (60b) suggests that the ditransitive *give X the creeps* does not show dative alternation. Only the NP NP construction is allowed. However, in (60c), the NP PP realization is allowed. Bresnan et al. convincingly argue that the NP PP realization in (60c) is allowed because it lets the complex constituent *people whose* ... follow the simpler *the creeps*.

(60) a. That movie gave [me] [the creeps]
b. *That movie gave [the creeps] [to me].
c. Stories like these must give [the creeps] [to people whose idea of heaven is a world without religion]... (their 9–10)

The effects of grammatical complexity have also been found in the NP domain, for instance in the genitive alternation (Rosenbach, 2005). In each of these cases, the researchers have found that postponing complex constituents is only one of the tendencies in English word order – there are additional factors besides complexity that influence word order.

In influential work, Hawkins (1994) proposes that many of the claimed word order universals (given before new, definite before indefinite, pronominal before full NP) fall out of the complexity effect. Hawkins constructs an intricate processing account of complexity effects. Summarized very concisely, Hawkins posits that languages prefer to minimize the domain needed to identify all sub-constituents of a constituent during processing. This leads to a preference for putting the heads of sub-constituents as close to the head of the main constituent as possible. In the English postverbal domain, this preference translates into the preference to move longer constituents to the back. However, for a consistently head final language, the account would predict that long constituents are fronted – a prediction that appears to be borne out for Japanese (Hawkins, 1994; Yamashita, 2002). The claim that other word order trends can be reduced to a complexity effect, however, cannot be maintained given the work on English above, and Dutch and German to be discussed below. In more recent work, Hawkins (2004) allows other factors besides a complexity effect.

Existing work on the effect of complexity on word order in German and Dutch deal almost exclusively with Mittelfeld phenomena and the results are not as clear as the English results cited above. Hawkins (1994) argues for the effect of complexity in the German Mittelfeld, and for a complexity effect on relative clause extraposition in German

and Dutch. As mentioned, Haeseryn et al. (1997) assume a Complexity Principle for Dutch. Corpus evidence for the influence of grammatical complexity on word order in Dutch and German comes from Shannon (2000). Shannon shows that the word order pronominal object before subject is more common as the subjects become longer. Heylen (2005) presents a corpus study showing that constituent length has an effect on certain cases of German scrambling.

In contrast, other corpus studies have found no evidence for a tendency to place complex constituents at the end in the German Mittelfeld (Fanselow, 2000) or have found only a very weak preference dominated by, for instance, definiteness effects (Kurz, 2000b). Unlike the work on the English dative alternation, Van der Beek (2005) concludes that there is no length effect on the order of direct object and indirect object in Dutch.

To the conflicting results in the literature on the effect of complexity on word order in the postverbal domain in German, Dutch and English, we can add uncertainty about what the effect of complexity on topicalization should be. For instance, Birner and Ward (1998, pp24ff) consider topicalization in English to be pragmatically driven, and suggest that complexity plays a minor role at best. Hawkins (2004) allows for the interaction of complexity with other (grammaticalized) performance considerations that favour for instance putting a topic in sentence initial position. However, it is not clear what the effect of complexity on Vorfeld positioning in Dutch should be in Hawkins' model.

There is only little empirical work on Dutch and German Vorfeld positioning that one could base a hypothesis on. Jansen and Wijnands (2004) present results of a small corpus study of Dutch newspaper articles. They find, amongst other things, that PPs and CPs avoid the Mittelfeld and occur either in the Vorfeld or the Nachfeld; that NPs that contain a relative clause appear mostly in the Nachfeld; and that NPs that contain PP modifiers appear as often in the Vorfeld as NPs without such modifiers. The examples in (61) illustrate the tendency of a CP (in boldface) to avoid the Mittelfeld (61b), but not the Vorfeld (61a) or Nachfeld (61c).

(61) a. **Dat het hoe dan ook gênant wordt,** is al vaak gebleken.
That it anyway embarrassing becomes is often proven
'It's been proven many times before that everything ends in embarrassment.'
b. ?Al vaak is **dat het hoe dan ook gênant wordt** gebleken.
c. Al vaak is gebleken **dat het hoe dan ook gênant wordt**.

The behaviour illustrated in (61) suggest that Vorfeld occupants may be more complex than Mittelfeld constituents. This contradicts the general Complexity Principle of Haeseryn et al. (1997) that would predict that Mittelfeld occupants are more complex than Vorfeld occupants because the Mittelfeld comes after the Vorfeld.

Looking at constituent length, rather than category or internal make-up, Jansen and Wijnands find that there is no significant difference in constituent length anywhere in

the sentence (Vorfeld, Mittelfeld), except for the last constituent. The last constituent is on average longer than constituents anywhere else. This would indicate that, although there is a strong link between complexity and right-peripheral placement (extraposition to the Nachfeld), there is no opposite correlation, that is, between the Vorfeld and lighter constituents. The conclusion would have to be that simple before complex is not a global word order trend, but that it is an effect directly related to the right periphery – just like the preference for important material is directly related to the Vorfeld.

Nevertheless, I have based the predictions that I test in Chapter 4 on the Complexity Principle of Haeseryn et al. (1997). That is, we predict that longer constituents are less likely to appear in the Vorfeld. In Section 4.4, I will, however, also consider the postverbal domain in more detail, as the results of Jansen and Wijnands suggest that complexity effects on word order in Dutch are restricted to an effect at the right periphery. The corpus also allows us to investigate to what extent trends in definiteness and argument order on the one hand, and length effects on the other are mutually explainable.

2.7 Conclusion

In this chapter I have introduced some aspects of word order in the Dutch main clause. I have described the direct preverbal position referred to as the Vorfeld in terms of form and function. I have also demarcated the area of investigation by contrasting Vorfeld occupation and topicalization with other constructions that involve the left periphery, like topic drop, left dislocation, hanging topic, the mirror construction, and the phenomenon of multiple Vorfeld occupation.

I have also presented predictions about Vorfeld occupation in a corpus. These predictions were generated from existing results on Mittelfeld word order in Dutch and German, and postverbal word order in English, on the hypothesis that global word order trends underly these domains and the Vorfeld alike. The relation of these global trends and Vorfeld occupation may be complicated by the status of the Vorfeld as a position for important material.

1. Canonical argument order In the Dutch Mittelfeld, canonical word order has been argued to be subject $\prec$ indirect object $\prec$ direct object. On the basis of this order, we expect that subjects have the highest chance of appearing in the Vorfeld, followed by indirect objects. Direct objects should have the lowest chance of occupying the Vorfeld.

2. Definiteness Scrambling over argument and adverb in the Mittelfeld in Dutch and German can be related to a scale pronoun $\prec$ definite full NPs $\prec$ indefinite full NPs. Elements on the scale show a stronger tendency to scramble than elements to their right. This leads to the prediction that pronouns have the highest chance of appearing in the

Vorfeld, and indefinite NPs the lowest. The relation between definiteness (NP form) and Vorfeld occupation is also influenced by the first-things-first principle. Because Vorfeld material is preferably important, (reduced) personal pronouns are not good Vorfeld occupants. We thus make conflicting predictions with respect to Vorfeld occupation by personal pronouns.

3. Grammatical complexity It has been claimed that there is a general preference to order light and simple constituents before heavy or complex material. This trend has not been clearly established for Dutch, but there is solid evidence that complexity plays a role in the English postverbal domain. The prediction that will be investigated is that less complex constituents have a higher chance of appearing in the Vorfeld.

These predictions are independent of each other; that is, we predict to see each of the effects above after controlling for the other two factors.

In addition to the results of investigating the predictions above, I will also present results on the relation between the presence of sentence adverbials and direct-object–initial word order.

I will introduce the spoken Dutch corpus in Chapter 3. This chapter also contains a discussion of the methods and tools used in the investigation. A portion of the chapter is dedicated to the definition of the Vorfeld in terms of the corpus, which does not contain direct annotation for topological fields. The results of the corpus investigation itself are presented in Chapter 4.

Chapter 3

Methods, Techniques & Material

Before proceeding to the corpus study in the next chapter, I will discuss some background issues in this chapter. I will begin by introducing the *Corpus Gesproken Nederlands*, used for the investigations in this dissertation. I will briefly describe its make-up, the type of annotations, and the amount of available material in Section 3.1. Section 3.2 introduces the syntactic annotation available for the corpus. I will discuss general properties of the annotation and the way in which the annotation may influence what we can learn from the corpus. Moreover, I will briefly consider how much of the information that we need for the investigations we have planned is directly retrievable from the corpus, and how much must be added to it. The discussion of the syntactic annotation also prepares us for Section 3.3. In this section, I will give a definition of the Vorfeld in terms of the syntactic annotation. Topological fields are not annotated directly in the Corpus Gesproken Nederlands, but the amount of syntactic annotation that is present allows us to give a good definition of the Vorfeld.

Finally, the last two sections introduce two important tools. Section 3.4 explains the use of the logic programming language Prolog for the purpose of querying the corpus. The section ends with a Prolog translation of the Corpus Gesproken Nederlands definition of Vorfeld. Section 3.5 provides a general introduction to the use of *logistic regression modelling*. This statistical technique is used to answer several important questions in the remainder of the dissertation.

3.1 About the *Corpus Gesproken Nederlands*

The *Corpus Gesproken Nederlands* (Spoken Dutch Corpus, CGN, 2004), is a corpus of approximately 9mln words, that contains spoken Dutch from adult native speakers in The

Table 3.1: Overview of available syntactically annotated material.

Component		Region	
		nl	fl
a	Spontaneous conversation	300 368	146 745
b	Interviews with Dutch teachers	25 687	34 064
c	Spontaneous telephone dialogues	69 933	19 886
d	Spontaneous telephone dialogues	0	6 257
f	Interviews/discussions/debates (bc)	75 106	25 144
g	(Political) Discussions/debates/meetings	25 117	9 009
h	Classroom recordings	25 961	10 103
i	Live commentaries (bc)	24 986	10 130
j	News reports/background (bc)	25 065	7 679
k	News (bc)	25 384	7 305
l	Commentaries/columns/reviews (bc)	25 071	7 431
m	Ceremonious speeches/sermons	5 184	1 893
n	Lectures/Seminars	14 913	8 143
	Total Used	642 775	293 789
e	Simulated business negotiations	25 485	0
o	Read speech	0	44 144
	Grand Total	668 260	337 933

Note: Counts are in words. Abbreviations nl: Netherlands; fl: Flanders; bc: broadcast material. Segments c and d differ only in recording method.
Source: CGN (2004)

Netherlands and Flanders. The complete corpus has received orthographic transcription, part-of-speech (POS) tagging, lemmatization and automatic phonetic transcription and alignment. About a third of the data is from Flanders. Meta-data about the background of the speakers and the situational context are also provided.

Parts of the corpus have received additional annotation. In this study, we are interested in lexical, syntactic and information structural properties of Vorfeld placement. However, there is no annotation for information structure. The constituent properties whose relation to the Vorfeld I will be investigating – grammatical function, definiteness/NP form and grammatical complexity – can be read from the combined morphological and syntactic annotation layers. There are about 1mln words of data available with this information.

The corpus is divided in different components that correspond to different sources and/or genres. The amount of material with morphological and syntactic annotation that is available per component is shown in Table 3.1. The table also shows how much data

in each component comes from which region. Two components are excluded from the selection: *Component e* was excluded because there was only material available from the Netherlands, and *component o* was excluded because it contains read speech from literary novels – data that we do not expect to be representative of spoken language. The remaining components are by no means homogeneous. They contain speech ranging from spontaneous to prepared, and speech produced in monologues or dialogues. The mixed nature of the data is an advantage because it allows us to better approximate the full range of spoken Dutch. Also, excluding data should be avoided as much as possible, because data sparseness may be a real problem. Note that *components c* and *d* are of the same genre but differ in recording method, so there is no reason to throw *component d* out.

It might be the case that there are subtle differences in Vorfeld placement between regions, registers, and genres. I fully expect that looking for regional variation in Vorfeld occupation in the corpus, and likewise looking for differences between the registers/genres, would prove to be interesting. However, in this dissertation I will not consider these issues, and treat all parts of the corpus just described on a par.

In the annotation, speech is divided into *utterances*. The selected corpus (*components a–d* and *f–n*) has ∼125k utterances, averaging 7.6 words per utterance. It should be kept in mind that many of the utterances, especially in the spontaneous speech components, are one or two word utterances, of the sort: 'uhm' 'uh uhm', etcetera.

The CGN has receive detailed syntactic annotation but there is no annotation for topological fields. An important task in this chapter is therefore to provide a definition of Vorfeld in CGN terms. Section 3.3 gives this definition in prose, and Section 3.4 discusses the implementation of this definition. In the next section, I will begin by introducing the syntactic annotation.

3.2 Syntactic annotation in the CGN

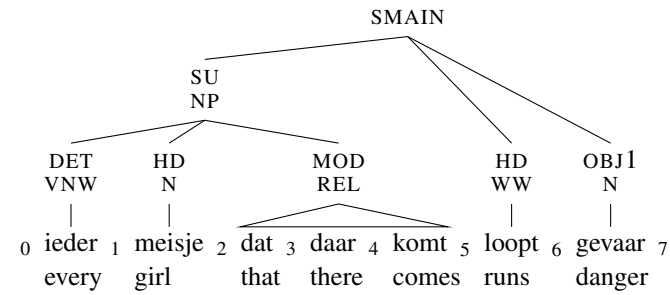
The syntactic annotation scheme employed in the CGN is a hybrid between two approaches to syntactic description. First and foremost, the syntactic structures in the CGN are dependency structures. However, CGN trees also contain phrasal nodes, which are not traditionally part of dependency structure, but come from a phrase structure approach to describing syntactic structure. In this section I will describe this hybrid structure, and discuss how the structure influences the kind of information we can extract from the corpus and the way in which we can do so (Subsection 3.2.1). In Subsection 3.2.2, I will briefly consider the special problem of multiple dependencies. In many cases one could argue that a complement is dependent on more than one head. However, annotating multiple dependencies leads to a less constrained data structure which is harder to use in corpus work. Fortunately, we will see that we can ignore multiple dependencies in our definitions and queries.

3.2.1 Dependencies and phrases

The CGN syntactic annotation mixes dependency structure and phrase structure. In a dependency structure, syntactic structure is specified in terms of grammatical relations between heads and complements. Grammatical relations are primary in a dependency structure. Traditionally, all nodes in a dependency tree are words. A syntactic structure is then a graph that connects all word in a sentence. In a CGN tree, however, nodes can be phrases, too. A CGN dependency relation is annotated between a mother phrase and a daughter constituent (a phrase or word). A special type of CGN dependency relation is the *head* relation (HD). Replacing phrasal nodes in a CGN tree with the word that is the head-daughter would result in a more traditional tree. Note that CGN does not require phrases to have heads, nor does it require them to be unique.

The hybrid of dependency structure and phrase structure allows for an effective division of labour: Formal information like the type of a clause or the category of a word can be annotated in the phrase labels, and grammatical functions can be annotated as dependency relations. The mixed scheme was originally put forward for German in Skut et al. (1997). The annotation guidelines for the CGN are given in Hoekstra et al. (2003, in Dutch). Guidelines for morphological analysis in the CGN can be found in Van Eynde (2003, in Dutch). English reports of the morphology- and syntax-annotation efforts can be found in Van Eynde, Zavrel, and Daelemans (2000) and Hoekstra et al. (2001), respectively. The annotation format is also used in the Alpino Treebank (Dutch, with a near identical annotation protocol, Van der Beek et al., 2002), and in the TiGer Treebank (German, Brants et al., 2002). An example CGN syntactic tree is given in (1).

(1)



'Every girl that goes there is in danger.'

The whole sentence in (1) is marked up as a V2 main clause (SMAIN). Nodes in the tree below this top-level have two labels: the dependency relation between the node and its mother, and the syntactic category (phrase type or part-of-speech) of the node. The sentence in (1) has an NP subject ($^{\text{SU}}_{\text{NP}}$), the finite verb in second position as its head ($^{\text{HD}}_{\text{WW}}$,

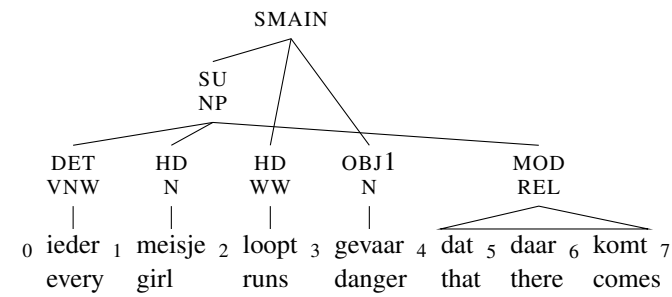
WW for *werkwoord* ‘verb’), and a noun direct object ($^{\text{OBJ1}}_{\text{N}}$). The subject NP has three daughters: a pronominal determiner ($^{\text{DET}}_{\text{VNW}}$, VNW for *voornaamwoord* ‘pronoun’), a head noun, and a modifying relative clause ($^{\text{MOD}}_{\text{REL}}$). The internal structure of the relative clause is not shown. A complete list of dependency relations and syntactic categories is given in Appendix A.

CGN syntactic trees do not encode linear order. The nodes in the dependency tree in (1) are not ordered with respect to each other. This is potentially problematic. Suppose we want to know whether the subject in (1) precedes the finite verb. We would like to be able to decide this by looking at the node that corresponds to the subject $^{\text{SU}}_{\text{NP}}$ and at the node that corresponds to the finite verb $^{\text{HD}}_{\text{WW}}$. However, since these are not ordered, we cannot say whether the former precedes the latter. In our corpus research, it is crucial that we are able to answer such a question. After all, one of the characterizing properties of Vorfeld material is that it precedes the finite verb that marks the left bracket in the topological fields template (Section 2.1).

As a rather straightforward solution to this, used for instance in the Alpino Treebank (Van der Beek et al., 2002), we can use the linear order of the words that are dominated by two nodes to determine the linear order of the nodes. A node begins where the leftmost word it dominates begins, and ends where the rightmost word it dominates ends. Under this definition of linear order of a node, the subject NP of (1) starts at 0 and ends at 5, and the head verb node begins at 5 and ends at 6. To answer our earlier question: The subject of (1) precedes the finite verb.

The fact that dependency structure does not encode linear order itself is convenient when one studies word order variation: If dependency structure is independent of linear order, word order variation does not change the syntactic structure of a sentence. Consider a word order variant of (1) where the relative clause that modifies the subject is in the postverbal domain. The result is (2).

(2)



‘Every girl that comes there is in danger.’

This constancy in syntactic analysis reflects a basic, defining assumption about word order variation, which is that there is a common core (of propositional content, thematic

role assignment, grammatical function assignment, etcetera) between word order variants.

The tree in (2) also shows another salient property of CGN trees: they may contain discontinuous phrases. In a discontinuous phrase, descendants are separated by material that is not descendant of that phrase. In (2), the head of the subject is separated from its sister relative clause by the finite verb and the direct object. Since the trees do not encode linear order directly, this discontinuity cannot be seen in the syntactic structure itself but depends on the linear order of the dominated words. A result of discontinuous phrases is that, given two nodes, we cannot always decide which one comes first. In (2), the subject NP ends at 7 and the finite verb starts at 2, so the subject does not precede the finite verb. However, since the finite verb ends at 3 and the subject begins at 0, the finite verb does not precede the subject either. Without further assumptions, we cannot consider the subject to be either preverbal or postverbal. This issue becomes relevant when we want to know whether the subject in (2) is in the Vorfeld or not. I will therefore return to this issue when I define Vorfeld occupancy in CGN terms in Section 3.3.

Now that we have seen some general properties of the CGN syntactic annotation, let us briefly consider what information is directly available in the annotation and what information we will have to add (automatically) in light of the research intentions stated in Chapter 2. In Section 2.6, I discussed three constituent properties whose relation to Vorfeld occupation will be investigated; grammatical function, definiteness or NP form, and grammatical complexity. I will give the details of the operationalization of each of these properties in terms of the corpus annotation in Chapter 4, when I discuss the results of the corpus investigations.

Grammatical function The first property we are interested in is grammatical function. We need to know the grammatical function of a constituent because we will only look at the behaviour of subjects, indirect objects and direct objects. Moreover, one of the questions to be investigated in the corpus that was raised in Section 2.6 was how grammatical function relates to Vorfeld occupation.

Grammatical function can be read fairly directly from the dependency relations: Subjects are SU-dependents, direct objects OBJ1-dependents and indirect objects OBJ2-dependents. However, there are two caveats. First, a minor point is that the OBJ1-relation is also used for objects of a preposition, and not just direct objects of a verb. Secondly, direct objects and indirect objects in a clause may be embedded under verbs without affecting their (intuitive) markedness as Vorfeld occupants. This is illustrated by (3). The object initial word order in (3a) is no less or more marked than (3b).

		OBJ1	HD	SU
(3)	a.	**Die boot**	zag	ik.
		that boot	saw	I
		'I saw that boat.'		

	OBJ1 VC VC	HD	SU	HD VC	HD VC VC
b.	**Die boot**	had	ik	willen	zien.
	That boat	had	I	want	see

'I had wanted to see that boat.'

As an indication of the syntactic structure I have given the *dependency path* above each word. A dependency path is the list of dependency relations that one crosses if one walks from the word up the tree to the SMAIN-node (Van der Beek et al., 2002). The dependency paths show how the initial direct object in (3a) is an immediate OBJ1 of SMAIN, whereas the direct object in (3b) is a dependent of a phrase embedded under two verbs (VC for verbal complement). It is tempting to just allow any OBJ1 descendant of an SMAIN at any level to count as a direct object in our investigation, since the embedding does not appear to make a difference for its fronting behaviour. However, this is too unrestricted. A direct object from an embedded clause is marked in the Vorfeld of the matrix clause (4).

	OBJ1 VC BODY VC	HD	SU	CMP VC	BODY VC		
(4)	?**Die boot**	hoopte	ik	dat	[ik	zou	zien].
	That boat	hoped	I	that	I	would	see

'I had hoped to see that boat.'

The dependency path of the Vorfeld occupant contains a BODY-relation, which (in this case) indicates that the object is extracted from an embedded clause. A way around this problem is to define *direct object* as a collection of dependency paths beginning with OBJ1 that do not differ in their fronting behaviour. The question of which paths can be pooled together is thus in part an empirical question, to be answered in Chapter 4. The same reasoning carries over to indirect objects and OBJ2. A positive side effect of defining grammatical relation in terms of dependency paths is that the problem that OBJ1 is also used for prepositional objects disappears – prepositional objects have different dependency paths.

Definiteness The second property of constituents that I will investigate is definiteness. The CGN syntactic annotation does not include an annotation for definiteness, so we will have to add this information ourselves. For our investigations, we have to be able to distinguish at least the following: the three definiteness levels (pronoun, definite full NP and indefinite full NP), and pronominal form (demonstrative, full or reduced). These distinctions can be fairly easily added using lists of distinguishing features. The pronominal forms can be categorized simply by listing all possible forms. The full NPs can be categorized into definite or indefinite mostly by looking at their determiners. More details of the classification can be found Section 4.3.1.

Grammatical complexity Finally, I will look at the relation between a constituent's grammatical complexity and Vorfeld occupation. One measure of grammatical complexity

is syntactic category (NP, PP, CP, etcetera). This can be read almost directly from the syntactic annotation. In the literature on the effects of grammatical complexity, the length of a constituent is also frequently used as a measure of its complexity (see Section 4.4). The length of a constituent or node is the number of words that are dominated by it – retrieving this information from the corpus is unproblematic.

The CGN annotation allows us to gather all the information about constituents that we are after. This information can either be read almost directly from the annotation (grammatical function, grammatical complexity) or we can expect to reliably add this information automatically on the basis of the existing annotation (NP form). That this process can be automated is important because it allows us to use the whole of the CGN without having to worry about the time it takes to manually annotate for the information.

3.2.2 Multiple Dependencies

CGN dependency trees are unlike context free phrase structure trees in that linear order is not encoded in the tree. However, dependency annotation allows for a more serious deviation from the tree as datastructure: Nodes may depend on multiple mothers. Multiple dependencies mean that the datastructure of a rooted tree does not suffice anymore, and that the more general *directed acyclic graphs* would have to be used to represent the corpus data. Although this in itself is not problematic, searching through trees and inferencing on trees is simpler than working with directed acyclic graphs. Below, I will discuss two situations in which one might want to annotate multiple dependencies. In the first case, the multiple dependencies are (supposedly) *lexically derivable* from one dependency – these are not annotated in the CGN. In the second case, the multiple dependencies are *not lexically derivable* – these are annotated by the CGN but do not contain any information that we need to rely on. In conclusion, we can ignore the problem of multiple dependencies and assume that the syntactic annotations are trees in our definitions.

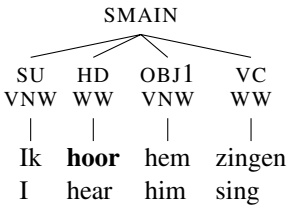
Lexically derivable multiple dependencies In cases of raising, control (equi), AcI, etcetera, one may argue that a constituent is a dependent of more than one node. For instance, in the AcI in (5a), the pronoun *hem* 'him' is an object of *horen* 'to hear' and in some sense also the subject of *zingen* 'to sing'. We might annotate this with a dependency OBJ1 from *hem* to the SMAIN, and a dependency SU from *hem* to the phrase headed by *zingen*. In the control-construction in (5b), a relation SU from *ik* 'I' to the SMAIN as well as a relation SU from *ik* to the phrase headed by *horen* 'to hear' is reasonable.

(5) a. Ik hoor hem zingen.
I hear him sing
'I can hear him sing.'

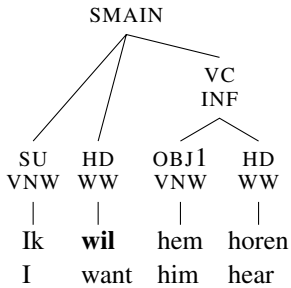
b. Ik wil hem horen.
I want him hear
'I want to hear him.'

The CGN only annotates the highest dependency in the tree. According to the annotation guidelines, the structure of the sentences in (5) is (6).

(6) a.



b.



Additional dependencies may be derived from a) the highest dependency, the one that is annotated, and b) information about the matrix verb. For instance, we know about the control verb *willen* 'to want' that its subject is also the subject of the infinitival VP *willen* takes as an argument.[1]

The question is whether we need the information that would be contained in the extra dependency links that the CGN has chosen not to annotate. For the investigations in this dissertation, we do not need the extra dependencies. For instance, the ungrammaticality of the reduced pronoun direct object *'m* in the Vorfeld in (7) shows that the fact that this direct object is also the subject of the embedded verb *horen* does not bring it the same privileges as being a subject of the finite verb would bring (see Section 2.3 for a discussion of reduced pronouns and Vorfeld occupation).

(7) *'m Hoor ik zingen.
him.RED hear I sing
'I can hear him sing.'

[1] It is not clear whether it is always the case that the matrix verb provides enough information to construct additional dependencies. For English, Pollard and Sag (1994, sec 7.4) point out that in cases like (i), it depends on the lower verb which of the arguments of the matrix verb is shared. In (ia), Kim is attending the party, whereas in (ib), Sandy is allowed to attend the party.

(i) a. Kim promised Sandy to attend the party.
b. Kim promised Sandy to be allowed to attend the party.

Investigating whether, and if so how often, this occurs in spoken Dutch is not the topic of this dissertation, so I will not consider this issue any further. Here, I would just like to point out that considering these extra dependencies as lexically derivable is a slight simplification.

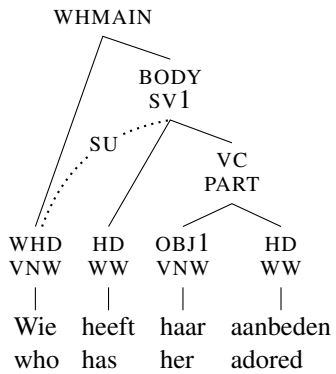
Apparently, only the dependency relation between *'m* and *hoor* is relevant in deciding whether *'m* can front. The other, currently not annotated, dependency relation would only add confusion, albeit easily resolved.

A possible downside of the omission of the additional dependencies is that an argument is not always a dependent of its main verb in the annotation. In Section 2.6, I mentioned that some researchers of German word order have claimed that different verbs prefer different base orders of their arguments. In order to investigate this in a corpus, we would have to know of each constituent what its main verb is. However, if the main verb is not the finite verb, a constituent may not be a dependent of its main verb in the CGN annotation. Adding the additional information of what the main verb is, and which dependency relation a constituent has to that verb, may be possible to do automatically but it adds another non-trivial step to the investigation.

Non-lexically derivable multiple dependencies There is a second group of constructions in which multiple dependencies may be called for. In this second group, the additional dependencies can generally not be derived from other parts of the annotation. These cases are annotated by the CGN. An example will follow below.

The CGN annotation formally distinguishes primary dependencies from secondary ones (following Skut et al., 1997). The primary dependencies form a tree. The trees that we have seen until now consisted of only primary dependencies. When there is a need for extra dependencies, these can be added as decorations to this tree (secondary edges). Secondary edges are employed for instance in the annotation of the syntactic structure of questions (8). The secondary edge is drawn as a dotted curve. The phrase labels WHMAIN and SV1 refer to main wh-question and verb-initial clause. The relations WHD and BODY hold between a wh-question and its wh-constituent and a wh-question and the non-wh-part, respectively.

(8)



'Who has adored her.'

CGN annotates the wh-word/phrase as the head of a wh-question. At the same time, this word may also be an argument in the clause. This argument relation is encoded with the secondary edge. In (8), the wh-word is a subject.

The corpus work in this thesis is not concerned with questions, so the extra annotation can be safely ignored. Other constructions that are annotated with secondary dependencies are (free) relative clauses and coordinations. Since none of these constructions is directly relevant for the Vorfeld in declarative sentences, we can ignore all annotation that uses the secondary dependencies. When we only look at primary dependencies, the annotations will always we trees. Therefore, we can safely assume in our definitions and queries that all dependency structures are trees. We can now turn to the definition that will be central to all corpus work in this thesis: the definition of Vorfeld in terms of the CGN annotation.

3.3 Finding the Vorfeld in CGN

The CGN syntactic annotation is detailed enough for us to retrieve information about a constituent's grammatical function, definiteness and length. What remains to be given, however, is a definition of the Vorfeld in terms of the CGN annotation. In this section I will develop that definition.

The Vorfeld is a region in a main clause. What we are after in this section is a definition of whether a constituent is in this region. We can look at this as a relation between a constituent and the main clause that the constituent is in. This relation is separate from grammatical function, and it is not restricted to certain syntactic categories. We have seen in Chapter 2 that nearly anything could appear in the Vorfeld. As a consequence, characteristics such as syntactic function or syntactic category are not very helpful when we try to decide whether a constituent is in the Vorfeld. On identifying topological fields in a corpus that does not annotate for them, Meurers (2005) writes the following.

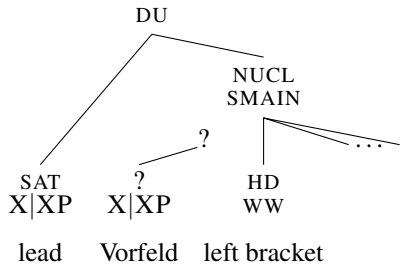
> Searching for material in fields with less characteristic membership [than the *verbal cluster* – gjb], such as the fronted material in the *Vorfeld*, the freely ordered mixture of elements in the *Mittelfeld*, or extraposed material in the *Nachfeld*, is practically impossible in a corpus without topological or structural annotation. Meurers (2005, p1631)

As we will see below, in spite of the fact that the CGN does not have topological annotation, the structural annotation that is present is enough to define the topological Vorfeld on.

Let us start by briefly repeating what the Vorfeld is from the introduction in Section 2.1. In a main clause in Dutch and German, we can define three linearly ordered regions – *Vorfeld*, *Mittelfeld*, and *Nachfeld* – that are separated by verbal material. In the left

periphery, the finite verb in main clauses marks the right boundary of the Vorfeld and is referred to as the *left bracket*. In addition to the Vorfeld and left bracket, the left periphery of a Dutch utterance may contain extra material that precedes the Vorfeld. This material is in a region known as the *lead*. The relation between the CGN syntactic tree and the topological fields in the left periphery is schematically represented in (9).

(9)



Phenomena like contrastive left dislocation and hanging topics (see Section 2.4.2), which involve material in the lead, are annotated as super-sentential or discourse phenomena. Therefore the whole utterance is marked as a *discourse unit* (DU). In this unit, the main clause (SMAIN) is the nucleus (NUCL), and the left dislocated material the satellite (SAT). Material in the lead may be words or phrases of various types (for the CGN), hence 'X|XP' as indication of the category for theSAT-dependent. The Vorfeld is however part of the SMAIN, so we know that it must be a descendant of the SMAIN node in the tree. What we do not know is at what level below the SMAIN-node the Vorfeld constituent occurs, what dependency relation the Vorfeld constituent will have, or of which syntactic type the Vorfeld constituent will be. Finally, the left bracket is the finite verb in the main clause: It is the head of the clause (HD daughter) and it is a verb (WW).

We can distill two properties that a constituent in the Vorfeld must meet from this description of the CGN annotation: A Vorfeld constituent is a descendant (of any generation) of an SMAIN node, and a Vorfeld must precede the HD of the SMAIN that it is a descendant of. Let us apply this definition to the example in (10). De finite verb is boldfaced.

(10) [SMAIN Jan des Bouvrie's achternaam **congrueert** niet in getal].
Jan des Bouvrie's surname agrees not in number
'There is something odd about JdB's last name.'

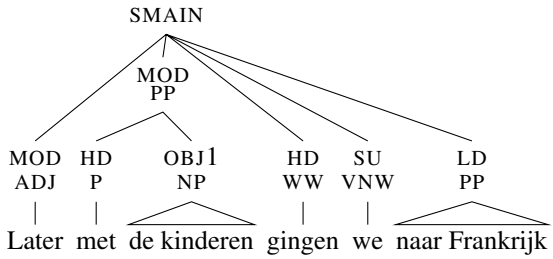
All of *Jan*, *des*, *Bouvrie*, *achternaam*, *Jan des Bouvrie's* and *Jan des Bouvrie's achternaam* are descendants of the SMAIN that precede the finite verb. However, we are only interested in one of them; the largest constituent *Jan des Bouvrie's achternaam*. Let us dub this the Vorfeld occupant. We now need a way to separate Vorfeld occupants from other Vorfeld constituents.

It is not correct to say that Vorfeld occupants are Vorfeld constituents that span the entire string between the beginning of the sentence and the finite verb. There are cases that have multiple Vorfeld occupants ('|' marks the boundary between the occupants), such as (11).[2]

(11) Later | met de kinderen **gingen** we naar Frankrijk.
later with the children went we to France
'Later with the children, we went to France.'

I will define Vorfeld occupant as a constituent in the Vorfeld that is not dominated by another constituent in the Vorfeld. This deals with both (10) and (11) in the desired fashion. For the latter, this is shown in the syntactic structure in (12).

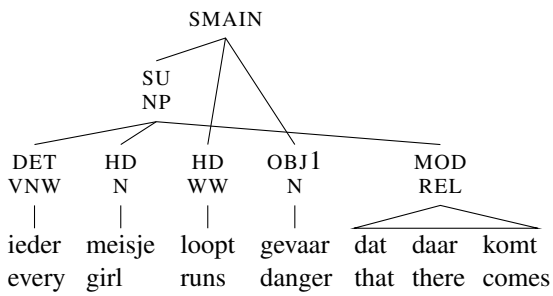
(12)



Of all the constituents in (12) that precede the head of SMAIN, [$_{\text{ADJ}}$ *later*] and [$_{\text{PP}}$ *met de kinderen*] are not dominated by other constituents also preceding the head. These two constituents are therefore Vorfeld occupants.

There is one final issue we need to address, which relates to discontinuous constituents. Consider example (2), repeated here as (13). The subject in the Vorfeld has an extraposed relative clause.

(13)



'Every girl that comes there is in danger.'

[2]In the CGN multiple Vorfeld occupancy is possible and our definition thus has to be able to deal with this situation. See Section 2.4.3 for discussion of multiple Vorfeld occupancy.

According to the definition thus far, the subject of (13) is not a Vorfeld occupant. The subject does not precede the head of SMAIN because the rightmost word under the subject node, *komt* 'comes', is not before the head of SMAIN but behind it in sentence final position. This has the unfortunate side effect that *ieder* and *meisje* do count as Vorfeld occupants. After all, they are Vorfeld constituents that are not dominated by another Vorfeld constituent.

There are several constructions that involve discontinuous phrases that start in the Vorfeld and end in the postverbal domain, of which we would like to say that the dependent is the Vorfeld. The problem occurs with stranded or extraposed modifiers (13), coordination, and certain discourse-level dependencies. The problematic constituents have in common that some key descendant like the head noun (in the case of a discontinuous NP), or the first conjunct (in cases of coordination) is properly in the Vorfeld.

We will solve this issue by loosening the precedence requirement. Instead of requiring that the complete phrase precedes the head of SMAIN, we will require that at least one of its key dependents does. Since the head noun is a key dependent of the NP in (13), we allow the NP to be a Vorfeld constituent, and thereby a Vorfeld occupant. The determiner and head noun *ieder* and *meisje* are no longer Vorfeld occupants under this definition. Note that we want to be restrictive in what dependent we allow as *pars pro toto*. For instance, an OBJ1 should not fulfill this role: having an object in the Vorfeld does justify classifying the whole VP as a Vorfeld constituent.

The topological notion of Vorfeld can be successfully defined in CGN terms. The definition can be summarized as in (14).

(14) A constituent V is a *Vorfeld occupant* when
  a. V is a descendant of an SMAIN clause S
  b. and V or a key direct dependent of V precedes the head of S
  c. and there is no constituent that meets criteria (a) and (b) of which V is a descendant

The selection of a corpus search tool, and the implementation of definition (14) in terms that fit this tool is the subject of the next section.

3.4 Implementation

The corpus searches in this dissertation will be conducted using the programming language Prolog.[3] This approach was chosen over using existing, dedicated corpus searching

[3] I used SICStus Prolog v3 (SICS, 2005), although nothing relies on this.

tools suited for the CGN like TiGer Search (TiGer, 2003) and DT_Search (Bouma and Kloosterman, 2002) because of a lack of flexibility in the dedicated tools.

DT_Search, developed in the context of the Alpino treebank (Van der Beek et al., 2002), uses the XPath query language to search through an XML representation of the corpus. Although DT_Search has been successfully used in CGN-related corpus studies before (for instance, Bouma, 2004, Van der Beek, 2005), (linguistically) intuitive formulation of more complicated queries is difficult because XPath does not offer a straightforward way of naming specific nodes for future reference, nor does it allow one to easily construct macros of queries. This means that complex queries cannot be easily composed out of less complex ones. Recently, Bouma and Kloosterman (2007) have proposed to use the more powerful XQuery language to solve these issues. An additional problem is that although DT_Search is good at retrieving fragments from the corpus on the basis of a query, it does not provide many facilities to analyze the data after retrieval. For instance, in Section 3.2.1, I explained that information about definiteness of an NP will have to be inferred because it is not part of the annotation of the corpus. This is the kind of classification that one would like to be able to do in a post-processing stage. Such data manipulation and analysis requires separate tools if one uses DT_Search. The more elaborate TiGer Search does provide the possibility to combine simpler queries into a more complex one, but the post-processing available in the programme itself is still fairly restricted. Using Prolog solves all issues mentioned above, since it is a full-featured programming language rather than a corpus search engine that interprets a query language. One can use variables to refer to the same node at different points in the query and one can define sub-functions (actually: relations) as reusable macros to hide complicated parts of the query. Finally, any manipulation of the data for post-processing can be done right inside Prolog.

In Prolog, there is no principled difference between the data, the queries, the processing and the post-processing. To give a concrete example of why this is useful and efficient from a developer's point of view, consider the concept of dependency paths, mentioned in Section 3.2.1. Recall that in general a dependency path is the list of dependency relations that we cross travelling from a node in the tree up to a node somewhere above it in the tree. Given our assumption that all data is represented as trees, there is always exactly one such path between two nodes. We can use dependency paths in different ways. For instance, we could look in the corpus for pairs of nodes that are connected by a certain dependency path. In this case, the dependency path is part of the query. Alternatively, we might want to know for a given pair of nodes what the dependency path between them is. Here, the dependency path is used in post-processing. If retrieval of nodes from the corpus and post-processing of retrieved nodes is separated from each other, one has to encode the definition twice, possibly using very different tools or languages. In our Prolog setup, we can define the concept of dependency path once, and use it in different ways at different

stages. Furthermore, although specifying a particular dependency path is probably trivial in a query language that is especially tailored to describe linguistic dependencies, it is more of an effort to collect them in post-processing without a dedicated language.

In comparison to many other general purpose languages, Prolog is well suited for corpus work because it is a language that – ideally, and in practice only to a certain extent – allows one to specify the solution to a problem in terms of logic, rather than describing how to get to that solution. This means that in order to find a certain type of linguistic construction in the corpus, one defines what properties the construction should have, and not how to find it. Note that this is also the purpose of having a query language that is interpreted by another programme, such as DT_Search and Tiger Search have. It allows the linguist to just describe the bits of the tree that they are after. The job of the programme is then to find these bits in a corpus.

In order to use the CGN in Prolog, the corpus had to be converted into a Prolog database: a collection of logical representations of facts and relations. Every node in the CGN receives a unique identifier. Associated with these identifiers is information about the node: Which node is its parent, what relation does it have to this parent node, of what category is the node itself, what part of the string falls under the node, etcetera. This representation format is very similar to one of the native formats the CGN comes in, so the conversion is fairly straightforward. In addition, the morphological annotation layer of CGN was merged with the syntactic layer in order to give easy access to detailed morphological information.

There are two concerns that one might have with the approach taken here: memory and speed. Regarding memory, the whole corpus is loaded into memory at once and this puts limits on the size of the corpus. Tools like DT_Search and TiGer Search do not run into this problem because they are capable of searching through corpora that reside on disc. Because the CGN is a moderately sized corpus, and because we are only interested in a subpart of the corpus, the memory limitations were not a problem. If, in the future, we would like to use Prolog to search the large >10mln or even >100mln word corpora that exist, we will have to find a more creative method of accessing the corpus. Secondly, one might be worried about the speed with which the queries are processed. Due to the way the corpus was represented and the fact that there was no file- or disc-access involved in searching the data, performance compares favourably to DT_Search.

The Prolog-version of the definition of Vorfeld occupant (14) is given in (15). A quick walk-through follows below.[4]

(15)
```
vf_occupant(S,V):- vf_constituent(S,V) ∧
                   ¬ ( parent(V,D) ∧
                       vf_constituent(S,D) ).
```

```
vf_constituent(S,V):- category(S,‘smain’) ∧
                      child(S,H) ∧ function(H,‘hd’) ∧
                      descendant(S,V) ∧
                      key_part(V,R) ∧
                      begin(H,BH) ∧ end(R,RE) ∧
                      RE ≤ BH.

key_part(N,N).
key_part(N,C)      :- child(N,C) ∧ function(C,‘hd’).
key_part(N,C)      :- child(N,C) ∧ function(C,‘conj’).
key_part(N,C)      :- ...
```

The (a) and (b) clauses of (14) are implemented in the definition of the relation `vf_constituent` (the constituent is in an SMAIN and partly or fully precedes the head of SMAIN). Vorfeld constituency is a relation between a node `S` and a node `V`: a descendant `V` of an SMAIN `S` is considered a Vorfeld constituent when some ‘key part’ of `V` precedes the SMAIN’s head daughter. Key part is defined as a separate relation, which holds between any node and itself, or between a node and its head daughter, or between a node and a daughter that is one of the conjuncts in a coordination, or between a node and some other type of child that I have not specified here (hence the ellipsis). The statement that any node is considered to be a key part of itself is not superfluous since nodes need not have any daughters. Finally, Vorfeld occupancy is a relation `vf_occupant(S,V)` between two nodes, such that `V` is a Vorfeld constituent of `S` and there is no parent of `V` that is also a Vorfeld constituent of `S`.

The other relations that are used in this definition need to be specified as well, but this is a trivial exercise. Some are atomic, that is, they are directly part of the corpus specification. Examples of atomic relations are `function`, `parent` and `category`, and for words `begin` and `end`. Others can simply be defined in terms of these: `child` in terms of `parent`, and `descendant` recursively in terms of `child`, and `begin` and `end` for non-lexical nodes in terms of the dominated lexical material.

Querying the corpus and post-processing the results is done in Prolog. In this section I have motivated this choice and I have shown how the CGN definition of Vorfeld is implemented. Searching the corpus and manipulating corpus data is only half of empirical

[4]The statements are to be read as follows: Capitals indicate variables, and variable scope is from left of the ‘`:-`’ to the full stop ‘`.`’. The relation named on the left hand side is considered to hold between its arguments when all conditions on the right hand side are met. In this snippet of code, `p(X)` is to be read as: ‘X is a p’, and ‘`p(X,Y)`’ means ‘Y is the/a p of X’. Relation predicates whose name occurs more than once on the left hand side can be proven to hold in more than one way. ‘`key_part(N,N).`’ means: everything is a key part of itself, unconditionally. Finally, anything that cannot be proven is considered false.

work involved in the corpus investigations. The other half is the statistical interpretation of the retrieved data. One of the statistical methods used will be explained below.

3.5 Statistical methods

Of the statistical methods used to analyze corpus data in Chapters 4 & 6, there is one that deserves a brief introduction: *logistic regression*.[5] Here and in the rest of the dissertation, I am assuming that the reader has some understanding of basic concepts such as probability, variance, and significance testing; and is familiar with contingency tables, Fisher's Exact test, χ^2 (test), and Wilcoxon's test. The description of logistic regression given below is purely meant to introduce its main ingredients and to give an intuitive understanding of the method. A reader familiar with logistic regression will not find anything new here, and the unacquainted but interested reader may find the description to be too superficial. In the latter case, I can refer the reader to the relevant chapters of Agresti (1996) and Rietveld and Van Hout (1993), and references therein, for introduction. Agresti (2002) provides a more rigorous mathematical discussion.

A natural way of displaying and analyzing count data is to use contingency tables. However, contingency tables become impractical and hard to interpret when there are more than two or three dimensions, or when the dimensions have many values. When one of the dimensions is continuous, a contingency table cannot be used at all without transforming the data. In all these cases, it may be advantageous to *model* the data in order to investigate them (Agresti, 2002). That is, we describe the data with a mathematical model, and make generalizations about the data on the basis of the *model parameters*. The type of model that we will use here, logistic regression, is a model that is suitable for describing data with a discrete dependent (*response*) variable and discrete or continuous independent (*explanatory*) variables. If we have a two-valued dependent variable whose chance of having value 1 (meaning true, yes, success, something being the case, etcetera) *given* the values of the independent variables is P, a logistic regression model looks as in (16).

$$\text{(16)} \quad \text{logit}(P) = \ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n$$

where α and β's are the parameters of the model, $1,\ldots,n$ are indices for the independent/explanatory variables, and x_i is the observed value of independent variable i for the instance that we are trying to predict the probability of.

[5] In linguistics, especially socio-linguistics, logistic regression has been used since the 70s in the form of VARBRUL (Cedergren and Sankoff, 1974).

To give a more concrete example, the dependent variable will be Vorfeld occupancy in Chapter 4. For each constituent in the dataset, we predict the probability of membership of the group of Vorfeld occupants on the basis of its properties – grammatical function, definiteness, and grammatical complexity. Of course, a constituent either is a Vorfeld occupant or not. The probability of a single constituent being a Vorfeld occupant is the proportion of all constituents with similar properties that occupy the Vorfeld.

The explanatory variables in a logistic regression model can be discrete or continuous. Continuous variables are used as they are but since we need numbers to do our calculations, discrete variables are mapped. A discrete variable that can take on n values is mapped to $n-1$ binary variables, each with numerical values 1 and 0. Each of these new variables represents a value of the original variable. The value of the original variable that is not mapped is referred to as the *base level*. It is represented by all the new binary variables being 0. Examples of mappings for a binary variable 'Winter' (true when it is winter, false when it is not) and a ternary variable 'Springmonth' (with values 'march', 'april' or 'may') are given in (17).

(17) Winter

$$\begin{aligned} &= \text{true} && \Rightarrow && \text{Winter} = 1 \\ &= \text{false} && \Rightarrow && \text{Winter} = 0 \end{aligned}$$

Springmonth

$$\begin{aligned} &= \text{march} && \Rightarrow && \text{Springmonth/march} = 1, \text{Springmonth/april} = 0 \\ &= \text{april} && \Rightarrow && \text{Springmonth/march} = 0, \text{Springmonth/april} = 1 \\ &= \text{may} && \Rightarrow && \text{Springmonth/march} = 0, \text{Springmonth/april} = 0 \end{aligned}$$

The binary variable Winter $\in \{\text{true}, \text{false}\}$ is represented by one binary variable Winter $\in \{1,0\}$. The ternary variable Springmonth $\in \{\text{march}, \text{april}, \text{may}\}$ will be represented by two binary variables Springmonth/march $\in \{1,0\}$ and Springmonth/april $\in \{1,0\}$. The base level of Springmonth is therefore the value 'may'.

Building a logistic regression model of the data gives us insight into patterns in the data because we can interpret the β-parameters.[6] The βs are weights that indicate how and how strongly each factor contributes. (Significantly) positive values mean that as the value of the independent variable increases, the predicted probability of a positive value of the dependent variable increases, too. Negative values indicate a negative correlation. The size of the weight tells us something about the strength of the factor, and can be understood in terms of *odds ratios*, which may bear some explanation.

Like probability, *odds* is a way of expressing the chance that something might happen. Whereas the probability of success P relates the occurrence of a certain event to the total

[6]The α-parameter is the intercept and is thus related to the overall probability of success at the base level of all factors, or put differently, when all xs are 0. As such it is not directly relevant.

number of events, the odds O express how often an event occurs in relation to how often it does not occur. So, if a day being rainy has odds of 3 – or 3 : 1 ("3 to 1") – this means that for every three days of rain, there is one day of sun. Odds and probability can be defined in terms of each other: $O = \frac{P}{1-P}$. So, odds of 3 : 1 correspond to a probability of .75 or 75%. Probability ranges from 0 to 1 (0%–100%); odds lie between 0 and infinity.

The *logit* of a probability P is defined as the (natural) logarithm of the odds O associated with P. The logit of P is the left hand side in (17). The *odds ratio* is the quotient of the odds of two different events: O_1/O_2. For instance, if the odds of a rainy day in August are 3 : 1, and the odds of a rainy day in July are 2 : 5, then the odds ratio is $\frac{3/1}{2/5} = 7\frac{1}{2}$. That is, the odds of it raining in August are 7.5 times the odds of it raining in July.[7] Because odds are always positive, odds ratios will also be positive. An odds ratio of 1 indicates no difference. Odds ratios relate to understanding the β-parameters in a logistic regression model in the following way: The ratio between the predicted odds for two values of variable i is e^{β_i}.

One of the reasons to not directly predict probability from the model parameters, that is, to not use a simple linear model, is that under the logit transformation, the predicted probabilities always fall between 0 and 1. Without the logit, the predictions might fall outside this range, which is conceptually impossible.

Let us turn to a toy example. Imagine a model that predicts whether a day will be rainy during the two summer holiday months July and August. It uses one independent variable (factor) August, which is true (or 1) when the day falls in the month of August, and false (or 0) when the day falls in July. Say we choose the parameters to be $\alpha = -0.89$ and $\beta_{\text{August}} = 1.95$. If we fill in the parameters in (16), we end up with (18):

$$\text{(18)} \quad \ln\left(\frac{P(\text{rainy day})}{P(\text{sunny day})}\right) = -0.89 + 1.95\text{August}$$

This model will for a day in July predict that the odds of a rainy day are $e^{-0.89+1.95*0} = 0.41$, which corresponds to a probability of $\frac{0.41}{0.41+1} = 29\%$. The odds of a rainy day in August are predicted to be $e^{-0.89+1.95*1} = 2.89$, a probability of $\frac{2.89}{2.89+1} = 74\%$. The ratio of the odds of a rainy day in August and the odds of a rainy day in July is $2.89/0.41 = 7$, which is $e^{1.95}$.

Estimating the parameters so that the model describes a given dataset as well as is possible with the information present in the used factors is called *model fitting*. Model fitting is handled by the statistical software (see below). Note that fitted models do not have to describe the data perfectly. It might be that one is missing relevant factors

[7]Note, by the way, that the odds ratio can differ radically from the quotient of the probabilities, called *relative risk*. In our example the relative risk is $\frac{3/(3+1)}{2/(5+2)} = 2\frac{5}{8}$. When saying something like: 'the chance of X happening is n times the chance of Y happening', we therefore need to know whether we are talking about chance in terms of odds or probability.

(explanatory variables), that there are errors in the data related to the way of measuring, or it might be that there is inherent variation in the data that cannot be explained.

After fitting the model, we can draw inferences from the model by looking at the model and several statistics. The parameter estimates themselves indicate the size and direction of the effect. They can be converted to odds ratios so that we can interpret them more easily. The estimates come with standard errors, which allows us to create confidence intervals for them and their associated odds ratios. *Wald's test* allows us to test whether the contribution of an effect is statistically significant, by testing whether a coefficient is significantly not 0.

In Chapters 4 and 6, I will report the results of model fitting in tables like Table 3.2. The tables present parameter estimates, 95% confidence intervals for the odds ratios, and p-values from Wald's test. We can therefore read an estimate of the size of each effect (the OR confidence interval) and the significance of the contribution (p-values) directly from the table.

A confidence interval is an indication of how certain we are about the true value of whatever we are trying to estimate from a sample. Wide confidence intervals indicate low certainty, narrow confidence interval high certainty. The width of a confidence interval depends on the sample and on the demanded level of confidence. If we have little data or much variation in the data, the confidence intervals will be wide. If we demand a high level of confidence (for instance 99%), the confidence interval will be wider than if we had demanded a lower confidence level (say, 95%).

Let us return to predicting rainy days during the summer holiday. Assume that the climate doesn't change and that there thus is such a thing as the true ratio between the odds of rain in August, and the odds of rain in July. We can fit a model predicting the chance of a rainy day in summer using the data from one particular summer. The results of fitting such a model on data from the summer of 2006 are given in Table 3.2. The table shows that on the basis of the available data, we are confident at the 95% level that the odds of a rainy day in August (in general) are between 2.2 and 21.5 times higher than the odds of a rainy day in July. The significance of the difference between August and July (represented by the variable August), can be read from the table in two ways. First, the p-value that comes from Wald's test in the last column is very low. Secondly, the 95% confidence interval for the odds ratio August=true/August=false does not include the value 1. Recall that an odds ratio of 1 indicates no difference.[8] We can therefore be certain that the odds of a rainy day in August are higher than the odds of a rainy day in July. If additional assumptions are met – the sample is representative, we did not ignore other relevant factors, the model predicts a variable that can be meaningfully predicted in this way – the logistic regression model would support the general claim that there are more rainy days in August.

[8]These are two ways of reading the result from the table, they do not represent two different ways of testing whether a factor contributes significantly.

Table 3.2: Predicting rain in the summer holiday of 2006

Parameter	Estimate	OR (lo–hi)		p
α	-0.8938			
August	1.9499	2.2	21.5	<.001

Source: Rain observation data from the archives at `http://www.knmi.nl/`. July 2006 had 9 days of rain and August, 23.

Another way of testing for the significance of parameters is to look at the contribution of a parameter in reducing the deviance (error) of the model. Two nested models[9] can be compared using a likelihood ratio test. If the model with a factor included is significantly better than the model without that factor, the extra factor contributes significantly in predicting the dependent variable. Dropping the factor August from the rain-prediction model leads to a worse fit, which means that August helps to explain variation in the data ($G^2 = 13.131$, $df = 1$, $p < 0.001$). The predictive value of the full model is indicated by the c-index ('c' for concordance). The c-index is an indication of how often the model predicts a higher probability of being true for an actual true outcome than for an false outcome. It is 0.5 when the model has no predictive value, and 1 when the model predicts the data perfectly. Finally, it may also be instructive to compare predicted and observed probabilities.

As mentioned, I will use logistic regression models to investigate what factors influence the probability of a constituent being a Vorfeld occupant. In Chapter 4, I will investigate several factors separately (grammatical function, definiteness and grammatical weight), mostly using contingency tables. However, logistic regression is a perfect tool to investigate the extent to which these factors influence Vorfeld occupation independently and the extent to which the effect of factors can be explained from other factors. Take grammatical function and definiteness as an example. We can expect that the majority of subjects is definite. Suppose we find that both being a subject and being definite increases the chance of appearing in the Vorfeld. We would then want to know whether the subject effect can explain the definiteness effect or vice versa. With a logistic regression model, we can answer this question. If both factors turn out to be significantly positive in the model, then we can conclude that the subjecthood and definiteness effects cannot be explained by each other. A reasonable conclusion would be that being a subject as well as being definite contributes towards appearing in the Vorfeld.

Moreover, we can formulate factors to investigate complicated hypotheses that are hard to test without modeling. In Chapter 6, for instance, I will use logistic regression to investigate whether the relation between definiteness of the subject and definiteness of the object influences Vorfeld occupation of the object.

[9]Two models are nested when the factors in the first model are a subset of the factors in the second.

Although Prolog is well suited to do corpus work with, it is not an obvious choice for statistical calculations. All calculations were done using *R*, (R Development Core Team, 2006). The logistic regression models were fitted using maximum likelihood estimation. Model fitting and inspection was done with the *Design* library (Harrell, 2003). An exception is the last model in Chapter 6. To overcome issues with sparse data and highly correlated factors, a special type of penalized maximum likelihood estimation was used, provided by the *logistf* library (Heinze and Ploner, 2003). See Harrell, Lee, and Mark (1996) and Baayen (forth.) for tutorials and guidelines for good practice. For the other statistics, functions from the standard packages of R where used.

3.6 Summary

This chapter has introduced the CGN and the techniques used to investigate it. In particular I motivated the use of Prolog as a tool in corpus linguistics, and I introduced the statistical method of logistic regression modelling, that will play an important role in Chapters 4 & 6.

I have paid special attention to the definition of the Vorfeld in terms of CGN annotation. A satisfactory definition of Vorfeld occupancy can be given, and moreover, we can easily translate this definition into Prolog. The ability to automatically identify Vorfeld constituents on the basis of the CGN syntactic annotation is a prerequisite for the large scale corpus investigations of the next chapters.

With the expectations about Vorfeld occupation that arose in the discussion in the previous chapter, and the tools and definitions that I have introduced here, we are ready to investigate the Vorfeld in spoken Dutch. In Chapter 4, I will present a corpus investigation that will help us answer the question of how a constituent's properties influence the chance that it is selected as the Vorfeld occupant. The second question to be answered in this dissertation introduced in Chapter 1, is how the chance of communicative success influences the choice of a Vorfeld occupant. Chapter 5 proposes a theoretical model of the influence of communicative success. The tools discussed in the current chapter will be employed to test predictions of this theoretical model in Chapter 6.

Chapter 4

A Corpus Study of the Vorfeld

In Chapter 2 I formulated expectations about the relation between constituent properties and Vorfeld occupation. I proposed that we should understand Vorfeld occupation as a combination of global word order trends and special properties of the Vorfeld as a highlighting position. In this chapter, I present a corpus study of the Vorfeld, and show that the expectations raised in Chapter 2 are to great extent borne out. The chapter thereby provides an answer to the first subquestion of the general research question, formulated in Chapter 1: How do properties of a constituent influence the chance that it is selected as Vorfeld occupant. In answering this question I will make extensive use of the tools and corpus that have been discussed in Chapter 3.

The chapter is organized as follows. Section 4.1 gives an overview of the Vorfeld in the spoken Dutch corpus CGN. This section also gives details of the dataset extracted from the CGN that is to be used in the rest of the dissertation. The four sections that follow form a single unit. Together they answer the question of how properties of a constituent are linked to its chance of appearing in the Vorfeld. In Section 4.2, I look into the relation between grammatical function and Vorfeld occupation. According to the hypothesis that global word order tendencies underly Vorfeld occupation, we expect that, of subjects, indirect objects and direct objects, subjects front most often and direct objects least often. Section 4.3 investigates the relation between definiteness and Vorfeld occupation. This relation is complex. On the one hand, the correspondence of Vorfeld occupation with Mittelfeld word order leads us to expect that pronouns front most often, followed by definite full NPs, and that indefinite full NPs front least often. On the other hand, we expect that personal pronouns are not welcome in the Vorfeld, because they do not realize important information (in the sense of Gundel, 1988). The third and final constituent property to be investigated is grammatical complexity. In Section 4.4, I investigate whether less complex material has an increased chance of appearing in the Vorfeld. I will also

Table 4.1: Top dependency path counts in the Vorfeld, with proportion estimates at the 95% confidence level. ▶

consider in some detail whether complexity is a global effect, or whether the effect of complexity on Vorfeld occupation is an epiphenomenon of the alleged tendency of complex material to occur at the right periphery of a clause. Section 4.5 wraps up the investigation of the relation between constituent properties and Vorfeld occupation. A logistic regression model of the findings of the preceding three sections will give further insight into how the three constituent properties grammatical function, definiteness and complexity combine to influence Vorfeld occupation.

In Chapter 2, I also briefly discussed the observation made in the literature that the presence of negation facilitates the occurrence certain non-canonical word orders. To follow up on this, I will look into the influence of negation, and sentence adverbs in general, on direct object fronting in Section 4.6. Section 4.7 summarizes the findings of this chapter.

4.1 Some first statistics

The subset of the Corpus of Spoken Dutch that was selected in the previous chapter (*components a–d* and *f–n*) contains 71 934 V2-clauses (tagged as SMAIN), with an average length of 8.6 words, not counting fillers, interjections and restarts.

From each of these sentences, the constituents that met our definition of Vorfeld occupant (Section 3.3) were collected, and the *dependency path* between a Vorfeld occupant and its main clause node was recorded. Recall from Section 3.4 that a dependency path is the list of dependency relations that we cross when we traverse the syntactic tree between a node and the top of the tree. Without abstracting away from the CGN annotation, dependency paths can give us a first impression of which items, in terms of grammatical relations, can appear in the Vorfeld, and how often they do so. In order to quantify the relative tendency of material in different grammatical relations to appear in the Vorfeld, I have also counted how often each recorded dependency path occurs postverbally.

Table 4.1 gives Vorfeld and elsewhere counts, as well as estimates of the proportion of Vorfeld occupancy for each dependency path.[1,2,3] The less obvious dependency

[1]The confidence intervals for proportion in Table 4.1 and further tables in this chapter are at the 95% level. The intervals give us an *indication* of significant difference between two proportions: Two non-overlapping 95% confidence intervals are significantly different *below* the 5% level. Two proportions that are significantly different at the 5% level may have slightly overlapping 95% confidence intervals. The intervals are exact binomial proportion intervals. See Section 3.5 for a discussion of confidence intervals in the context of logistic regression.

Path	Vorfeld yes	Vorfeld no	Prop est (%) lo	pt	hi
SUP	1 839	892	65.5	**67.3**	69.1
SU	46 216	24 932	64.6	**65.0**	65.3
MOD	15 016	55 260	21.1	**21.4**	21.7
OBJ1 PC SU	13	73	8.3	**15.1**	24.5
OBJ1 VC	1 193	6 964	13.9	**14.6**	15.4
OBJ1	2 634	15 777	13.8	**14.3**	14.8
OBJ1 PC BODY VC	19	121	8.4	**13.6**	20.4
OBJ1 VC VC	104	670	11.1	**13.4**	16.0
OBJ1 PC VC	189	1 239	11.5	**13.2**	15.1
OBJ1 PC PREDC	58	386	10.1	**13.1**	16.6
OBJ1 PC OBJ1	38	256	9.3	**12.9**	17.3
LD	665	4 988	10.9	**11.8**	12.6
OBJ1 PC	259	2 153	9.5	**10.7**	12.0
OBJ1 PC VC VC	14	122	5.7	**10.3**	16.7
OBJ1 BODY VC	49	437	7.6	**10.1**	13.1
POBJ1	21	199	6.0	**9.5**	14.2
OBJ1 LD	300	3 251	7.6	**8.4**	9.4
PREDM	87	1 421	4.6	**5.8**	7.1
OBJ2 VC	25	416	3.7	**5.7**	8.3
OBJ1 MOD VC VC	15	250	3.2	**5.7**	9.2
continues on right…					

Path	Vorfeld yes	Vorfeld no	Prop est (%) lo	pt	hi
continued from left…					
LD VC	88	1 780	3.8	**4.7**	5.8
OBJ1 LD VC	59	1 247	3.5	**4.5**	5.8
MOD VC	263	6 111	3.7	**4.1**	4.6
OBJ1 MOD OBJ1	32	782	2.7	**3.9**	5.5
OBJ2	23	569	2.5	**3.9**	5.8
OBJ1 PREDC	42	1 041	2.8	**3.9**	5.2
OBJ1 MOD VC	78	2 035	2.9	**3.7**	4.6
TAG OBJ1	22	596	2.2	**3.6**	5.3
OBJ1 MOD SU	22	589	2.3	**3.6**	5.4
MOD VC VC	24	673	2.2	**3.4**	5.1
OBJ1 MOD OBJ1 VC	16	454	2.0	**3.4**	5.5
PC VC	49	1 533	2.3	**3.1**	4.1
PREDC VC	34	1 199	1.9	**2.8**	3.8
PREDC	459	16 831	2.4	**2.7**	2.9
OBJ1 MOD	144	5 323	2.2	**2.6**	3.1
PC	62	2 566	1.8	**2.4**	3.0
Rest *(119 paths)*	314	82 316	0.3	**0.4**	0.4
Total	70 485	245 827	22.1	**22.3**	22.4

Abbreviations: In the table head *pt* stands for point estimate ($\frac{\text{vorfeld}}{\text{vorfeld+elsewhere}}$, boldfaced), *lo/hi* for the lower/upper boundary of the 95% C.I. In the dependency paths SUP/POBJ1 means 'preliminary' subject/direct object, PC/VC prepositional (i.e., oblique)/verbal complement, PREDC/PREDM predicative complement/modifier, LD locative or directional adverbial, TAG discourse tags, BODY core of a CP or VP.

relation abbreviations are explained in the legend below the table. A full list of CGN abbreviations can be found in Section A.2, in the appendices. Appendix B gives an example sentence for each path in the Table 4.1. Below, I will begin by discussing, on the basis of Table 4.1, the exact data selection for the experiments the rest of the chapter. I will give illustrative examples of the data that is excluded from the selection (Section 4.1.1). After that, Section 4.1.2 discusses how grammatical function in a sentence is related to the dependency paths.

4.1.1 Data selection

In Table 4.1 there is a total number of 70485 Vorfeld occupants. This number is lower than the earlier mentioned number of SMAINs in the corpus for two reasons. First, a fair number of SMAINs lack a head-daughter, material in the Vorfeld, or both. This often occurs in coordinations which may share a head or another dependent using so-called secondary edges (see Section 3.2.2). An example of a coordinated SMAIN 'missing' a head, that is, missing a left-bracket, is in (1). In the first SMAIN, *Stalin* is in the Vorfeld, and *is* is the verbal head (boldfaced). Gapping in the second conjunct results in a headless SMAIN. There is no independent way of establishing whether *Adolf Hitler* is in the Vorfeld or not.

(1) [SMAIN Stalin **is** twee] en [SMAIN Adolf Hitler één]
Stalin is two and Adolf Hitler one
'Stalin comes in second, and Adolf Hitler first.' (Nl-l 7388:30)

All 1426 such cases were excluded.

Secondly, SMAINs with multiple heads were not counted. Such SMAINs are instances of the mirror construction (Section 2.4.4). An example is given in (2). The two heads are given in boldface.

(2) ik **kwam** om één uur **kwam** ik thuis
I came at one o clock came I home
'I came home at one.' (Nl-a 254:530)

[2]Although path counts are independent of each other in the case of Vorfeld tokens, they are not in the case of elsewhere tokens. Every token of a certain path is also a token of a suffix of that path. So, elsewhere counts for OBJ1 VC (*direct object* under a *verbal complement* dependent) are necessarily higher than those for just VC. This need not be a problem at this level of analysis, but we need to keep in mind that suffixes tend to have lower proportions in the Vorfeld, compared to their superpaths.

[3]The interpretation of the proportion given for the total (22%) is not so clear. It would seem to express that constituents that have a dependency path that is observed to lead to the Vorfeld at least once occur in the Vorfeld around 22% of the time. On the (naïve) assumption that being allowed to occur in the Vorfeld is equal to occurring there at least once, we could say that the prior probability of Vorfeld occupation for a constituent that *could* appear in the Vorfeld is 22%. The fact that this probability is so high is due to the high number of subjects and modifiers amongst the constituents.

Mirror constructions were excluded too, because for these the notion of Vorfeld is hard to define: It appears as if the construction has two Vorfelds. Also, in this example, there is uncertainty about whether the subject is in the Vorfeld or not, since it appears twice: once before the first head, and once directly after the second.

The remaining 70485 Vorfeld occupants are divided over 69917 SMAINS. There are 499 sentences with more than one Vorfeld occupant.[4] Inspection of this set shows that sentences with more than one Vorfeld occupant were often annotation or extraction errors. All cases with 4 and 5 Vorfeld occupants were errors. However, (correctly marked up) sentences with 3 Vorfeld occupants do occur in the data, for instance the example in (3). The Vorfeld occupants are separated by '|', boldface marks the head. The subject is *ze* 'they', and appears in the Mittelfeld.

(3) in het begin | voor ieder vak | voor die mannen **moeten** ze dan hun
in the beginning for each subject for those men must they then their
boek bijhebben
book have
'In the beginning, these men want them to bring their books for each subject.'

Also quite common is a Vorfeld containing a focus particle and another constituent (see Section 2.4 for discussion). I will exclude these cases from the dataset in the rest of the chapter, *except* in Section 4.6, where excluding these cases would bias the data too much.

The definition of Vorfeld occupant in Section 3.3 considers a constituent to be in the Vorfeld if a key part of it precedes the head of SMAIN. This means that the Vorfeld in (4) is occupied by the subject.

(4) HD SU: [tw eentje] is MOD SU: [bw d'rvan] kapot
one is of it broken
'One of them is broken.' (Nl-a 363:140)

[4]There is one exception in the dataset. The sentence in (i) features one constituent that is the Vorfeld occupant of multiple SMAINS according to our definition. The heads directly after the respective Vorfelds are in boldface, the constituent in question is between square brackets.

(i) [in een heidegebied] **wordt** xxx gezegd vennen natte heide droge heide **zijn** beschermd
in a heath area is NOISE said lakes wet heath dry heath are protected
'In a heathy area, it is said, small lakes and wet and dry heath are protected.' (Vl-n 400016:46)

The complete sentence is annotated as an SMAIN headed by *wordt*. The SMAIN headed by *zijn* is a direct object of *gezegd*. The fronted PP modifies this latter SMAIN. It is completely unclear to me whether this is a reasonable annotation, whether this sentence is grammatical or not, and whether *in een heidegebied* should be in the Vorfeld of the first or of the second SMAIN. One might speculate that this sentence is a blend of 1) a sentence with a finite comment clause or parenthetical *wordt gezegd* 'is said' (Schelfhout, 2005), and 2) a sentence containing direct speech *er wordt gezegd: vennen etc.* 'it is said: small lakes etc.'. I will leave further speculation to the reader.

The numeral (TW, telwoord) *eentje* 'one' heads the discontinuous subject NP and has the adverbial (BW, bijwoord) as a dependent modifier. As we start to investigate different aspects of these structures, discontinuity of this kind becomes a problem. One of the things we would like to know about Vorfeld occupants is their length, but there is no straightforward answer to how long the Vorfeld occupant in (4) is. Discontinuous cases like (4) are therefore excluded from the dataset (132 in total). Note that discontinuous constituents that are completely postverbal are not excluded.

In Table 4.1, the dependency relation that shows the highest tendency of appearing in the Vorfeld is the expletive or 'preliminary' subject (SUP). The SUPs mark the expletive subject in a subject extraposition (5a). Extraposition of the direct object may be accompanied by insertion of a preliminary object, marked POBJ1. An example is in (5b). In both cases, the logical argument keeps the normal dependency relation. In (5a), the logical subject is a type of infinitival clause (OTI, '*om*... *te*-infinitive'). In (5b), the logical object is a CP.

(5) a. SUP **'t** voert een beetje ver [OTI SU **om dat helemaal te doen**]
EXPL leads a bit far COMP DEM completely INF do
'Doing that completely would be taking it a little too far.' (Nl-a 280:58)

b. 'k vind POBJ1 **'t** altijd tof [CP OBJ1 **als ik zo veel verschillende dingen kan doen**]
I find EXPL always cool when I can do so many different things
'I always think it is cool to be able to do so many different things.'
(Vl-a 400454:6)

An extraposed subject/object will not appear in the Vorfeld, so the inclusion of sentences with extraposed subjects/objects would lead to a bias against Vorfeld occupation. Hence, sentences that contain a SUP or a POBJ1 descendant were removed.

Incidentally, inspection of the excluded items shows that SUP in a good portion of the excluded items is used for the expletive subject of an existential construction. An example is given in (6).

(6) SUP **d'r** is [NP SU **geen contact tussen hun en mijn muur**]
EXPL is no contact between their and my wall
'There is no contact between their wall and mine.' (Nl-a 250:68)

The CGN annotation guidelines require the expletive subject of an EC to be marked as a modifier, however. Therefore, excluding sentences with SUP dependents does not completely remove ECs from the dataset.

Finally, as discussed in Section 2.4, I excluded left dislocation and hanging topics. These constructions differ from topicalization in form and function. In the CGN, left-dislocated material and hanging topics are put in the topological *lead* – preceding the clause proper. The complete utterance is a discourse unit (DU), the lead material is a satellite dependent (SAT), and the main clause the nucleus (NUCL)

(7) [DU [NP de benedenbuurman bij mij] [SMAIN die rookt ook]]
SAT NUCL
the downstairs neighbour at me DEM smokes too
'My downstairs neighbour, he's a smoker, too.' (Nl-a 250:256)

In order to get rid of left-dislocation and hanging topics, SMAINs that have a SAT sibling were excluded from the dataset.

To summarize, the dataset that I use in the rest of the chapter is described in (8).

(8) All SMAINs in the CGN that...
a. occur in *components a–d* or *f–n*, and
b. have exactly one head, and
c. have exactly one Vorfeld occupant, properly contained in the Vorfeld, and
d. do not contain expletive subjects or direct objects, and
e. do not occur together with a SAT node in a DU (discourse unit).

About 62k sentences meet the requirements in (8).

4.1.2 Subjects and objects in a sentence

Part of the expectations about Vorfeld occupation that I formulated in Chapter 2 is phrased in terms of subject, direct object and indirect object. These grammatical functions translate roughly to the CGN dependency relations SU, OBJ1 and OBJ2 (see also Section 3.2.1). A problem with simply equating subject, direct object and indirect object with the dependency relations SU, OBJ1 and OBJ2 is that we do not know at what level in the syntactic tree a dependency relation occurs. This information is important. For instance, the Vorfeld occupant of (9) has the dependency relation SU, but it is the subject of the embedded clause, not of the matrix clause. The long dependency path shows this. The dependency path of a matrix subject would be simply SU.

(9) dat hoop ik [CP dat [SSUB ook ter sprake zal komen]]
SU BODY OBJ1 SU OBJ1 BODY OBJ1
DEM hope I CMP too will be discussed
'I hope that that will be discussed, too.' (Nl-a 594:121)

The sentence is decidedly marked, or even marginal. The matrix subject *ik* is directly to the right of the finite verb. Sentence (9) should not be considered to be subject-initial. Nodes with a SU dependency relation at a embedded level should not be pooled with matrix subject data.

This problem is overcome by defining subject and object in terms of their dependency paths. Matrix subjects are always direct dependents of the matrix clause, so the dependency path SU (as in Table 4.1) will be used to refer to subjects. Objects however may be objects of the finite verb, or of a non-finite verb. To illustrate, we find pairs like the following in the data.

(10) a. OBJ1
dat verwachten wij niet
that expect we not (Nl-a 724:208)

b. OBJ1 VC HD VC
dat hadden we niet verwacht
that had we not expected
'We did not expect that to happen' (Nl-a 308:314)

We have no reason to believe that embedding under a temporal, modal or any other (auxiliary) verb as in (10b) has an influence on the amount of object fronting. Indeed, if we compare the proportions of OBJ1 (14.3%) and OBJ1 VC (14.6%, difference not significant, $p = 0.495$, 2-tailed Fisher's) in Table 4.1, we can see that they are very similar. It seems therefore to be safe to pool the data for dependency paths OBJ1 and OBJ1 VC, and refer to them collectively as direct objects. We will do the same for indirect objects, which refer to the paths OBJ2 (3.9% in Vorfeld) and OBJ2 VC (5.9%, difference not significant, $p = .182$, 2-tailed Fisher's).

Indeed, for all arguments in Table 4.1, the proportion of Vorfeld occupation is hardly affected by an extra level of VC. For instance, objects in oblique PPs (OBJ1 PC, direct object under prepositional complement), oblique arguments (PC), and predicative complements (PREDC) all show similar Vorfeld proportions, whether they are below a VC or not.

A direct advantage of pooling data for the objects is that we have more data available. This is especially relevant for the indirect object data. In Table 4.1, there are only about 1k OBJ2 (VC)s. In addition, just a small proportion of OBJ2 (VC)s appears in the Vorfeld. Data sparseness is a real danger in the indirect object data. If we did not pool the data, this problem would be even worse.

Another positive consequence of pooling the dependency paths is that we also pool sentence types. For instance, a sentence with an OBJ1 VC direct object always has a non-finite verb. This verb resides in the verb cluster, the topological right bracket. Schematically, sentences with an OBJ1 VC dependent are of the form $XV_{fin}XV^*$. A

sentence with a direct object that is a OBJ1 *may* lack a right bracket, and therefore be of the form $XV_{fin}X$. The presence or absence of a right bracket will turn out to be relevant in the investigation of grammatical complexity in Section 4.4.

We have had a first look at the Vorfeld in the CGN. I have used the overview presented in Table 4.1 to motivate the data selection. In Sections 4.2 through 4.5, I will use this data to show that grammatical function, definiteness and grammatical complexity each influence the chance that an argument occupies the Vorfeld.

In Section 4.6, I will use a slightly larger selection of data to show that certain adverbs are associated with topicalization of direct objects. The different data selection used in that section will be motivated there.

4.2 Arguments

With the corpus, and subject and object defined as in the previous section, we can take a first look at the influence of grammatical function on Vorfeld occupation. In Section 2.6.1, I hypothesized that subjects show the strongest tendency to appear in the Vorfeld, followed by indirect objects. Direct objects would have the lowest tendency to do so. This expectation is based on the canonical word order in the Dutch Mittelfeld, which is subject $\prec$ indirect object $\prec$ direct object. The results in this section will, however, not fully answer the question whether canonical Mittelfeld order is reflected in Vorfeld occupation. In order to answer this question, we need to control for the (potential) influence that other constituent properties have on Vorfeld occupation. We will have to wait until Section 4.5 before we can properly assess the influence on grammatical function independently of potential interfering variables.

4.2.1 Corpus results

Table 4.2 gives details of subject, direct object and indirect object fronting. The counts and proportions differ slightly from Table 4.1 because of the data selection explained in the previous section.

At the very general level of Table 4.2, the expectation that subjects appear most frequently in the Vorfeld is borne out. About 70% of the subjects occupy the Vorfeld. In the case of subjects, we may also turn this number around: Every sentence has a subject, so about 70% of the Vorfelds contain a subject. However, there is no sign yet of indirect objects topicalizing more frequently than direct objects. In fact, Table 4.2 shows that on average, direct objects topicalize a lot more often than indirect objects do. This may reflect the findings of Lamers (2001) and Thrift (2003), who found that indirect objects in the Vorfeld are marked compared to direct objects on the basis of amongst

Table 4.2: Summary of Vorfeld occupation of arguments.

	Vorfeld		Prop est (%)		
Argument	yes	no	lo	pt	hi
subject	43 523	18 597	69.7	**70.1**	70.4
direct object	3 418	20 432	13.9	**14.3**	14.8
indirect object	38	815	3.2	**4.5**	6.1

Note: subject = SU, direct object = OBJ1 + OBJ1 VC, indirect object = OBJ2 + OBJ2 VC.

Table 4.3: Classification after part-of-speech and syntactic category.

Category	CGN labels
nominal	NP, N, VNW, MWU (when proper names)
prepositional	PP, VZ
verbal	TI, OTI, AHI, INF, WW, PPART
clausal	CP, WHSUB, WHQ, WHREL, REL, SVAN, SMAIN, SSUB, SV1

Note: See Appendix A for explanation of the CGN POS/Cat-labels. Conjunctions/lists of one category are also assigned that category. Other POS-types (notably adjectives and adverbs) were assigned to a rest category.

other things questionnaire data. We will see in later sections that there is more to the difference between direct and indirect objects than meets the eye, however. If we take the effect of definiteness on Vorfeld occupation into account, the difference between direct objects and indirect objects is not as big as Table 4.2 suggests.

Subjects and objects can be a of a wide variety of categories. We can divide the data of Table 4.2 into four main categories: *nominal*, *prepositional*, *verbal* and *clausal*. The translation between CGN-tags and the four categories is given in Table 4.3. The categories *nominal* and *prepositional* should be self-explanatory. The difference between *verbal* and *clausal* is that clausal constituents are finite, and contain all arguments of the verb, whereas verbal constituents are non-finite or do not contain all arguments of the verb. Tables 4.4–4.6 show how each of the grammatical functions breaks down into these categories. Below, I will illustrate the data with some examples for each grammatical function. The nominal data will considered in more detail in the section on definiteness (Section 4.3).

Subjects Vorfeld occupation of subjects per category is detailed in Table 4.4. The proportion of subjects in the Vorfeld is high in each category, although clausal subjects appear to have a slightly reduced chance of appearing in the Vorfeld.

Table 4.4: Subject fronting per category.

	Vorfeld		Prop est (%)		
Category	yes	no	lo	pt	hi
nominal	42 973	18 307	69.8	**70.1**	70.5
prepositional	18	3	63.7	**85.7**	97.0
verbal	79	21	69.7	**79.0**	86.5
clausal	206	147	53.0	**58.4**	63.6
Rest	247	119	62.4	**67.5**	72.3
Total	43 523	18 597	69.7	**70.1**	70.4

Table 4.5: Direct object fronting per category.

	Vorfeld		Prop est (%)		
Category	yes	no	lo	pt	hi
nominal	3 247	14 825	17.4	**18.0**	18.5
prepositional	13	74	8.2	**14.9**	24.2
verbal	16	106	7.7	**13.1**	20.4
clausal	78	4 331	1.4	**1.8**	2.2
Rest	64	1 096	4.3	**5.5**	7.0
Total	3 418	20 432	13.9	**14.3**	14.8

Table 4.6: Indirect object fronting per category.

	Vorfeld		Prop est (%)		
Category	yes	no	lo	pt	hi
nominal	36	687	3.5	**5.0**	6.8
prepositional	2	124	0.2	**1.6**	5.6
Rest	0	4	0.0	**0.0**	60.2
Total	38	815	3.2	**4.5**	6.1

The nominal data contains some sentences with more than one subject. Certain dialects of Dutch allow subject doubling. In these dialects, the subject can be expressed more than once by non-inflectional subject markers, such as clitics or (free) pronouns (De Vogelaer, 2005). In the CGN, all these expressions are marked up as a subject dependent. There were 169 sentences with two subjects, 21 sentences had three, and 1 sentence showed 'subject quadrupling', as illustrated in (11). The subjects are boldfaced.

(11) **ik** zijn **'k** **ik** nog altijd bezig **ik**
I am I.RED I still busy I
'I'm still busy, me.' (Vl-a 400727:88)

In (12), we find examples of verbal (a) and clausal (b) subjects in the Vorfeld.

(12) a. **dingen op maat** **maken** kost altijd tijd
things at measure make costs always time
'Making things to measure always takes time.' (NL-a 303:118)
b. **hoe ze** **d'r uitzien** maakt niet uit
how they look makes not PRT
'What they look like makes no difference.' (NL-a 303:149)

As was to be expected, the numbers in these categories are heavily influenced by my previous choice to exclude sentences with an expletive subject SUP. If we had included these sentences, verbal subjects would have a proportion in the Vorfeld of 25% (79/316), and clausal subjects of $\sim$ 20% (206/988). However, even in the data as given in Table 4.4, the group of clausal subjects has a relatively low chance of appearing in the Vorfeld: Whereas the proportion of nominal, prepositional and verbal subjects that fronts is 70% or higher (point estimate), the proportion of clausal subjects is around 60%. This suggests that one does not need an expletive subject to move a clausal subject to the end of the sentence. Inspection of the dataset shows that the material in the Vorfeld in expletive-less extraposition are typically modifiers. However, in (13), an adjectival predicate takes up the Vorfeld.[5]

(13) (In a book review:)
nieuw is **dat hij politiek en politici van vroeger en nu vergelijkt**
new is that he politics and politicians of past and now compares
'New [about the book] is that he compares politics and politicians of the present and the past.' (Vl-l 601074:50)

Clausal constituents can be considered to be complex constituents. This reduced chance of Vorfeld occupation is a first indication that grammatical complexity influences Vorfeld occupation. Section 4.4 will consider the effect of grammatical complexity in more detail.

[5]The example has a journalistic flavour to it. See Birner and Ward (1998) for similar examples also involving V2 in English.

Prepositional subjects are mostly of an elliptic nature. In (14), a motion verb is dropped, resulting in a bare PP.

(14) **naar Praag** kan in één keer
to Prague can in one go
'You can go to Prague directly.' (Vl-a 400392:173)

Direct objects Table 4.5 breaks down the direct object data into categories. As was the case with subjects, the majority of direct objects is nominal. In the direct object data, we find sentences with more than one direct object, too (62 in total). On top of repetitions or elaborations, these are sentences with an upstairs and a downstairs object, as in (15).

OBJ1 VC SU OBJ1
(15) dat hoort u mij niet zeggen
'You will/do/did not hear me say that.' (Nl-g 8:49)

Verbal direct objects in the Vorfeld are direct objects of *vinden* 'find, consider' or *doen* 'do', as in (16), since Dutch has (optional) do-support for fronted verbs.[6]

(16) **voorlezen** kun je maar beter wel doen
read out can you PART better AFF do
'It is better to read to your children.' (Nl:f 7250:117)

Verbs of saying and knowing are the reason for the large proportion of clausal direct objects, compared to the subject data. An example of a clausal direct object is in (17).

(17) **dat de satiricus op dit soort momenten verstomd is** kan ik begrijpen
that the satiricist on this kind moments silent is can I understand
'That the satiricist is silent on occasions like this, I can understand.' (Nl-l 7209:45)

The group of clausal direct objects in the Vorfeld also contains some tag-like constructions, in which the direct object in the Vorfeld is a main clause, and the matrix head and subject

[6]The do-support in the example is interesting, because it cannot have been triggered by the necessity of filling the left bracket (cf. Grimshaw, 1997, who proposes English do-support is triggered by the need to have a head for the CP). The ungrammaticality of the *doen*+verb in situ shows that it is a kind of do-support.

(i) Je kunt maar beter wel voorlezen (*doen)
You can PART better AFF read out do
'It is better to read to your children.' (constructed)

Variation like this also shows the limits of taking the dependency path as a characterization of a constituent's role. In the Vorfeld *voorlezen* is an OBJ1 VC, and *doen* is HD VC. If *voorlezen* were in in the verbal cluster it would be a (HD) VC. This holds mainly for verbal constituents, which will not be the focus of this dissertation. Still, future work should keep this in mind.

follow it. In these cases, it is not clear whether we are dealing with simple word order variation, as a canonical SVO variant is not available (18).

(18) a. [$_{\text{smain}}$ **dit kan ik nog wel lezen**] zie ik
this can I PART AFF read see I
'I see I can still read this.' (Nl-l 21:5)
b. *Ik zie dit kan ik nog wel lezen. (constructed)

As was the case with clausal subjects, clausal objects have a reduced chance of appearing in the Vorfeld compared to other types of direct objects. We can again take this as preliminary evidence for an effect of grammatical complexity on Vorfeld occupation.

Verbs of saying are also responsible for the large group of direct objects that do not fall into one of the four categories: They may take discourse units (DU) as their direct object.

Prepositional direct objects are often corrections, continuations or clarifications embedded under a verb, but they can also be cases of partitives and head noun ellipsis. Unlike with prepositional subjects, ellipsis of motion verbs is not very common here. The example in (19) is a case of sluicing in an embedded context.

(19) **in welke volgorde** wil je nu natuurlijk weten
in what order want you now obviously know
'Now, of course, you want to know in which order.' (Vl-a 400327:6)

Indirect objects The breakdown of indirect objects is in Table 4.6. The lack of clausal and verbal objects is to be expected. Indirect objects are generally beneficiaries, experiencers or recipients, and mostly human. Verbal and clausal arguments typically denote actions, events, propositions or utterances. The high number of prepositional indirect objects is due to the fact that the roles associated with indirect objects are often expressed with PPs headed by *voor* 'for' or *aan* 'to'. In the data, prepositional indirect objects always have nominal internal objects.

For many indirect objects, there is the possibility of *dative alternation*. The indirect object can be realized as a NP or an PP (see Van der Beek, 2005, for a corpus investigation of the dative alternation in Dutch). For instance, the corpus contains the following pair (20), an example of the dative alternation with the prototypical ditransitive *geven* 'give'. Note that (20b) has a direct object in the Vorfeld, whereas (20a) has an adverbial in the Vorfeld and its direct object in the Mittelfeld. The indirect object is in boldface.

(20) a. dan geef ik dat ook maar weer **aan hem**
then give I DEM too PART PART to him
'Then I will just give him that, too.' (Nl-a 625:290)
b. die geef 'k **'m** ook
DEM give I.RED him.RED too
'I will give him that too.' (Nl-a 322:139)

Dative alternations can in principle occur with indirect objects in the Vorfeld, too. Since there are so few prepositional indirect objects in the Vorfeld, however, finding a good minimal pair is not possible. The data does contain the following pair, showing an NP indirect object and a PP indirect object of the same verb in the Vorfeld. The examples have been shortened.

(21) a. **aan de bevolking** is eenvoudigweg gezegd dat [+12w]
to the people is simply said that
'The people were simply told that . . . ' (Vl-j 600551:36)

b. **ons** hadden zun ook wel gezegd dat [+9w]
us had they too PART said that
'We had been told that . . . ' (Vl-a 400092:196)

The verb with the highest (token) frequency in the indirect object data is *lijken* 'seem/appear'. This verb does not allow for an indirect object PP, and it is not a ditransitive. Interestingly, of the 123 occurrences of a indirect object in the context of *lijken*, only two are in the Vorfeld. One of these is in (22). The subject of *leek* is *'t*, and that *iedereen* is the indirect object:

(22) **iedereen** leek 't echt superranzig
everybody seemed it really super gross
'Everybody thought it was really disgusting.' (Nl-a 391:176)

The numbers in the indirect object data are too low to give any reliable answers to the questions whether prepositional indirect objects appear in the Vorfeld less often than their nominal counterparts, or whether having the possibility of dative alternation influences Vorfeld behaviour. These are valid questions that require a bigger dataset to be answered. Another difficulty with the indirect object data is that it is quite heterogeneous. For instance, the data contains prototypical transfer ditransitives, but also dative experiencer transitives. In the future, it would be interesting to investigate whether these two (and other) indirect object types differ in their Vorfeld behaviour. In this dissertation I will not do so, because it involves finding out which verb selected the indirect object. This information cannot in general be easily obtained from the CGN (Section 3.2.2).

Although the data in Table 4.6 allows us to compare PP indirect objects with NP indirect objects, it does not tell us anything about the popularity of preposition stranding in the indirect object data. Looking back at Table 4.1, we can see that, on the whole, preposition stranding is very common. Take the dependency paths PC (VC), which refer to oblique (prepositional) complements. There are 49 PC + 62 PC VC = 111 oblique complements in the Vorfeld. In these cases the whole PP occupies the Vorfeld. The paths OBJ1 PC (VC) refer to the internal objects of oblique complements. Of those internal objects, 259 OBJ1 PC + 189 OBJ1 PC VC = 418 are in the Vorfeld. There are almost four

times as many objects cases of preposition stranding than cases of in which the whole oblique PP is taken to the Vorfeld. In the indirect object data, there is only one case of preposition stranding (that is, OBJ1 OBJ2 (VC) in Vorfeld), namely the example in (23). The preposition and its object are in boldface.

(23) **daar** hoef je niet zoveel **aan** uit te leggen
there have you not so much to explain.INF
'One doesn't have to explain him/her/them so much.' (Nl-b 140:165)

Table 4.6 shows that 2 of the indirect object PPs are in the Vorfeld. If indirect object PPs were anything like (other) oblique arguments PPs, we would expect there to be about 8 cases of preposition strandings in the indirect object data. Inspection of the prepositional indirect objects provides us with a possible explanation of the unpopularity of preposition stranding in this group. Amongst the 126 prepositional indirect objects, there are only 3 with an R-pronoun, that is, there are only 3 prepositional indirect objects that readily allow preposition stranding (see Section 2.2.3). One of these cases is (23), the other two have the complete prepositional direct object in the Mittelfeld, and are given in (24).

(24) a. 'k geef **daar** geen les **aan**
I give there no teaching to.
'I will not teach him/her/them.' (Vl-a 400288:23)

b. je kan **er** altijd iets **aan** vragen en zo
you can there always something to ask PART
'One can always ask him/her/them things.' (Vl-b 400155:299)

Overall, there is not enough evidence to decide whether preposition stranding is more or less common with indirect objects than with other prepositional complements. The fact that only one example shows up could be explained by the low number of R-pronouns as PP-internal objects in the indirect object data. There might be a dispreference for using R-pronouns for human referents, which in turn would explain this low number.[7]

4.2.2 Summary

On average, subjects have the strongest tendency to appear in the Vorfeld amongst subjects, direct objects and indirect objects. Indirect objects show the lowest average proportion of topicalization. However, this does not permit us to conclude much about the influence of canonical word order on Vorfeld occupation yet, because we have not factored out other influences. We have also seen first indications that grammatical complexity plays a role

[7] There certainly seems to exist a *prescriptive rule* against using R-pronouns for human referents, especially in formal registers. See e.g. `http://taaladvies.net/taal/advies/vraag/887/` (in Dutch), for 'language advice' on the use of *waarmee* 'wherewith' vs *met wie* 'with whom'. Even though I am skeptical about the linguistic reality behind this rule, it is hard to judge from the current annotations.

in Vorfeld occupation. For both subjects and direct objects, clausal constituents appear less frequently in the Vorfeld than constituents of other categories. As for the indirect object data, the low number of indirect objects in the corpus and the low overall tendency for topicalization means that data sparseness is likely to become a real problem in more detailed analysis.

4.3 Definiteness

In Section 2.6.2, I formulated expectations about the effect of definiteness of nominal constituents on Vorfeld occupation. Under the hypothesis that definiteness globally exerts influence on word order, we can expect that indefinite full NPs do not front often, that definite full NPs are somewhat more likely to front, and that pronouns front most frequently. These expectations are based on the behaviour of NPs in the Mittelfeld of German and Dutch, where pronouns tend to be realized early, and indefinite full NPs tend to be realized late (see Section 2.6.2 for discussion and references).

I also hypothesized that we see a second, partly conflicting trend in Vorfeld occupation. This trend would be due to the status of the Vorfeld as a position for important material, in accordance with Gundel's (1988) first-things-first principle. What exactly counts as important material and what does not is hard to establish, especially in a corpus study such as this. However, on the basis of the relation between predictability of mention and importance (Givón, 1988; Gundel, 1988), and on the basis of existing results on the interpretation of personal and demonstrative pronouns, I hypothesized that reduced personal pronouns show infrequent fronting because their predictability of mention is high. Section 4.3.1 gives details of the way I categorized NPs into the different definiteness levels. Section 4.3.2 presents the corpus results.

The prediction that reduced pronouns front less often than full personal pronouns and especially demonstrative pronouns applies across the board. For objects, this trend would not be surprising: In general, reduced pronoun objects are ungrammatical in the Vorfeld. However, we expect to see the effect with subjects, too. For German, Gärtner and Steinbach (2003) also claim that reduced personal pronoun subjects front less frequently than full personal pronouns and demonstrative pronouns. They propose prosodic reasons for this effect. Section 4.3.3 investigates whether there is any empirical support for the alleged behaviour of reduced personal pronoun subjects in Dutch.

4.3.1 Operationalizing definiteness

In order to begin to answer the questions of Section 2.6.2, we need at least to be able to categorize nominal subjects and objects into pronoun, definite full NP and indefinite full NP, and the pronominal data into demonstrative and personal pronouns. Table 4.7 gives

Table 4.7: Operationalization of definiteness.

Level	Form	Description and examples
indefinite full NPs		
	indefinite determiner	NPs with a indefinite determiner. *een X* 'a X', *drie Xen* 'three Xs', *veel Xen* 'many Xs'
	bare noun	determinerless NPs with a noun as head. *geluid* 'sound', *mensen die dat willen* 'people.PL who want this'
definite full NPs		
	definite determiner	NPs with a definite determiner. Strong quantifiers. *de/het X* 'the X', *alle Xen* 'all Xs'
	proper name	proper names, possibly modified. Titel+PN. *Benno Baksteen*, *minister Banyamwabo* *Beatrix, vorstin der Nederlanden* 'B., queen of the Netherlands'
pronouns		
	demonstrative	demonstrative pronouns, possibly modified. *die* 'they/them/those', *die van hiernaast* 'the people/things next door'
	personal	personal and reflexive pronouns, possibly modified. *ik/jij/hij/zij* 'I'/'you.NOM'/'he'/'she', *zij die dat willen* 'they who want this'

details of a form-based operationalization of these distinctions. In the case of full NPs, one way to determine their definiteness level is to look at the determiner (the indefinite and definite forms in Table 4.7). When there are no determiners, definiteness level is determined by the head noun: Proper names are definite full NPs, any other NPs are indefinite full NPs (that is, bare nouns). Pronouns can be personal or demonstrative pronouns. The distinction between full and reduced pronouns will only be made for subjects, in Section 4.3.3.

I will report counts and proportion in the Vorfeld for the three definiteness levels, and I will also give details for each of the six NP form distinctions. The distinction between pronominal forms will play an important role, because demonstrative pronouns are both pronoun *and* important information, whereas personal pronouns may not be important information. Demonstrative pronouns should therefore front more often than any other type of nominal constituent. With respect to definite full NPs, researchers like Aissen (2003) place proper names higher on the definiteness scale than other definite full NPs. We may be able to see this difference reflected in the Vorfeld behaviour.

Giving separate counts for bare nouns is motivated by shortcomings in the operationalization outlined above. In some cases, bare nouns should arguably be classed as definite full NPs. For instance, some bare nouns are used to express generics, and it can be argued that they should be treated as definite NPs. This holds for generically used NPs with an indefinite determiner, too, but these cases are not as common. We have no way of distinguishing generically used bare nouns from other bare nouns. Another case of possible definite full NPs are proper name-like nouns and titles (like *mama* 'mum', *dokter* 'doctor', *meneer* 'mister/sir'). In the CGN, these are marked up as common nouns, and the scheme of Table 4.7 therefore classes them as bare nouns. However, if they are used as proper names, they should ideally be classified as definite full NPs. Again, we have no way of distinguishing these uses automatically. Further situations in which bare nouns really seem to be definite are coordinations, as discussed in Heycock and Zamparelli (2003). For instance, in (25), the mother and child referred to in the subject (boldfaced) have been explicitly mentioned in the preceding discourse (not shown), and should hence be considered definite.

(25) **moeder en kind** zijn nog altijd weg volgens mij
mother and child are PART PART gone according to me
'Mother and child are still missing, I think.' (Nl-a 458:207)

More broadly, the availability of bare noun realization may be tied to the constituent appearing in the Vorfeld. Hoeksema (2000) points out that article drop is allowed when a nominal argument is in the Vorfeld, even when a postverbal realization does not allow dropping the article. This may happen with definite as well as indefinite full NPs. An example from the CGN is given in (26a), where the bare noun subject is in boldface. The constructed alternatives (26b) and (26c) show that article drop is not allowed when the subject is in the Mittelfeld, and that either a definite or an indefinite article may have been dropped.

(26) a. **nadeel** is dat heel vaak de schets mooier getekend is
disadvantage is that very often the sketch nicer drawn is
'Disadvantage is that the sketch very often is nicer drawn.' (Nl-j 7333:129)
b. dat heel vaak de schets mooier getekend is, is *(een/het) **nadeel**
c. (een/het) **nadeel** is dat heel vaak de schets mooier getekend is

The coupling of Vorfeld positioning and the availability of article drop may have as a result that bare noun fronting will be overestimated. The calculation of Vorfeld proportion per definiteness level assumes that Vorfeld placement itself does not influence definiteness level. Bare nouns like in (26a) end up in the bare noun group, even though they could be considered Vorfeld instances of definite/indefinite determiner NPs, like in (26c). To summarize, the bare noun data may not only contain bare noun indefinite NPs, as intended

Table 4.8: Definiteness summary, arguments combined.

		Vorfeld		Prop est (%)		
Level	Form	yes	no	lo	pt	hi
indefinite full NPs	indef. det.	763	5 736	11.0	**11.7**	12.5
	bare noun	818	1 941	27.9	**29.6**	31.4
	Total	1 581	7 677	16.3	**17.1**	17.9
definite full NPs	def. deter.	4 263	4 507	47.6	**48.6**	49.7
	proper name	1 774	847	65.9	**67.7**	69.5
	Total	6 037	5 354	52.1	**53.0**	53.9
pronouns	demonstrative	13 746	2 492	84.1	**84.7**	85.2
	personal	24 677	17 895	57.5	**58.0**	58.4
	Total	38 423	20 387	64.9	**65.3**	65.7
Rest		215	401	31.1	**34.9**	38.8
Total		46 256	33 819	57.4	**57.8**	58.1

in the classification, but it may also contain material that is definite-like (generics), or truly definite (the coordination examples). Finally, bare nouns in the Vorfeld may be the result of article drop in definite or indefinite full NPs. Each of these cases will lead to a boost of the proportion of bare nouns that appear in the Vorfeld compared to the average indefinite full NP.

Nevertheless, using the operationalization of definiteness given in Table 4.7, we can automatically classify the NPs in the corpus, and investigate to what extent the expectations formulated Section 2.6 are met in the corpus. In discussing the results, I will try to take the shortcomings of the operationalization into account.

4.3.2 Corpus results

The Vorfeld behaviour of nominal subjects and objects combined is summarized in Table 4.8. By summing the Vorfeld and elsewhere counts (sums not shown in table), we can see that, of the 80k nominal subject/object constituents in the dataset, the majority – nearly 60k – are pronominal. There are only slightly more definite full NPs than indefinite full NPs (11k vs 9k).

We can see a neat increase in fronting frequency between the three definiteness levels in Table 4.8: Around 17% of the indefinite full NPs appear in the Vorfeld, around 53% of definite full NPs, and 65% of pronouns. On the six-point scale formed by the NP

Table 4.9: Association between grammatical function and definiteness (PMI).

	Grammatical function		
Definiteness level	subject	direct object	indirect object
indefinite full NP	-1.4245	1.6718	-2.2165
definite full NP	-0.1777	0.4881	-0.1777
pronoun	0.1651	-0.7912	0.1966

Note: All PMIs are based on cell counts above 500, except for indefinite full NPs (18) and definite full NPs (91) in the indirect object row. The total table count is 79459.

forms, the Vorfeld percentages also steadily increase from indefinite determiner NPs (11%) to demonstrative pronouns (84%). Personal pronouns, however, show a reduced Vorfeld percentage (58%). The combined data clearly shows the predicted complex effect: Constituents higher on the definiteness scale front more often, except for personal pronouns,

Of course, the data in Table 4.8 hides an important confound: grammatical function. In Table 4.9, we can see how grammatical function is correlated with the three definiteness levels. The numbers in the table give a measure of association known as *pointwise mutual information* (PMI).[8] PMI is an indication of what we learn about the value of one variable when we know the value of the other. For instance, if we know that the constituent in question is a subject, Table 4.9 tells us that the chance that the constituent is also a pronoun has increased compared to the average chance of seeing a pronoun (positive PMI). The measure is fully symmetrical, so we can learn about definiteness on the basis of grammatical function, and we can equally can learn about grammatical function on the basis of definiteness. A positive PMI means that a combination is favoured, a negative PMI means that a combination is disfavoured. A PMI of zero means that the chances of seeing a combination of values are exactly what we would expect if the variables were not associated.

Table 4.9 shows that the association between the variables definiteness and grammatical function is such that the average effect of definiteness in Table 4.8 might be explained by grammatical function. I will ignore indirect objects: There are so few of them that they will not affect the average that much. When we move up the definiteness scale, the mix of subjects and direct objects changes. There is an increasing amount of subjects in the data, and a decreasing amount of direct objects. Since subjects front often, and direct objects do not, we will see that the average constituent fronts more often than a constituent that

[8]Pointwise mutual information (PMI) in Table 4.9 is calculated as follows.

(i) $\mathrm{PMI}(gf, df) = \log_2 \frac{P(gf \wedge df)}{P(gf) \times P(df)}$

where gf and df are values of the variables Grammatical function and Definiteness level.

is lower on the definiteness scale, even if there is no difference between the definiteness levels within subjects and direct objects. The obvious way to investigate whether this is the case is to separate the data into subjects, direct objects and indirect objects, and to look at the Vorfeld proportions per definiteness level in each group.

Before I will turn to a discussion of the breakdown of definiteness data, I would like to point out that Table 4.9 has linguistic relevance besides being a motivation for more detailed scrutiny of the data. Aissen (2003) noted that being a subject and being highly definite is an unmarked situation. She also argues that we should consider what is unmarked for subjects to be marked for objects. This is referred to as *markedness reversal*. Markedness reversal can be observed for subjects and direct objects in Table 4.9. If we start low on the definiteness scale in the subject column and move up, we see increasingly positive PMIs: We move towards an unmarked combination. In contrast, if we start *high* on the definiteness scale in the object column and move down, we see increasingly positive PMIs. The corpus data reflects that what is marked for a subject in terms of definiteness is unmarked for an object. Interestingly, we do not see markedness reversal between subjects and indirect objects. The relation between definiteness and indirect objects is very similar to the relation between definiteness and subjects. In Chapters 5 and 6, I will use the definiteness markedness reversal between subjects and direct objects. In those chapters, I investigate whether the recognizability of constituents as subjects or objects influences word order, and I will use definiteness as one of the sources of information about the grammatical function of a constituent. Now, however, it is time to look at definiteness and Vorfeld occupation in more detail.

Subjects Table 4.10 shows how Vorfeld occupation of subjects varies across definiteness levels. On the 6-point scale of NP forms, the data looks very similar to the overall picture in Table 4.8, apart from the fact that the percentages are consistently higher. NPs with an indefinite determiner have the lowest chance of appearing in the Vorfeld (43%), demonstrative pronouns the highest (92%). Compared to demonstrative pronouns, the drop in the proportion of personal pronouns in the Vorfeld is considerable. Personal pronoun subjects are about as likely as bare noun subjects to appear in the Vorfeld (63% and 66%, respectively). On the 3-point definiteness scale, we see the predicted increase from indefinite full NPs (52%) to definite full NPs (74%). However, pronouns on average front slightly less often than definite full NPs (70%). The low average of the pronoun data is caused by fact that most of the subject pronouns are personal pronouns.

Inspection of the dataset shows that a good part of the indefinite subjects appears in an existential construction (EC), with *er*, *'r* or *d'r* in the Vorfeld. Non-EC sentences that have postverbal indefinite subject often have an adverbial in first position. Two examples are in (27). The subject is boldfaced.

Table 4.10: Definiteness summary for subjects.

		Vorfeld		Prop est (%)		
Level	Form	yes	no	lo	pt	hi
indefinite full NPs	indef. determiner	674	896	40.5	**43.0**	45.4
	bare noun	715	362	63.5	**66.4**	69.2
	Total	1 389	1 258	50.6	**52.5**	54.4
definite full NPs	def. determiner	4 003	1 551	70.9	**72.1**	73.3
	proper name	1 714	461	77.0	**78.8**	80.5
	Total	5 717	2 012	73.0	**74.0**	74.9
pronouns	demonstrative	11 010	962	91.5	**92.0**	92.5
	personal	24 662	13 971	63.4	**63.8**	64.3
	Total	35 672	14 933	70.1	**70.5**	70.9
Rest		195	104	59.5	**65.2**	70.6
Total		42 973	18 307	69.8	**70.1**	70.5

(27) a. dan kookt vanzelf **een ander**
then cooks by itself an other
'Someone else will cook after a while.' (Nl-b 131:352)

b. (Warning the addressee to do something about flies coming in:)
daar staat **een vliegenraam**
there.DEICT stands an insect screen
'There's an insect screen over there.' (Vl-a 400083:239)

Indefinites that are in the Vorfeld, but not in an EC, can be regular indefinites (28a) but are more commonly partitives (28b) or generics (28c).

(28) a. **twee Brakelaars** waren op elkaar ingereden
two Brakelers have collided
'Two inhabitants of Brakel collided.' (Vl-b 400165:408)

b. (A teacher says about students:)
een heleboel namen dat idee over
a great amount took that idee PRT
'Many [of the students] adopted that idea.' (Vl-a 400407:45)

c. **een flesje** is nul komma drie liter
a bottle is zero point three litre
'A bottle is 0.3l.' (Nl-a 818:204)

Table 4.10 shows that there is quite a large difference between indefinite determiner NPs

and bare nouns. The bare noun data contains mass nouns, as in (29a), and generics (29b). It may be that these types of subjects are more likely to appear in the Vorfeld, so that the difference in fronting between the two indefinite NP forms reflects underlying differences in definiteness and is not just an artifact of the classification scheme. Of course, a non-generic interpretation is possible for bare nouns. Sentence (29c) is an example of this. Note that the bare noun subject appears postverbally.

(29) a. **soep** is zestig cent of zo
soup is sixty cent or something
'Soup is something like 60 cents.' (Nl-a 535:114)

b. **lemmingen** plegen geen massaal zelfmoord
lemmings commit no collectively suicide
'Lemmings do not collectively commit suicide.' (Nl-a 610:204)

c. op de Zuidpool leven **dieren**
on the South Pole live animals
'On the South Pole there are animals.' (Nl-a 563:172)

Future research into Vorfeld occupation that takes aspects like countability and genericity into account may be able to make the relation between genericity, bare nouns and Vorfeld occupation more concrete.

Further inspection of the sentences also suggests that the difference between indefinite determiner NPs and bare nouns is boosted by the classification scheme itself. First, there is the misclassification of words like *papa* and *mama*, and also of mentioned words (for instance when their meaning is explained). Secondly, article drop in the Vorfeld is quite frequent. This holds not only for the type mentioned in the previous section, where the effect was supposedly restricted to a small set of nouns such as *disadvantage*, *fact*, etcetera, but also in very economical journalist language (headlines and sports commentaries), and in the spontaneous dialogue parts. Of the latter type, (30a) is an example.

(30) a. **keuken** wordt afgeplakt
kitchen becomes masked
'The kitchen is being masked (for painting).' (Nl-a 280:224)

b. Morgen ('tomorrow') wordt *(de) **keuken** afgeplakt.

The Vorfeld occupant *keuken* cannot appear postverbally without an article, as the reconstruction (30b) demonstrates.

Within the group of definite full NPs, proper names front more often than definite determiner NPs. As with the difference between indefinite full NP forms, there may a relevant underlying difference between definite determiner NPs and proper names, or there may be an effect of article drop. To start with the former, it may be that proper

names rank higher on a fine grained definiteness scale than other definite full NPs (see, for instance, Aissen, 2003). That proper names front more often would then reflect this difference in rank. It might also be that proper names are more often animate than other definite full NPs, and that animacy has a positive influence on fronting. There is evidence that animacy promotes early realization in the German Mittelfeld (Kempen and Harbusch, 2004). The line of reasoning that I have followed when I formulated the expectations for the corpus research is that word order factors in the Mittelfeld surface as statistical trends in Vorfeld occupation. As a result, we could speculate that the animacy effect on Mittelfeld word order surfaces as a positive influence on Vorfeld occupation. On the other hand, the difference between proper names and definite determiner NPs could be explained by article drop, too. The example in (30) shows that a postverbal NP may be a definite determiner NP (30b), while the preverbal counterpart is a bare noun. This does not apply to proper names, which do not have an article to drop when they are in first position. As a result, definite determiner NPs may appear to front less often than proper names, because some of the definite determiner NPs 'become' bare nouns when they front.

The approach to investigating the effect of definiteness on the Vorfeld does not allow us to solve the problem that Vorfeld occupation has an effect on the NP form classification. I will therefore not be able to do much more than point out that the statistics may be polluted by this effect: Vorfeld occupation by indefinite determiner NPs and definite determiner NPs may be systematically underestimated and Vorfeld occupation by bare nouns overestimated. However, it should be emphasized that article drop affects the interpretation of the frequency differences between indefinite determiner NPs and bare nouns, and between definite determiner NPs and proper names, but not the interpretation of the frequency difference between indefinite and definite full NPs. In spite of the presumably overestimated Vorfeld occupation of bare nouns, definite full NPs still front more often than indefinite full NPs.

I will look into the behaviour of subject pronouns in more detail in Section 4.3.3.

Direct objects Table 4.11, p112, shows the relation between definiteness and Vorfeld occupation in the direct object data. The general trends are the same as in the subject data. Topicalization becomes more frequent as we move up on the six-point NP form scale and on the three-point definiteness hierarchy. The exception in this trend is again formed by personal pronouns, which have a lower tendency to topicalize. The direct object data differs in two respects from the subject data. First, at each level, the percentages are lower. This is in line with our expectation that direct objects inherently front less often than subjects. Secondly, the difference between demonstrative pronouns and personal pronouns is dramatic compared to the subject data. Personal pronoun direct objects have an almost 0% chance of appearing in the Vorfeld, whereas demonstrative pronoun

Table 4.11: Definiteness summary for direct objects.

		Vorfeld		Prop est (%)		
Level	Form	yes	no	lo	pt	hi
indefinite full NPs	indef. determiner	86	4 833	1.4	**1.7**	2.2
	bare noun	102	1 572	5.0	**6.1**	7.3
	Total	188	6 405	2.5	**2.9**	3.3
definite full NPs	def. determiner	255	2 897	7.2	**8.1**	9.1
	proper name	57	362	10.5	**13.6**	17.3
	Total	312	3 259	7.8	**8.7**	9.7
pronouns	demonstrative	2 723	1 527	62.7	**64.1**	65.6
	personal	4	3 342	0.0	**0.1**	0.3
	Total	2 727	4 869	34.8	**35.9**	37.0
Rest		20	292	4.0	**6.4**	9.7
Total		3 247	14 825	17.4	**18.0**	18.5

direct objects topicalize more often than not. The fact that definiteness is correlated with Vorfeld occupation in similar ways in the subject and direct object data means that the difference in Vorfeld occupation between subjects and direct objects observed in Section 4.2 cannot be explained by definiteness and that the average relation between definiteness and Vorfeld occupation cannot be explained by grammatical function. Rather, grammatical function and definiteness each influence a constituent's chance of becoming a Vorfeld occupant.

Let us walk through the data with some examples. Like in the subject data, the bare noun Vorfeld data contains cases of article drop, such as (31). The sentences (31a) and (31c) cannot be reconstructed as canonical word order sentences without inserting an article, (31b) and (31d), respectively.

(31) a. **visrestaurant** vind ik iets te riskant
fish restaurant find I somewhat too risky
'I find (going to) a/the fish restaurant a bit too risky.' (Vl-a 400089:162)

b. Ik vind *(een/het) **visrestaurant** iets te riskant (constructed)

c. **achternaam** noemen we niet
last name mention we not
'We will not mention the last name.' (Nl-a 714:117)

d. We noemen *(de) **achternaam** niet. (constructed)
e. We noemen geen **achternaam**. (constructed)

The constructed variant in (31e) shows that sentence negation may be incorporated in the object NP in Dutch. Sentence negation in (31e) is expressed in the object NP by *geen* 'no'. Incorporation of negation is one more reason why the Vorfeld proportion of bare nouns may be overestimated. Consider the contrast in (32). When the direct object is in the Vorfeld, as in the attested (32a), it is a bare noun indefinite NP. However, when it is in the Mittelfeld (32b), the direct object incorporates negation, and the result is an indefinite determiner NP.

(32) a. **winterwortels** hebben we hier niet
winter carrots have we here not (Nl-f 7255:1)
b. We hebben hier geen **winterwortels** / *niet **winterwortels.**
We have here no carrots / not carrots.
'We do not have winter carrots here.' (constructed)

The effect of negation incorporation on the statistics is the same as the effect of article drop – the bare noun Vorfeld proportions are overestimated. In this case, however, the reason is not that the bare noun count in the Vorfeld is inflated, but rather that the bare noun count in the postverbal domain is in some sense lower.

Inspection of the definite determiner NP data and the indefinite determiner NP data reveals that there are large differences within these NP forms. Let us start with the indefinite determiner NPs. The effect of negation incorporation sketched above is clearly seen by the fact that of the 402 indefinite full NPs that have *geen* 'no' for a determiner, only 2 appear in the Vorfeld (0.5%). The large number of postverbal *geen*-NPs is due to negation incorporation. For comparison, of the 1 946 NPs with the indefinite article *een* 'a' as determiner, 31 (1.5%) are in the Vorfeld. An extreme on the positive side is *zo'n/zulke* 'such a'/'such': 9.8% (15/153) of the occurrences are in the Vorfeld. Although I have classified *zo'n/zulke* NPs as indefinites, they have a clear anaphoric component. This definite-like property of *zo'n/zulke* NPs may be the reason that they front so frequently. The definite determiner NPs object also show differentiation. There are striking differences between NPs with a definite article and NPs with a demonstrative determiner. Of all NPs with definite articles *de/het/'t* 'the', 5.6% (86/1521) appear in the Vorfeld. In contrast, 18.7% (109/582) of object NPs with *die/dat/deze/dit* 'that/this' appear in the Vorfeld. The difference between NPs with a (regular) definite article and NPs with a demonstrative determiner parallels the difference between personal pronouns and demonstrative pronouns. I consider the behaviour of definite full NPs with a demonstrative determiner further evidence of the first-things-first nature of the Vorfeld.

As Gundel (1988) points out, first-things-first also motivates focus topicalization, because it puts the important information (focus) before the unimportant, predictable

information (background). Indeed, focus topicalizations can be found in the data. Sentence (33) is an example.

(33) **twee kindjes** heeft ze
two children.DIM has she
'She has two small children.' (Nl-a 773:161)

Jansen (1981) has found focus topicalization to be rare in Dutch. Because I do not systematically investigate the information structure of sentences in the corpus, I cannot deny or confirm this claim.

One of the corpus research questions raised in Chapter 2 was whether reduced object pronouns are acceptable in the Vorfeld in (non-standard) Dutch. Weerman (1989) claims that object *'t* 'it' in the Vorfeld is acceptable. In addition, for German, Gärtner and Steinbach (2003) suggest that one has a greater chance of coming across topicalized reduced object pronouns in colloquial or dialect data. If so, a spoken language corpus with data from different regions would be a good place to find them.

In the direct object data in Table 4.11 there are only four cases of topicalized personal pronouns. One of these actually is a reduced personal pronoun. The sentence, with a bit of context included, is in (34).

(34) (Two men discuss the costs of an extra window in the roof, which involved professionals installing leaden sealing)
A So, who payed for that?
B Well, nothing – those guys.
A So they put those lead strips on there for free?
B Yeah, of course. It's all charged to the association of owners.
A dus **'t** heeft de veRENiging betaald
so it.RED has the association payed
'So the OWNers' association payed for it [i.e., it wasn't free].' (Nl-a 320:512)

The postverbal subject *de vereniging* is realized with a prominent late rise. The subject is in focus because of the contrast with the implicit proposition that nobody payed for the work. Note that the reduced pronoun in question is *'t* 'it', claimed to be grammatical in the Vorfeld by Weerman (1989).

Informally asking some informants showed that not everybody accepted the example in (34). Some informants called it ungrammatical or awkward, some immediately corrected it to the neuter demonstrative *dat*, and some even seemed to hear *dat*. Listening to the recording has convinced me, however, that only *'t* is pronounced. The fact that people will hear *dat* even when only *'t* is pronounced is actually problematic: If corpus annotation is susceptible to the same auditive illusion, it might be that other instances of Vorfeld object *'t* exist, but that they have been annotated as *dat*. The fact that annotation was

semi-automatic makes this a little less likely, but not impossible. As a consequence, working at the annotation layer that I am using in this dissertation might not be the right approach if one wants to find exceptions to the 'rule' that reduced object personal pronouns cannot appear in the Vorfeld. Alternatively, one could argue that it really is a (phonetically) reduced *dat*, although, as Gärtner and Steinbach (2003) point out, this raises the questions of why other demonstratives could not receive the same treatment and what the prerequisites for such reduction are. All in all, I am not convinced about the general availability of object *'t* in the Vorfeld. Only one example turned up in the corpus, which was furthermore not a well accepted one. However, there is the possibility that one has to search at annotation levels below lexical annotation to find more examples. Looking at these lower levels – phonological transcription, for instance – is a topic for further research.

A more surprising result than the lack of good examples of reduced topicalized personal pronouns is the finding that full personal pronouns do not topicalize frequently either. The remaining 3 personal pronouns of the 4 fronted pronouns in the data are full pronouns of first or second person. An example is (35).

(35) **jou** moest ik hebben
you must.PST I have
'You I was looking for.' (Nl-a 411:39)

This result is surprising because full pronouns in the Vorfeld are generally considered to be fully grammatical in the literature. If they are, one wonders why they are so rare in spoken Dutch. In this context, it is probably not a coincidence that the observed cases of full personal pronouns in the Vorfeld are first and second person. A third-person pronoun in the Vorfeld sounds somewhat marked. Consider the example in (36b), which is constructed by replacing the demonstrative pronoun in the Vorfeld of the attested (36a).

(36) a. (Talking about people the speakers know with epilepsy)
die had ik ook 'ns een keer aangetroffen toen ze een aanval had
DEM had I too PART found when she an attack had
'Her, too, I once found having a fit.' (Nl-a 773:106)
b. **haar** had ik ook 'ns een keer aangetroffen... (constructed)

The use of a full personal pronoun in the Vorfeld in (36b) sounds stilted compared to the demonstrative pronoun in (36a). It may be that spoken Dutch favours the use of demonstrative pronouns over stressed or otherwise not-reduced personal pronouns. The finding that the full pronouns that do appear in the Vorfeld are first and second person can be attributed to the fact that a demonstrative pronoun is third person only. For first and second person pronouns, the option of using a less marked form does not exist. However, a better understanding of the use of demonstrative pronouns in Dutch discourse and

Table 4.12: Definiteness summary for indirect objects.

		Vorfeld		Prop est (%)		
Level	Form	yes	no	lo	pt	hi
indefinite full NPs	indef. deter.	3	7	6.7	**30.0**	65.2
	bare noun	1	7	0.3	**12.5**	52.7
	Total	4	14	6.4	**22.2**	47.6
definite full NPs	def. deter.	5	59	2.6	**7.8**	17.3
	proper name	3	24	2.4	**11.1**	29.2
	Total	8	83	3.9	**8.8**	16.6
pronouns	demonstrative	13	3	54.3	**81.3**	96.0
	personal	11	582	0.9	**1.9**	32.9
	Total	24	585	2.5	**3.9**	5.8
Rest		0	5	0.0	**0.0**	52.0
Total		36	687	3.5	**5.0**	6.8

the relation between demonstrative pronouns and the Vorfeld is required before we can conclude that the speculations I give here are on the right track.

Indirect objects Finally, Table 4.12 gives the indirect object data. The data sparseness makes it difficult to make solid claims about the effect of definiteness on Vorfeld occupation in the indirect object data. At first sight, it would appear that there is an anti-definiteness effect: Indefinite full NPs topicalize more frequently than definite full NPs. However, the extremely wide confidence intervals show that no such conclusion is warranted. The proportion of topicalized indefinite full NPs may be as low as 6.4% and the proportion of topicalized definite full NPs may be as high as 16.6%.

Nevertheless, conclusions can be drawn about the difference between demonstrative pronouns and definite full NPs. Demonstrative pronouns topicalize more frequently than definite full NPs. This is in line with the results for subjects and indirect objects.

The data in Table 4.12 sheds new light on the average difference between direct objects and indirect objects we observed in Section 4.2. Overall, 18% of the nominal direct objects appeared in the Vorfeld versus only 5% of the indirect objects. This was the opposite of what was hypothesized on the basis of canonical argument order in Section 2.6. Yet, if we compare Table 4.11 (direct objects) and Table 4.12 (indirect objects), we see that the large overall difference is caused by the fact that only a very small percentage of the indirect objects falls in the demonstrative pronoun category. If we try to factor out the effect of definiteness, there is no evidence that direct objects topicalize more often than

indirect objects. Between 2.5% and 3.3% of the indefinite full NPs are topicalized in the direct object group, whilst between 6.4% and 47% are topicalized in the indirect object group. Of definite full NPs, 7.8%–9.7% are topicalized in the direct object group, versus 3.9%–16.6% in the indirect object group. For full NPs, there is no evidence that being a direct object has a positive influence on topicalization in the full NPs. In fact, for the indefinite full NPs, indirect objects appear to have a slightly higher tendency to appear in the Vorfeld. If we look at the pronoun data, we can conclude that there is no evidence for a difference between demonstrative pronoun direct objects and demonstrative pronoun indirect objects (62.7%–65.6% versus 54.3%–96.0%). In the personal pronoun group, indirect objects again appear to topicalize slightly more often: direct objects (0.0%–0.3%), indirect objects (0.9%–32.9%). We should not read too much into these numbers, but we can conclude that the large average difference between direct and indirect objects is primarily caused by the difference in distribution of NP form. In Section 4.5, I will come back to this issue, and I will show that, given some reasonable assumptions about the effect of definiteness on Vorfeld occupation, it is likely that indirect objects in fact front slightly more often than direct objects do.

Finally, let us consider the case of reduced personal pronouns in the Vorfeld. If indirect objects really have an inherently higher rate of topicalizing than direct objects, our chances of finding a reduced topicalized pronoun might be better in the indirect object group. However, none of the indirect object personal pronouns in the Vorfeld is reduced. Data inspection also shows that the 11 full personal pronouns that appear in the Vorfeld are either objects in an impersonal passive (37a) or dative experiencers (37b).

(37) a. **haar** werd verteld gewoon dat bepaalde dingen aan haar zouden liggen
her was told PART that certain things were due to her
'She was told that she should blame herself for certain things.' (Nl-a 935:93)

b. **mij** boeit dat helemaal niet
me binds that totally not
'I am completely uninterested in that.' (Nl-a 679:261)

Recipient indirect objects in the Vorfeld are demonstrative pronouns, as in (38).

(38) ja **die** moeten we ook water geven morgen
yes DEM must we too water give tomorrow
'Yes, we should water it [a plant] tomorrow, too.' (Nl-a 594:67)

Of the 13 demonstrative indirect pronouns in the Vorfeld, 8 realize recipients. Investigation of a larger indirect object dataset will have to show if there is a systematic difference in the definiteness distributions of the thematic roles, and whether the thematic roles show different fronting behaviour.

There is a clear effect of definiteness on fronting in the subject data and the direct object data. Indefinite full NPs in both groups front relatively rarely. Definite full NPs front more often. This is in line with behaviour of indefinite full NPs in the Mittelfeld and is therefore support for the thesis that there is a global tendency to realize definite information early in the sentence. The indirect object data might be compatible with this trend, too, although there is too little data to actually observe this trend for indirect objects.

The interpretation of the pronoun data is more complicated. For all three grammatical functions, demonstrative pronouns show a greatly increased chance of appearing in the Vorfeld compared to both definite full NPs and personal pronouns. In contrast, personal pronouns seem to 'shy away' from the Vorfeld. In the subject data, personal pronouns are less likely to front than definite full NPs, and in the object data, personal pronouns are less likely to front than indefinite full NPs. This complex effect can be seen as the result of a combination of the global trend to realize material high on the definiteness scale early and the special properties of the Vorfeld as a position for important material (first-things-first).

4.3.3 Pronouns in the Vorfeld

Throughout the discussion about definiteness data, we have seen that pronominal NPs played a special role. Thus far, we have made the following observations:

- Personal pronouns have a relatively low chance of appearing in the Vorfeld, whereas demonstrative pronouns have a high chance of appearing in the Vorfeld.
- The difference between demonstratives and personal pronouns is more pronounced for objects than for subjects.
- There is only one case of a reduced personal pronoun in the Vorfeld in the object data.

Let us summarize the first and third point in the following schematic way (39). The two scales represent the negative effect of pronominal form on the chance of appearing in the Vorfeld. An prominal type on the scale appears less often in the Vorfeld than a type to its right:

(39) a. Subjects: personal $\succ$ demonstrative
b. Object: reduced $\succ$ full $\succ$ demonstrative

These scales are meant to be a rough indication of fronting behaviour. The fact that the difference between demonstrative and personal pronouns is more pronounced in the object data (60% versus $<$1%) than in the subject data (90% versus 60%) is not described by them.

The clear finding that demonstratives front most frequently in both the subject and object data is explained by appealing to the first-things-first principle. Personal pronouns

tend to realize predictable material, and are therefore not important (Givón, 1988; Gundel, 1988). The difference between full and reduced personal pronouns could also be related to the first-things-first principle. Full personal pronouns could be stressed (expressing contrast), or they might refer to less predictable material compared to reduced personal pronouns – in either case, the full personal pronouns would qualify as more important information than reduced pronouns.

However, we can ask a further question: Do personal pronoun subjects form a homogeneous group? On the basis of the first-things-first principle, we expect that this is not the case. The preference of the Vorfeld for important material applies across the board and therefore also extends to subjects. We can already observe its effect in the subject data, in the difference between demonstrative and personal pronouns. If the first-things-first principle is also what underlies the difference between full and reduced pronouns in (39), the subject data should show a similar split. Gärtner and Steinbach (2003) also claimed that reduced personal pronoun subjects appear less frequently in the Vorfeld than full personal pronoun subjects. Gärtner and Steinbach give a prosodic explanation. According to Gärtner and Steinbach, a reduced pronoun forms one prosodic word with the finite verb (cliticization). Because encliticization is preferred over procliticization in German and Dutch, reduced pronouns are best placed after the verb. Full personal pronouns and demonstrative pronouns do not cliticize, so they are not subject to a prosodic constraint on cliticization.

The prediction that reduced pronoun subjects do not like to appear in the Vorfeld is one that is hard to test using intuition. A reduced pronoun subject in the Vorfeld is grammatical and does not 'feel' marked at all. But, if reduced pronoun subjects have a preference for the Mittelfeld over the Vorfeld, it is reasonable to expect to see a quantitative difference. To investigate this prediction, I split the personal pronoun subject data into groups of full and reduced pronouns. This cannot be done for all pronouns, since some personal pronouns only have full forms (for instance, *jullie* 'you.PL'), and other personal pronouns only have reduced forms (*het/'t*, 'it'). See Table 2.2, p32, for the complete paradigm. The result of splitting the personal pronoun subject data is given in Table 4.13.

Table 4.13 shows that, by and large, full personal pronoun subjects show a stronger tendency to front than reduced ones. The strongest difference between the full and reduced form is found in the 3.SG.M column. The form *hij* 'he.FULL' has a demonstrative-like tendency to appear in the Vorfeld, whereas the forms *-ie* and *'m* 'he.RED' are tied to the postverbal domain like real clitics. The difference between *gij* 'you.FULL'and *ge* 'you.RED' is only small, and not significant. For all but one of the reduced/full pairs in Table 4.13 the prediction about the reduced/full distinction is borne out. This part of the pronominal subject data is compatible with both Gärtner and Steinbach's prosodic account and the first-things-first explanation.

Table 4.13: Distribution of full and reduced forms of subject personal pronouns.

		Paradigm cell					
Type		1.sg	2.sg	2(.sg)	3.sg.m	3.sg.f/3.pl	1.pl
full	#	*ik* 12 437	*jij* 440	*gij* 154	*hij* 1 694	*zij* 378	*wij* 1 019
	%Vf	**68.1**	**71.6**	**57.8**	**90.2**	**78.8**	**68.6**
reduced	#	*'k* 3 607	*je* 4 870	*ge* 910	*-ie/'m* 750	*ze* 3 127	*we* 3 184
	%Vf	**78.2**	**44.3**	**47.6**	**0.0**	**54.4**	**48.0**

Note: Percentage in Vorfeld boldfaced. All differences in Vorfeld occupation between reduced and full pronouns significant at $p < .001$ (2-t Fisher's), except *gij/ge*, not significant ($p = .22$).

The notable exception in this data is the first person singular subject *ik/'k*, which is deviant in two ways. First, the number of full forms *ik* is consistently higher than the number of reduced forms *'k*. For the other pronouns, the reduced form is at least the most frequent form in the Mittelfeld, and for most, it is also the most frequent form in the Vorfeld. Secondly, the reduced form *'k* shows a stronger tendency to appear in the Vorfeld than the full form *ik* does. This behaviour is the opposite of the behaviour of other pronouns. We can update the scales in (39) with the information in Table 4.13. The result is (40).

(40) a. Subjects: reduced+*ik* $\succ$ full+*'k* $\succ$ demonstrative
b. Object: reduced $\succ$ full $\succ$ demonstrative

The behaviour of the first-person subjects is unexpected under the first-things-first approach, and I do not see an obvious way to explain their behaviour under this approach. However, a prosodic account may be able to accommodate the deviant behaviour of *ik/'k*: Of the pronouns in Table 4.13, *'k* is the only pronoun that does not necessarily form a syllable. The written form *'k* may be pronounced as [k]. One needs the functional explanation in terms of first-things-first to explain the difference between demonstrative pronouns and personal pronouns, but the behaviour of *'k* is a possible indication that prosodic factors influence the distribution of reduced and full forms in the Vorfeld, too.

4.3.4 Summary

Nominal subject and object constituents show consistent differences in fronting depending on definiteness. Indefinite full NPs front less frequently than definite full NPs. Pronouns show a split behaviour depending on whether they are demonstrative pronouns (which front very often), full personal pronouns (which front less frequently) or reduced personal pronouns (which front even less often). These trends are clearly observable in the subject

and direct object data, and the indirect object data is largely compatible. The definiteness effects are not equally pronounced in the subject and object data, however. The differences between demonstrative and personal pronouns and the differences between demonstrative pronouns and full NPs are smaller in the subject data than in the direct object data.

We may summarize the findings in this section by drawing two scales that partly conflict. The first scale (41a) describes the positive effect of appearing high on the definiteness scale on Vorfeld occupation. The second scale (41b) describes the dispreference of the Vorfeld for (reduced) personal pronouns.

(41) a. *Positive relation between definiteness and Vorfeld occupation:*
pronoun $\prec$ definite full NP $\prec$ indefinite full NP
b. *Negative relation between pronominal form and Vorfeld occupation:*
reduced personal pronoun $\succ$ full personal pronoun $\succ$ demonstrative pronoun

The definiteness scale in (41a) has been observed in word order in the Dutch and German Mittelfeld. The fact that its effects can been seen in the Vorfeld, too, supports the hypothesis that there is a global tendency to realize material that is high on the definiteness scale early in the sentence.

The pronominal form scale in (41b) can be related to the first-things-first principle of Gundel. Elements to the left of the scale realize increasingly predictable/unimportant information, and they are therefore increasingly less suitable Vorfeld occupants. This tendency is so strong that personal pronouns do not adhere to the definiteness scale of (41a), and front less frequently than definiteness NPs. Demonstrative pronouns are highest on the definiteness scale (pronouns), and they are important material: a combination of properties that makes them extremely suitable as Vorfeld occupants. An exception to the pronominal form scale was formed by the first person singular subjects, however. In this group, the reduced form fronts more often than the full form. We may take this as an indication that prosody also influences the relation between pronominal form and Vorfeld occupation (Gärtner and Steinbach, 2003).

In sum, the complex relation between definiteness and Vorfeld occupation is the result of a combination of a global word order trend and a specific property of the Vorfeld.

The study of the effects of definiteness on fronting has also taught us more about the effect of grammatical function on Vorfeld occupation. The large difference in fronting that we could observe between direct objects and indirect objects disappears if we take NP form into account. The low overall tendency of indirect objects to appear in the Vorfeld is explained by the fact that there are very few demonstrative pronoun indirect objects, whereas demonstrative pronouns are relatively frequent in the direct object data. Thrift (2003) claimed that demonstrative pronoun indirect objects are ungrammatical in Dutch. I would not go this far, but the corpus data do certainly suggest that they are marked.

4.4 Grammatical complexity

We have seen that Vorfeld occupation is influenced by grammatical function and definiteness. In this section, I will discuss a third factor: grammatical complexity. In Section 2.6.3, I reviewed previous results on the relation between word order and grammatical complexity. On the basis of this discussion, I proposed to investigate the hypothesis that there is a sentence-wide, general trend in Dutch to put syntactically simple material before complex material (cf. the *complexity principle* of Haeseryn et al., 1997). Since the Vorfeld comes early in the sentence, the effect of this tendency should be that the Vorfeld contains on average less complex material than the postverbal domain.

We have already seen some corpus evidence for the effect that is predicted by the complexity principle. In Section 4.2, we could observe that clausal constituents (which are by definition complex) front less often than nominal constituents (which are simple) in the subject and in the direct object data. In this section we will see that the effect of complexity on Vorfeld occupation is not only observed as a contrast between clausal and nominal constituents, but can also be seen within the nominal, verbal and clausal categories. However, I will also show that the conclusion that there is a direct effect of complexity on Vorfeld occupation is not warranted. Rather, the complexity effect on Vorfeld occupation is a side effect of the tendency to put complex material at the right periphery in Dutch.

I have not yet addressed the question of how complexity should be measured. Previous work on the relation between linear order and complexity has shown that complexity has a gradual effect (Wasow, 2002; Hawkins, 2004). A binary distinction like NP versus CP is therefore not a suitable measure of complexity. In his work on the English postverbal domain, Wasow (2002) investigates several measures of complexity: node count (in the syntactic tree), phrasal node count, the number of words, the presence or absence of relative clauses in an NP, etcetera. He finds that a) the data suggests a gradient measure and b) the gradient measures that he considers correlate to the extent of being practically interchangeable. I will therefore use the number of words in a constituent as a measure of its complexity or weight. Counting words has the advantage of being easy, theory-neutral, and robust. The latter two points mean that peculiarities of or errors in the CGN are less likely to have a big influence. Note that, like Hawkins and Wasow, I am concerned with *grammatical complexity*. There is no reason to expect that measures at a lower level, for instance counting syllables, phonemes, etcetera, will yield interesting results over and above the results that are brought forward by looking at the number of words.

4.4.1 Corpus results

Table 4.14 compares constituent complexity in the Vorfeld with constituent complexity in the postverbal domain. In contrast to factors investigated in previous sections, I will

Table 4.14: Summary of constituent length in words, combined data.

	Vorfeld				¬Vorfeld				
Category	#	Avg	$Q_{1,3}$	Mx	#	Avg	$Q_{1,3}$	Mx	p
nominal	46 191	**1.28**	1–1	47	33 829	**1.94**	1–2	83	<.001
n ∧ ¬pron	7 811	**2.62**	1–3		13 405	**3.37**	2–4		<.001
n ∧ ¬p ∧ ≤10	7 680	**2.42**	1–3	10	12 858	**2.83**	2–3	10	<.001
verbal	95	**2.58**	1–4	15	129	**7.37**	4–9	40	<.001
clausal	284	**6.12**	4–7	27	4 483	**9.09**	5–11	71	<.001

Note: # raw counts, *Avg* mean of lengths, $Q_{1,3}$ first and third quartile, *Mx* maximum length (minimum length is 1, except in the clausal categories, where it is 2), *p* result of 2-tailed Wilcoxon rank sum test on length of constituents in the Vorfeld vs length of constituents elsewhere.

consider the variable of complexity (constituent length in words) to be a continuous variable. Therefore, tables like Table 4.14 do not look like the contingency tables of the previous sections.

The numbers in boldface show that, in all categories, the average length in words of a Vorfeld constituent is less than the average length of a postverbal constituent. All length differences are highly significant. Note that, although it is significant, the difference in the nominal data is only of a modest size – about half a word. The first and third quartile[9] of constituent length and the maximum constituent length are provided to give an impression of the distribution of lengths in each group. In the nominal data, the quartiles reveal that at least 75% of the Vorfeld constituents are of length 1 (the third quartile ends at 1). We can safely assume that pronouns are to blame for this. To make sure that we are not simply seeing the high tendency of pronouns to appear in the Vorfeld in the overall data (65% of the pronouns in Table 4.8 appear in the Vorfeld), I have provided a summary for nominal, non-pronominal data in the pseudo-category 'n[ominal] ∧ ¬pron[ominal]'. The difference in length between Vorfeld occupants and postverbal constituents is still significant in this data. Finally, to make sure that the effect is not due to the strong skew in the data, I have included a second pseudo-category. The row 'n ∧ ¬p ∧ ≤10' shows that the length effect is persistent in the nominal data even when we only consider non-pronominal constituents of 10 words or shorter.

The large maxima in the Vorfeld (47 words for nominal constituents and 27 words for clausal constituents) suggest that there is no absolute ban on long (and thus complex) constituents in the Vorfeld, in spite of the (apparent) length effect. Inspection of the dataset shows that the 47-word nominal Vorfeld constituent is a subject. The sentence,

[9]Quartiles divide the dataset into four parts, so that (at least) 25% of the data lies between two subsequent quartiles. Between the first and third quartile lies therefore (at least) the central 50% of the data.

given as (42a), is from the political debate component of the CGN. The longest postverbal constituent (83 words) is also a subject and comes from the same component (42b). The political debates component contains speech that is formal and prepared, which may help explain the extreme lengths of the subjects. Both subjects contain multiple coordinations and CPs and PPs and can therefore be considered not only long, but also syntactically heavy by any standard. I have neither glossed nor translated the examples, but instead given a schematic indication of the structure and content of the sentence. The finite verb in second position is given in boldface.

(42) a. [$_{\text{CONJ}}$ [$_{\text{NP}}$ de toevallige omstandigheid dat de Hoge Raad een andere interpretatie gaf dan het Hof] [$_{\text{NP}}$ het feit dat de Tweede Kamer de ontstane rechtsongelijkheid zo spoedig mogelijk heeft willen wegnemen] en [$_{\text{NP}}$ het gegeven dat het kabinet buitensporig veel tijd nodig had om op de laatste motie te reageren]] **mogen** niet zoals het kabinet nu wil doen geloven tot het beeld leiden als zou er nu sprake zijn van een beperking van het gemeentelijk belastinggebied
'The circumstance of X, the fact that Y and the given that Z, should not lead to the conviction that P, contrary to the Cabinet's intentions.' (Nl-g 177:57)

b. aan de orde **zijn** [$_{\text{NP}}$ de samengevoegde interpellaties [$_{\text{PP}}$ van mevrouw Dua tot de heer De Batselier minister vice-president van de Vlaamse regering Vlaamse minister van Leefmilieu en Huisvesting over het mestactieplan] [$_{\text{PP}}$ van de heer Van Looy tot de heer Van Den Brande minister-president van de Vlaamse regering Vlaamse minister van Economie KMO Wetenschapsbeleid Energie en Externe Betrekkingen over het landbouwbeleid van de Vlaamse regering na de goedkeuring van het Sint-Michielsakkoord en de uitvoering van het mestactieplan] en [$_{\text{PP}}$ van de heer Denys tot minister De Batselier over het mestactieplan]]
'On the agenda are the combined questions of Ms D. to Mr De B. about X, of Mr Van L. to Mr Van Den B. about the Y and the execution of X, and of Mr D. to De B. about X.' (Vl-g 600014:1)

In sentence (42a), there is no obvious way of rearranging the constituents so that the subject is not in the Vorfeld, without degrading acceptability. One reason for this may be that the postverbal part of the sentence is quite complex itself. For the sentence in (42b), there is such an option. The PP *aan de orde* 'on the agenda' is a very good Vorfeld constituent, and there is no other material that would have to follow the postverbal subject. There is more to be said about the pragmatics of (42b), but before I discuss that, I will explore the reality and locus of the observed differences in length in more detail. To begin with, I will consider whether grammatical function or definiteness can explain the overall complexity effects.

Table 4.15: Summary of constituent length in words, per argument.

	Vorfeld				¬Vorfeld				
Arg./Categ.	#	Avg	$Q_{1,3}$	Mx	#	Avg	$Q_{1,3}$	Mx	p
subject									
nominal	42 917	**1.28**	1–1	47	18 368	**1.40**	1–1	83	<.001
n ∧ ¬pron	7 286	**2.64**	2–3		3 398	**3.12**	2–3		<.001
n ∧ ¬p ∧ ≤10	7 159	**2.43**	1–3	10	3 288	**2.68**	2–3	10	<.001
verbal	79	**2.62**	$1–3\frac{1}{2}$	15	22	**5.00**	$1–7\frac{3}{4}$	16	.014
clausal	206	**5.95**	4–7	19	148	**10.61**	6–13	39	<.001
direct object									
nominal	3 238	**1.28**	1–1	20	14 774	**2.67**	1–3	48	<.001
n ∧ ¬pron	513	**2.37**	1–3		9 905	**3.46**	2–3		<.001
n ∧ ¬p ∧ ≤10	509	**2.26**	1–3	10	9 469	**2.88**	2–3	10	<.001
verbal	16	**2.38**	1–4	8	107	**7.85**	4–10	40	<.001
clausal	78	**6.59**	4–8	27	4 335	**9.04**	5–11	71	<.001
indirect object									
nominal	36	**1.47**	1–1	7	687	**1.22**	1–1	13	.044
n ∧ ¬pron	12	**2.42**	$1–2\frac{1}{4}$		102	**2.46**	$1\frac{1}{4}–2\frac{3}{4}$		.748

Length by grammatical function In the investigation of the effects of definiteness on Vorfeld occupation, we saw that the grammatical functions differed radically in the distribution of definiteness. We have no reason to assume that constituent length is constant between the arguments either. Before we conclude that there is a meaningful difference in length, we should look at the data by grammatical function. Table 4.15 gives the details of the length of Vorfeld constituents and postverbal constituents by grammatical function. We can see that the Vorfeld constituents are consistently shorter than postverbal constituents in each category in the subject and direct object data. The indirect object data shows the opposite trend if we do not exclude pronouns: Vorfeld indirect objects are on average longer than postverbal indirect objects. This is due to the relatively high proportion of personal pronouns, which do not topicalize regularly, in the indirect object data. When we remove the pronouns, the difference in length disappears.

Length by definiteness level We have already seen that the length effect holds within the subject and direct object data, and that the observed difference in length is not solely due to pronouns. Still, the length effect could be caused by the definiteness effect observed in the previous section. In section 4.5, I will use a logistic regression model to investigate

Table 4.16: Average length per argument and definiteness level

	Indefinite full NPs			Definite full NPs		
Argument	ind. det.	bare	Total	def. det.	pn	Total
subject	3.43	2.21	2.89	3.05	1.47	2.64
direct object	3.45	2.58	3.23	3.71	1.40	3.50
indirect object	2.60	1.87	2.16	2.58	1.10	2.23
Combined	3.44	2.42	3.12	3.28	1.45	2.90

whether a definiteness effect can explain the length differences (or, for that matter, whether a length effect can explain the definiteness differences). Here, I will just point out that definiteness and length do not covary in the straightforward way that would be needed for either to explain the effect of the other on Vorfeld placement.

The average length per definiteness level and grammatical function is given in Table 4.16. There is not much of a difference between the lengths of definite full NPs and indefinite full NPs. If anything, definite full NP direct objects are a fraction longer than indefinite full NP direct objects. Table 4.15 showed that nominal, non-pronominal direct objects in the Vorfeld are shorter than those in the postverbal domain. This length difference cannot be caused by the fact that definite full NPs topicalize more often than indefinite full NPs, because this would result in longer Vorfeld occupants in the direct object case.

Table 4.16 does show that bare nouns are on average shorter than indefinite determiner NPs and that proper names are shorter than definite determiner NPs. This may seem a rather obvious point; after all, the NP forms are defined by whether they have a determiner or not. Bare nouns and proper names may be expected to be one word shorter on average. However, this difference in length offers an additional possible explanation for one of the patterns in the data we came across before. Section 4.3 showed that proper names front more often than definite determiner full NPs and that bare nouns front more often than indefinite determiner full NPs. I have offered two (types of) explanations for these differences. The first was that there is a real effect of NP form on fronting. Proper names have been argued to be higher on the definiteness scale than definite descriptions (=definite determiner NPs, Aissen, 2003). The bare noun data may contain more generic indefinite NPs than the indefinite determiner NP data. The second possible explanation for the differences in fronting behaviour was that we see the result of Vorfeld article drop. Article drop reduces the Vorfeld counts in the definite determiner NP and the indefinite determiner NP groups, and it increases the Vorfeld counts in the bare noun group. The supposed effect of grammatical complexity on Vorfeld occupation would offer a third explanation. Assume, for the sake of argument,

that there is such an effect. Bare nouns and proper names are shorter than indefinite determiner NPs and definite determiner NPs. As a result, bare nouns front more often than indefinite determiner NPs, and proper names more often than definite determiner NPs because they are less complex. The logistic regression model in Section 4.5 will be able to show whether complexity alone can explain the increased fronting of bare nouns and proper names.

Let me summarize the findings thus far. Vorfeld occupants are on average shorter than postverbal occupants. This difference in length is observed across different categories and grammatical functions, although the difference in the nominal data is only small. The difference in length cannot be explained by grammatical function. It is also unlikely that the difference is caused by an effect of definiteness on fronting, because definiteness level and length do not covary in the required way.

4.4.2 Locating the complexity effect

We have established that there is a length difference between Vorfeld occupants and postverbal constituents that cannot be explained by any of the factors that we have seen thus far. In this subsection, we will see that claiming that there is a direct effect of complexity on Vorfeld occupation would also be somewhat misleading.

I have operated on the assumption that there is a global, sentence-wide tendency to order simple material before complex material. This is suggested by the complexity principle of Haeseryn et al. (1997). However, others have claimed that the effect of complexity on word order in Dutch is restricted to the right periphery (Jansen and Wijnands, 2004; Van der Beek, 2005). According to these researchers, complex constituents have a tendency to appear at the right periphery in Dutch, but there is no concomitant tendency to order simple before complex material in the rest of the sentence.

A tendency to place complex material at the right periphery (call it the *complex-last principle*) could be on its own responsible for the observed differences. By measuring length in the Vorfeld versus length in the postverbal domain, we would observe a difference in complexity simply because the postverbal domain contains the right periphery. Schematically, the situation is as in (43). We have observed the difference indicated in (43a). The observed difference may be caused by the underlying distribution of simple/complex constituents given in (43b). The location of the complex material in (43b) is only approximate, because the right periphery does not correspond to a specific topological field.

(43) a. simple complex (Observed)

Vorfeld | left bracket | Mittelfeld | right bracket | Nachfeld

b. simple complex (Complex-last)

In order to investigate whether the distribution of complex material is as complex-last predicts, I have grouped constituents according to their position in the topological model of the sentence (see Section 2.1), and recorded their length. A constituent is at the right periphery when its right edge aligns with the right edge of the sentence. The right periphery may correspond to different topological fields, depending on whether there is material in the right bracket (verbal cluster) or not. When there is material in the right bracket, material in the right periphery can be either properly contained in the Nachfeld, or it can be a discontinuous constituent that starts before the right bracket, and ends after it. Non-Vorfeld material that completely precedes the right bracket is fully in the Mittelfeld. When there is no right bracket, a constituent is at the right periphery when there is nothing that follows it. I will refer to these constituents as *postverbal/right*. A postverbal constituent in a sentence with no material in the right bracket that is not right-peripheral will be referred to as *postverbal/left*.

The classification scheme is illustrated in (44a–f). The classified constituents are in boldface. The left and right bracket are between square brackets. The grammatical functions are mentioned for clarity, but they do not play a role in the classification.

(44) a. Subject in Vorfeld (vf):
de toekomst [ligt] open.
the future lies open
'The future lies open.' (Nl-a 289:446)

b. Subject properly in Mittelfeld (mf):
toen [gingen] **Rob en ik** boven [schilderen]
then went Rob and I upstairs paint
'That is when Rob and I went to paint upstairs.' (Nl-a 250:94)

c. Object postverbal, non-final (pv/l):
ik [heb] ook **al mijn aantekeningen van de universiteit** nog
I have too all my notes from the university still
'And I still have all the notes I took at university.' (Nl-a 587:88)

d. Object postverbal, final (pv/r):
ik [koop] wel **vrij veel boeken**
I buy AFF rather many books
'I do buy a fair amount of books.' (Vl-b 400117:7)

Table 4.17: Length of nominal, non-pronominal constituents per position and argument.

		Position					
Argument		vf	mf	pv/l	pv/r	disc.	nf
subject	Avg	2.64	2.48	2.35	4.42	10.14	4.25
	#	7286	1191	1238	863	70	4
direct object	Avg	2.36	2.29	2.36	4.25	8.38	5.43
	#	513	3630	1406	4364	440	51

Note: Not right-peripheral positions are *vf* Vorfeld, *mf* Mittelfeld, and *pv/l* postverbal-left. Right-peripheral positions are *pv/r* postverbal/right, *disc.* discontinuous, and *nf* Nachfeld. See (44) for examples of each.

e. Discontinuous subject with part in Mf and part in Nf (disc.):
hier [worden] dus **onderdeeltjes** [gemaakt] **die in medische**
here are DPART parts made that in medical
instrumenten terechtkomen
instruments end up
'Here parts are made that end up in medical instruments.' (Nl-j 7245:9)

f. Object in Nachfeld (nf):
ik [schrijf] [op] **elastiek voor skipas**
I write up elastic for ski permit
'I'll write down "elastic band for ski permit".' (Vl-a 400092:361)

Table 4.17 tabulates the number of constituents and their average lengths for the nominal, non-pronominal subject and direct object data per position. We see that there are no differences in length between Vorfeld occupants, Mittelfeld constituents, and the postverbal/left constituents of either grammatical function. Therefore, there is no difference in length between the constituents that are not at the right periphery. The average postverbal/right constituent is about two words longer than the constituents that are not at the right periphery. Discontinuous constituents, that start in the Mittelfeld and end in the Nachfeld are longer still. Constituents properly contained in the Nachfeld are shorter again, but still longer than the Vorfeld, Mittelfeld and postverbal/left material. That is, material at the right periphery is longer than material that is not at the right periphery. The data in Table 4.17 show exactly the pattern that Jansen and Wijnands (2004) and Van der Beek (2005) have also found: The only effect of complexity on word order in Dutch is that complex material is moved to the right periphery. Long nominal constituents are avoided to the same extent in the Vorfeld as in the Mittelfeld or in postverbal/left positions.

I conclude that the difference in length that we observed between the Vorfeld and the postverbal domain was caused by the fact that the right periphery falls in the postverbal domain. This also explains why the length differences between Vorfeld and postverbal material were only small, even though they were statistically significant. The effect of complexity was 'watered down' by the other short constituents in the postverbal group. The contrasts in size in Table 4.17 are much sharper.

The fact that the difference between Mittelfeld and discontinuous is larger than the difference between postverbal/left and postverbal/right is likely to be caused by a similar watering down effect. Postverbal/right constituents may be short constituents that happen to be at the right periphery. An example is (45).

(45) ik zie **het**
I see it
'I see.' (Nl-a 260:120)

The direct object *het* cannot help but be at the right periphery. Its position, and thus its classification as a postverbal/right constituent, is unrelated to its complexity.

To sum up, we can say that our working hypothesis has been partially confirmed. Vorfeld constituents are, overall, less complex than constituents elsewhere if we take length in words as a measure of complexity. This difference in complexity is caused by an effect of grammatical complexity on word order in Dutch, as it cannot be explained by grammatical function or definiteness. However, the effect is only indirectly related to the Vorfeld. The influence of grammatical complexity on word order is confined to the right periphery. In the next section, I will briefly discuss what the consequences of this finding might be for a theory of Dutch word order.

4.4.3 Two tentative proposals for theoretical consequences

Van der Beek (2005) suggested that some constituents in the position that we have classified as postverbal/right (pv/r) are actually *extraposed*. That is, the postverbal/right constituent in (46a) occupies the same structural position(s) in the sentence as the discontinuous constituent in (46b). In both cases, the relative clause occupies the same position.

(46) a. Jan reed op **een bromfiets die gestolen was**
Jan rode on a moped that was stolen was
'Jan rode a stolen moped.'

b. Jan kwam op **een bromfiets** aan **die gestolen was**
Jan came on a moped VPART that stolen was
'Jan arrived on a stolen moped.'

In (46b), the relative clause is in the Nachfeld. The Nachfeld is separated from the Mittelfeld by the verb particle *aan* that forms a right bracket. Obviously, it would be hard to prove in a theory-neutral way that the relative clause in (46a) is also in a Nachfeld-like position, because we cannot see the formal counterpart of the right bracket. If one, however, has a syntactic theory of German and Dutch that assumes that even right-bracketless sentences have a (phonetically null) VP-head somewhere in the end of the sentence – that is, if XVX and XVXV sentences have essentially the same structure – one might be able to specify a position to the right of this head, to which to extrapose material, just as one would if there was an explicit right bracket.

This is not the place to explore the theoretical side of this proposal in detail. However, we can have a look at whether the data would support this in terms of the presence of extraposable material. In Dutch, it is typically CP relative clauses and PP postmodifiers that are extraposed out of a nominal constituent. Indeed, over 80% of the postverbal/right constituents have at least one such dependent. In the Vorfeld, Mittelfeld and postverbal/left field, about 10% have at least one PP/CP dependent. So, even if we cannot decide whether postverbal/right constituents contain extraposed dependents, we can conclude that a solid portion of the postverbal/right constituents is of the right kind to allow extraposition of one of its dependents, and that this proportion is much lower in the other groups. Assuming that extraposition is the common cause of (part of) the postverbal/right and (all of) the discontinuous constituents is compatible with the corpus data.

An alternative approach would appeal to performance factors, and builds on the work of Hawkins (2004, and earlier work). Hawkins proposes that an important principle of efficiency influencing the shape of utterances is *Minimize Domains*, defined in (47).

(47) *Minimize Domains* is the preference to 'minimize the connected sequences of linguistic forms and their conventionally associated syntactic and semantic properties in which relations of combination and/or dependency are processed.'
(Hawkins, 2004, p31)

That is, dependencies, of all types, preferably cover short distances. One consequence of this principle is *Early Immediate Constituents*, defined in (48), where the notion of immediate constituent (IC) is comparable to what I have been calling a direct dependent.

(48) *Early Immediate Constituents (EIC)*: '[T]he human parser prefers linear orders that [...] maximize their IC-to-word ratios [...].' (p32, o.c.)

An IC is considered recognized when uniquely defining material of that IC is encountered. This is typically the head: When we are looking for a PP, we know we will find one when we encounter a P; when we are after an NP, we are satisfied with an N or a determiner, etcetera. Maximizing the IC-to-word ratio means minimizing the number of word needed

to recognize all ICs. In his most recent work, Hawkins remains neutral about whether Minimize Domains is a speaker or hearer preference, or both. Now consider an example in Dutch, based on a longer German example (Hawkins, 2004, p138).

(49) hij heeft [$_{VP}$[$_{NP}$ het boek [$_{CP}$ dat de prof verloor]] gisteren gevonden]
he has he book that the professor lost yesterday found
'Yesterday he found the book the professor had lost.'

We are interested in the process of constructing the VP and the object NP. The VP has three ICs: the argument NP, starting with and recognized at *het*, the temporal modifier *gisteren*, and the head *gevonden*. The NP also has three ICs: the determiner *het*, the noun *boek*, and the relative clause, recognized at *dat*.

In the word order in (49), repeated in (50a), we can recognize the three VP ICs in 8 words, and the NP ICs in 3, so the IC-to-word ratios are 3/8 and 3/3 respectively. Two relevant alternative word orders are the unscrambled Mittelfeld order (50b), and extraposition (50c).

(50) hij heeft...

a. het boek dat de prof verloor gisteren gevonden
NP 3/3
VP 3/8

b. gisteren het boek dat de prof verloor gevonden
NP 3/3
VP 3/8

c. gisteren het boek gevonden dat de prof verloor
NP 3/4
VP 3/4

The scrambled version in (50a) has the same IC-to-word ratios as the unscrambled version of (50b). This is because the VP has its head at the end, but the other ICs are all recognized at the first word. No amount of scrambling will change the fact that the whole VP has to be scanned for all ICs to be found. Therefore, in the German and Dutch Mittelfeld, the EIC does not predict a length effect on scrambling alone. However, example (50c) shows that extraposition of the relative clause does have an advantage. The NP deteriorates slightly, but the VP improves greatly. EIC predicts that (50c) is preferred over (50a&b) for processing reasons. Hawkins finds this prediction confirmed in German corpus data in terms of frequency of extraposition and length of the extraposed material.

Fanselow (2000) and Kurz (2000a; 2000b) point out that Hawkins' model predicts that German and Dutch behave like English when there is no material in the verb cluster at the end of the clause. In that case, the VP is head initial (on the surface). This would lead to a length effect on postverbal ordering. An example to show that EIC predicts this is given

in (51)/(52). I will ignore the IC-to-word ratio of the NP, because it does not change. The IC-to-word ratio of the VP improves by moving the long direct object to the end of the sentence into the postverbal/right position. I assume that the structure of the sentence is as in (51).[10]

(51) hij [$_{VP}$ vond [$_{NP}$ het boek dat de prof verloor] gisteren]
he found the book that the professor lost yesterday
'Yesterday, he found the book that the professor had lost.'

(52) hij ...

a. vond het boek dat de prof verloor gisteren
└──────────────────────────┘ VP 3/8

b. vond gisteren het boek dat de prof verloor
└──────────┘ VP 3/3

To summarize, the EIC would predict that, in the absence of a right bracket, there is a length effect on scrambling in Dutch. This fits our findings very well. It would fit the findings of Van der Beek (2005) on the order of direct and indirect object in the Mittelfeld, too. She finds that an effect of length on this order is only observed when there is no right bracket.

I have sketched two ways in which one might approach the observed difference in constituent length between right-peripheral constituents and constituents elsewhere in the sentence. Apart from differing in spirit, the two approaches differ crucially in whether they assume that simplex and complex sentences in Dutch, and presumably also German, are structurally alike (the first, formal approach) or different (the second, functional, Hawkinsian approach). However, as they stand, neither has a story about how the Vorfeld fits in. Recall that, although the Vorfeld constituents were not shorter or longer than the Mittelfeld ones, they were considerably shorter than those in postverbal/right and the discontinuous ones. If one could motivate that being attracted to the right periphery is mutually exclusive with moving to the Vorfeld, one would have explained the length effect on the Vorfeld.

I will not pursue these issue any further here. However, I will present one more example that a theory about the effect of complexity on extraposition and Vorfeld placement will have to deal with. The example in (53) is in some way 'trying to have its cake and eat it, too' by having both the head of the subject in the Vorfeld and a clause final relative

[10]One might argue that the finite verb is not part of VP in the XVXV sentences, and should not be in the XVX ones, either. For the XVXV sentence, I have followed the structure as given by Hawkins. Since the verb on which the argument NP and the temporal modifier depend is the same in the XVXV sentence as in the XVX sentence, I have include this verb in the relevant domain in the simplex sentence, too.

clause. Note that such examples are not in the dataset, because I excluded clauses with discontinuous Vorfeld occupants.

	HD SU				MOD SU			
(53)	**iedereen**	is	een	judas	**die**	**kandidaat**	**was**	
	everybody	is	a	Judas	who	candidate	was	
	'All candidates were Judases.'							(Nl-f 7151:21)

The existence of extraposition in combination with Vorfeld occupation, although rare, shows that a simple account based on mutual exclusion between extraposition and Vorfeld placement will not suffice.

4.4.4 Two types of Nachfeld occupation

In Table 4.17, we saw that arguments that are fully in the Nachfeld are actually shorter than the discontinuous constituents, that begin in the Mittelfeld and end in the Nachfeld. Unlike the postverbal/right group, the reason for the shorter length of Nachfeld constituents cannot be that the average length is pulled down by short constituents that have no place else to go. Inspection of the Nachfeld group however, quickly shows what is going on.

There are two types of Nachfeld constituents. The first type we have seen and discussed at length, and I have referred to these constituents as extraposed constituents. Extraposition involves, and is probably driven by the presence of, complex constituents such as CPs and PPs. Because I have only considered nominal subjects and objects in detail, the PPs and CPs in the Nachfeld have been postmodifiers of an NP. The result of placing such a PP/CP in the Nachfeld is a discontinuous NP, as in (54).

(54) in Kortessem hebben we **een luxeappartement** gezien **met drie slaapkamers**
in Kortessem have we a luxury apartement seen with three bedrooms
'In Kortessem we saw a luxury apartement with three bedrooms.'
(Vl-a 400067:67)

Extraposition of the postmodifier in (54) is optional, as is the extraposition of many PP arguments and PP sentence modifiers. When a subject or object is a CP, Mittelfeld placement is very marginal, as seen in the contrast between (55a), and the constructed alternatives (55b) and (55c).

(55) a. hij gaf wel aan **dat er een netwerk was**
he gave AFF VPART that EXPL a network was
'It did indicate there was a network.' Nl-a (260:277)
b. *?hij gaf wel **dat er een netwerk was** aan
c. **dat er een netwerk was** gaf hij wel aan

Clausal constituents were not investigated in detail in the previous section, but we can expect a very high average length of Nachfeld CP constituents. What Nachfeld positioning of argument CPs and the extraposition of relative clauses and NP modifiers have in common is that (grammaticalized) complexity considerations alone are enough to trigger them. They are pragmatically and prosodically neutral: Neither do they require accenting or deaccenting the extraposed constituent, nor do they demand information structural focus or backgrounding.

However, when we look at the nominal arguments in the Nachfeld, we can discern a second type of Nachfeld occupant. Many of the sentences with a nominal argument in the Nachfeld are not pragmatically neutral at all. They are used to achieve a clear presentational effect. The following examples are typical of this construction. Example (56a) involves a Nachfeld subject. The sentence is uttered by a TV presenter introducing a new item. The second is also broadcast material, in which a museum piece is discussed.

(56) a. hier bij Tijs Van Den Brink en mij is inmiddels aangeschoven
here with Tijs Van Den Brink and me is now joined
Benno Baksteen
Benno Baksteen
'BB has in the mean time joined TvdB and me.' (Nl-l 7228:1)

b. op de bovenste kun je lezen **'t woord Medemblik**
on the top can you read the word Medemblik
'On the top one, you can read the word "Medemblik".' (Nl-j 7418:75)

In text, the desired effect could be indicated by a colon. There is no reason to expect that this Nachfeld positioning has a systematic connection with complexity (see Birner and Ward, 1998, for remarks about complexity and inversion in English, which has some similar properties). The fact that the nominal Nachfeld constituents are on average longer than Mittelfeld and Vorfeld ones can be expected from a) the constituents containing on average a lot of information about what is presented, and b) there being no other place further to the right for possible postmodifiers to go. The presentational construction shares the latter property with the postverbal/right group. Since postverbal/right is at the right periphery, this position may also be the target of presentational post-positioning. The sentence in (42b), p124, with the 83-word subject, might be a good example of this. The Speaker of the Lower House announces topics on the agenda, which is an almost prototypical instance of presentation. In the sentence uttered by the speaker, placing the subject in the Nachfeld in the postverbal domain is motivated by both length considerations and the pragmatics of the construction.

4.4.5 Summary

In this section I have shown that there is a robust difference in length between Vorfeld occupants and constituents elsewhere that cannot be attributed to pronouns or differences in grammatical function and is unlikely to be caused by definiteness of the constituents. Grammatical complexity can therefore be considered a third factor in Vorfeld occupation, next to grammatical function and definiteness.

However, we have also seen that complexity does not influence Vorfeld occupation directly. Rather, all the action concerning complexity is at the right periphery: Sentence-final and discontinuous constituents are longer than Mittelfeld and Vorfeld constituents. Measuring length of Vorfeld constituents against material in the postverbal domain therefore results in a difference.

I have argued that not all placement at the right periphery is driven by grammatical complexity. Dutch has a presentational construction that involves placing material at the right edge of the sentence, which may be the Nachfeld or the postverbal/right position, depending on whether a sentence has material in the verbal cluster or not.

The effect of complexity on the Vorfeld is a third type of effect next to the global word order trends (grammatical function, the definiteness scale), and the first-things-first nature of the Vorfeld. Like the latter, complexity is a word order effect that is tied to a certain position. However, the complexity effect targets the right periphery. Findings about the complexity of Vorfeld constituents are 'only' a side effect of an effect that targets a position at the end of the clause.

Although we have seen that grammatical complexity is not of direct influence on Vorfeld positioning, I will nevertheless treat it as a factor in the rest of this chapter. Even if the effect is only indirect, it may be the case that grammatical complexity explains other trends that we have observed. The increased rate of fronting of bare nouns and proper names compared to indefinite and definite determiner NPs, respectively, might be due to the indirect effect of complexity on fronting. The logistic regression model in the next section allows us to investigate this issue in an insightful and systematic way. A logistic regression model will not only tell us which factors contribute to the Vorfeld occupation, it also allows us to quantify the impact of each factor independent of other factors.

4.5 Grammatical function, definiteness and complexity

Until now, we have found that the chance that an argument appears in the Vorfeld varies along with its grammatical function and definiteness level and, indirectly, with its grammatical complexity. The previous three sections also suggest that these factors cannot be reduced to each other. We observed the effect of definiteness even after splitting the

data into subject, direct object and indirect object (the latter with reservations). We also observed complexity effects in both subjects and direct objects. Therefore, the definiteness and grammatical complexity effects cannot be reduced to grammatical function, nor can the effect of grammatical function be reduced to either of the other two. We have also seen that it is unlikely that the definiteness and complexity effects can be reduced to either one of the two. The two do not appear to covary in the required way. This means that we have made progress in answering the question that is central to the first half of the dissertation: Which constituent properties influence Vorfeld occupation, and how? However, the investigation thus far has raised a couple of questions that have not been answered yet. In this section I will be able to answer the following three questions.

1. Are effects less pronounced for subjects? The trends in the subject and (direct) object data are alike, but the contrasts appear to be subdued in the subject data. For instance, demonstrative pronoun subjects show an increased tendency to appear in the Vorfeld compared to personal pronoun subjects and full NP subjects, but the difference between demonstrative pronoun and the other groups is much more pronounced in the direct object data (Section 4.3). Similarly, the length difference between Vorfeld direct objects and postverbal direct objects is greater than the length difference between Vorfeld subjects and postverbal subjects (Section 4.4). It seems as if subject- and object-fronting are influenced in roughly the same way by the same factors, but that object fronting shows greater variation. In this section we will get a firmer handle on the observed differences.

2. Can complexity explain differences within the definiteness levels? In Section 4.3, we saw that indefinite full NPs front less often than definite full NPs. However, we have also seen that there are differences within the definiteness levels if we look at NP form: Bare nouns front more often than indefinite determiner NPs, and proper names front more often than definite determiner NPs. In both definiteness levels, it is the shorter constituent that fronts more often: Bare nouns are shorter than indefinite determiner NPs, and proper names are shorter than definite determiner NPs. A logistic regression model that includes both length and NP form as factors will be able to answer whether the difference between the forms can be explained by the related length differences.

3. Do indirect objects front more often than direct objects? It is clear that subjects show a stronger tendency to appear in the Vorfeld than objects. However, the relation between direct object fronting and indirect object fronting is less obvious. In Section 2.5, I hypothesized that indirect objects front more often than direct objects do. On a macro-level, it seemed that the opposite was true: On average, nearly 18% of the direct objects appear in the Vorfeld and only 5% of the indirect objects (Section 4.2). A breakdown of the object data by NP form showed that these averages are strongly influenced by differences

in the NP form distributions in direct and indirect objects. Especially demonstrative pronouns, which front frequently, are rare in the indirect object data and relatively frequent in the direct object data. The question that remains is whether we indeed have reason to believe that our initial hypothesis that indirect object front more often than direct objects holds. I caution that a definite answer will have to await further research. A logistic regression model will not solve the sparseness of the indirect object data. However, I will be able to give a preliminary answer here.

The three questions above will be answered in this section by fitting a logistic regression model. The model will also serve as a way to reconfirm some of the earlier findings of this chapter. In Section 4.5.1 I will outline the model that will be fitted and examined. Section 4.5.2 discusses the results of fitting the logistic regression model.

4.5.1 Model definition

We are interested in finding out what properties of a constituent influence Vorfeld occupation, and how these properties do so. Therefore, we will fit a logistic regression model that predicts for a constituent what its chances are of appearing in the Vorfeld, based on its properties (grammatical function, definiteness and complexity).

Grammatical function is a variable with three values: subject, indirect object and direct object. On the basis of canonical argument order in the Mittelfeld, I proposed to view grammatical function as a scale subject $\prec$ indirect object $\prec$ direct object. Appearing left on the scale corresponds to a higher chance of appearing the Vorfeld. In a regression model, two of the three values will be used directly. The third is the base level to which the impact of having a certain property is compared. For grammatical function, we choose direct object as the base level. This means we expect that being a subject and being an indirect object will have a positive impact on Vorfeld occupation in the model.

Instead of the three definiteness levels, I will use the six levels of NP form: indefinite determiner NP, bare noun, definite determiner NP, proper name, demonstrative pronoun and personal pronoun. The base level will be indefinite determiner NP. In Section 4.3, we have seen that the differences in Vorfeld behaviour between the NP forms can be quite large. The effect of NP form on Vorfeld occupation is a mix of the effect of the definiteness scale pronoun $\prec$ definite full NP $\prec$ indefinite full NP and the effect of the first-things-first nature of the Vorfeld. Moreover, to answer the second question formulated above – whether the effect of complexity on Vorfeld occupation can explain the differences between certain NP forms – we need to have information about the NP forms in the model.

In Section 4.4, complexity was measured as constituent length in words. To better satisfy a linearity requirement of logistic regression models, I will use the natural logarithm

of the length in words, instead of length directly. One might object that the use of complexity as a factor in predicting Vorfeld occupation is slightly misleading. In Section 4.4, I concluded that the observed length effect on Vorfeld occupation is a side effect of the effect of length on placement at the right periphery. However, this indirect effect may still explain some of the other observations we have made, such as the difference between proper names and definite determiner NPs. Therefore, I include complexity in the model as a factor, but, in interpreting the model, we should not forget that the influence of complexity is only indirect.

Finally, one of the questions to be answered is whether definiteness and complexity have less of an impact on subjects. One way to assess this in a logistic regression model is to include additional parameters for subjects. I will include two such parameters: Subject NP Form and Subject Complexity. The parameter Subject NP Form will model how the impact of NP form on Vorfeld occupation should be adjusted or corrected for subjects. Like 'regular' NP Form, the parameter Subject NP Form has six values, and the value indefinite determiner NP will be the base level. The parameter Subject Complexity models how the impact of complexity on Vorfeld occupation should be adjusted for subjects.[11]

The model to be fitted can be summarized as in (57), where c is a nominal constituent.[12]

$$\text{(57)} \quad \ln\left(\frac{P(c \text{ in Vorfeld})}{P(c \text{ postverbal})}\right) = \alpha + \beta_{1,2} \text{ Grammatical Function}(c) + \beta_3 \text{ Complexity}(c) + \beta_4 \text{ Subject Complexity}(c) + \beta_{5\ldots9} \text{ NP Form}(c) + \beta_{10\ldots14} \text{ Subject NP Form}(c)$$

Note that some of the factors in (57) are short for two or more binary variables (indicated in the βs), depending on how many levels the factor has. For each level an estimated effect size will be given. See Section 3.5 for explanation of the mapping from multi-levelled

[11]Note that the Subject NP Form and Subject Complexity parameters model the two-way interaction between the *dummy* variable Grammatical Function=subject and the true variables NP Form and Complexity. Studying this interaction is motivated by the observations about the subject data made in the preceding sections, as explained in the text. Inclusion of the two-way interactions that involve the true, three-valued variable Grammatical Function would lead to more model parameters (a superset of those investigated now) capturing very specific situations, which will lead to data sparseness, a loss of generalization and a model that is hard to interpret.

[12]There is an assumption in logistic regression that is violated in our use. Logistic regression assumes that the predicted probabilities are not dependent upon each other. However, since some of the constituents may come from the same sentence (say, a subject and an object) and each sentence only has one Vorfeld occupant, the chance of one constituent appearing in the Vorfeld may be in principle be related to the chance of the other appearing in the Vorfeld. I will not address the issue here. However, in Chapter 6 I fit models that do not violate this assumption and that replicate part of the results in the current section. Therefore, I will assume that violating the independence assumption is not harmful.

Table 4.18: Model 1. Predicting Vf-occupation with NP form and subject correction.

Parameter		Estimate	OR (lo–hi)		p
α		-3.40394			
Gram. Function	direct object *(base level)*				
	indirect object	**0.77109**	1.39	3.35	<.001
	subject	**3.25778**	19.75	34.21	<.001
Complexity		**-0.68625**	0.42	0.60	<.001
Subject Complexity		**0.54592**	1.43	2.08	<.001
NP Form	indefinite determiner *(base level)*				
	bare noun	**0.95200**	1.91	3.50	<.001
	definite determiner	**1.59671**	3.85	6.32	<.001
	proper name	**1.70334**	3.84	7.86	<.001
	demonstrative pronoun	**3.98889**	41.95	69.49	<.001
	personal pronoun	**-2.32257**	0.05	0.18	<.001
Subject NP Form	indefinite determiner *(base level)*				
	bare noun	-0.06844	0.66	1.32	.696
	definite determiner	**-0.36903**	0.52	0.91	.008
	proper name	-0.19822	0.55	1.21	.317
	demonstrative pronoun	**-1.40452**	0.18	0.33	<.001
	personal pronoun	**3.03730**	11.58	37.51	<.001

Note: The leftmost numbers give the parameter estimates (the βs). The rightmost column shows whether the parameter contributes significantly (is significantly not zero, Wald's test). Boldfaced parameter estimates also indicate significant parameters (cf. Wald's test). The middle columns give an indication of the effect size in terms of a 95% confidence interval of the odds ratio.

factors to binary variables. The results of fitting the model in (57) are discussed in the next section.

4.5.2 Modelling results

The model schematically represented in (57) was fitted to 79454 nominal constituents, of which 46041 occupied the Vorfeld. The parameter estimates for Model 1 are given in Table 4.18, p140. Unsurprisingly, Model 1 explains the data a lot better than simply predicting the average probability for each constituent irrespective of its properties ($G^2 = 29048.77$, $df = 14$, $p < .001$). Model 1 is a moderately good predictor of Vorfeld occupation (c-index $= .755$). The model does not show signs of overfitting. I will consider Model 1 in more detail as I go through the answers to the three questions given in the introduction.

1. Are effects less pronounced for subjects? If we look at the parameter estimates for the values of NP Form in Table 4.18, we can see that the chance of appearing in the Vorfeld increases from indefinite determiner NPs through demonstrative pronouns – NP Form has positive parameter estimates that increase in size. A demonstrative pronoun is at least 42 times (lower OR confidence limit) more likely to occur in the Vorfeld than an indefinite determiner NP. Personal pronouns, on the other hand, are much less likely to occur in the Vorfeld than indefinite determiner NPs (at least by a factor 0.18, upper OR confidence limit). However, if a constituent is also a subject, we have to take the Subject NP Form parameter into account when calculating the effect of NP form on Vorfeld occupation. The Subject NP Form estimates up to demonstrative pronoun are increasingly negative. This means that the positive effects of NP Form are dampened in the subject data. For instance, a demonstrative pronoun subject is 'only' about $e^{3.98889+-1.40452} = 13$ times more likely to occur in the Vorfeld than indefinite determiner NP subjects are (as opposed to >42 times for non-subject demonstrative pronouns). Personal pronouns had a decreased chance of appearing in the Vorfeld compared to indefinite determiner NPs, but personal pronoun subjects have an increased chance (OR: $e^{-2.32257+3.03730} = 2$). For NP form, therefore, we can say that the contrasts are indeed less pronounced in the subject data: The increase from indefinite determiner NP through to demonstrative pronoun is present but less strong, and the 'dip' for personal pronouns is much less pronounced.

The same observation can be made for Complexity. The chance of any constituent appearing in the Vorfeld goes down with length. For every added log-word[13] the odds of appearing in the Vorfeld fall by at least a factor 0.60. However, when the constituent is a subject, the negative effect is smaller. A subject's odds of appearing in the Vorfeld decrease only by a factor of $e^{-0.68625+0.54592} = 0.9$ for every added log-word. The effect of complexity on subjects is therefore smaller than the effect of complexity on non-subjects. Even though the effect of complexity on subjects is very small, it is still significant. A model that only allows complexity to have an effect on non-subjects yields a worse fit ($G^2 = 13.63$, $df = 1$, $p < .001$).

The answer to the first question is therefore 'yes'. The contrasts in the subject data are less pronounced.

2. Can complexity explain differences within the definiteness levels? Even when complexity is controlled for, the odds of a bare noun occurring in the Vorfeld instead of in the preverbal domain are at least 1.91 times the odds of an indefinite determiner NP appearing in the Vorfeld. Thus, complexity does not explain this difference. However, the estimates for definite determiner NPs and proper names are very similar in size, and the OR confidence intervals largely overlap (3.85–6.32 and 3.84–7.86). This suggests that

[13]That is, for every time the length of the constituents grows with a factor e: 1 word = 0 log words, 3 words = 1 log word, 7 words = 2 log words, 20 words = 3 log words, 55 words = 4 log words, etcetera.

the differences observed between these two forms of definite full NPs can be explained by complexity. Comparison to a model that only uses one parameter for the two values confirms that the difference between definite determiner NPs and proper names is not significant ($G^2 = 0.42$, df $= 1$, $p = 0.517$).

This result does not carry over to the subjects, however. The constructed OR for a proper name subject is $e^{1.70334+-0.19822} = 4.5$, which is higher than the OR for a definite determiner subject $e^{1.59671+-0.36903} = 3.4$. The difference between definite determiner NP subjects and proper name subjects is significant, as shown by comparison of Model 1 with a model in which neither NP Form nor Subject NP Form distinguishes between the two forms of definite NP ($G^2 = 19.1$, df $= 2$, $p < .001$).

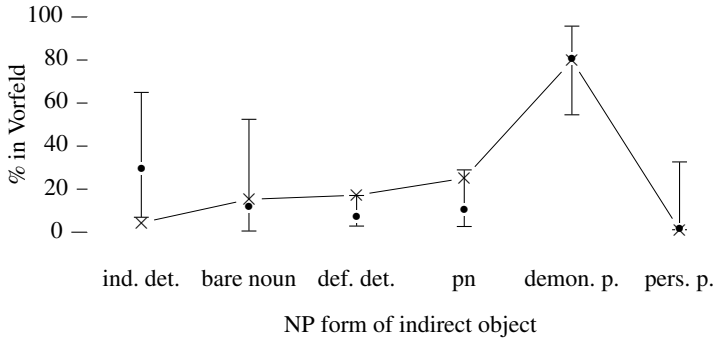
The second question is therefore answered in the positive for definite full NP objects, and in the negative for indefinite full NP objects and subjects of any definiteness level.

3. Do indirect objects front more often than direct objects? The final question to be answered by Model 1 is probably the most interesting one, because it is ultimately part of the larger question of whether behaviour in the Mittelfeld and Vorfeld is similar. If so, we expect that indirect objects *inherently* front more often than direct objects. According to Model 1, the odds of an indirect object fronting are at least 1.39 times the odds of a direct object fronting. The difference is small but significant. It seems therefore that the answer to the third question is also 'yes'. However, this answer comes with some reservation. Although Model 1 allows NP Form to have different effects on subjects and non-subjects, the assumption is that the effect of NP Form on direct objects and indirect objects is the same. This assumption is not clearly supported by the data, but it is not completely at odds with the observations either. Figure 4.1 shows this by plotting, for indirect objects, the observed effect of NP form on Vorfeld occupation and the modelled effect of NP form on Vorfeld occupation. The predicted probabilities of occupying the Vorfeld fall inside the confidence intervals of the observed data, except for the indefinite determiner NPs (too low) and definite determiner NPs (too high). Recall that the indirect object data suffers both from being sparse and from being heterogeneous. This prevents us from making solid statements about the data. However, as a preliminary answer to the question, I submit that indirect objects front slightly more often than direct objects do on the assumption that indirect objects and direct objects are equally influenced by NP form.

4.5.3 Summary

In this section, we have looked at the combination of grammatical function, definiteness (NP form), and complexity as factors in Vorfeld occupation. Analysis of a logistic regression model predicting Vorfeld occupation of nominal constituents reproduced the main findings of Sections 4.2–4.4: Subjects front more often than objects; constituents

Figure 4.1: Comparison of fitted and observed indirect object behaviour, by NP form.



Note: The connected crosses '×' indicate the average predicted probability of an indirect object appearing in the Vorfeld, according to Model 1. The bullets '•' and 95% confidence intervals indicate the observed Vorfeld proportions, as taken from Table 4.12, p116.

that are higher on the definiteness scale front more often than constituents lower on the definiteness scale, with the exception of personal pronouns; and less complex material is more likely to end up in the Vorfeld.

We have also been able to find answers to the three open questions. We have seen that the contrasts in the subject data are less pronounced than in the object data, for both definiteness and complexity. I also concluded that complexity can explain the difference in fronting between definite determiner NP objects and proper name objects – the latter are more likely to be Vorfeld occupants because they are shorter. However, neither the difference between definite determiner NPs and proper names in the subject data, nor the differences between indefinite determiner NPs and bare nouns of any grammatical function are explained by complexity.

Finally, we saw that indirect objects front slightly more often than direct objects on the assumption that NP form has the same influence on indirect objects and direct objects. This is as was predicted by the assumption that canonical order of arguments is a global word order trend, which applies to the Mittelfeld as well as to the Vorfeld. It must be emphasized that the equal influence of NP form is an assumption and that it is not a conclusion that is strongly supported by the data. More indirect object data are needed to confirm these preliminary findings.

4.6 The presence of negation, and other modifiers

Until now we have investigated Vorfeld occupation by looking at properties of the potential Vorfeld occupants themselves. In this short section, I will look at the influence of the environment of potential Vorfeld occupants. The question to be addressed in this section is whether the presence of negation in a sentence, and possible other sentence adverbials, facilitates topicalization.[14]

In Section 2.2.2, I mentioned that topicalization of predicates is facilitated by the presence of negation (see also Birner and Ward, 1998, for observations about English). An example is given in (58).

(58) een verkeerschaos wordt het ?(niet)
a traffic chaos becomes it not
'There will not be a traffic chaos.' (Nl-k 4671:3)

The observed variant with negation is a perfectly natural example of a topicalized predicate. Dropping the negation leads to reduced grammaticality. In the previous section we have seen many examples of object topicalization without negation. This suggests that object topicalization does not necessarily need the help of negation to be acceptable. However, it may still be the case that the presence of negation has a positive influence on object fronting.

Instead of looking at negation (*niet*, 'not') alone, or at a restricted set of negative adverbs, I will take a data-driven approach and look at all sentence adverbs and their relation to object fronting. This approach is helpful because we know so little about the phenomenon. It may be that there is quite a wide group of sentence adverbs that 'promote' object fronting, that it is only direct negation in the form of *niet*, or that the effect does not exist at all. If we take a very broad look at the data, we can point out the sentence adverbs that are positively correlated with object fronting, if there are any.

Sentence adverbs in the CGN are always direct MOD dependents of the SMAIN. Therefore, I will use the dependency path MOD as an operationalization of sentence adverb. Note, however, that not *all* direct MODs are sentence adverbs. I will look at the individual adverbs later on in this section. The fact that the operationalization of sentence adverb is too wide therefore need not worry us.

Apart from a potential positive effect on object fronting, the presence of a sentence adverb will in any case have a negative effect on Vorfeld occupation of other constituents

[14]The section is of a more preliminary nature than the previous sections, and I will not be able to do much more than present the results, leaving the effects found largely unexplained. However, a direct motivation for investigating the issue is the often-made, but, as far as I am aware, undocumented observation that *all kinds* of topicalization, including object fronting, sound better with a negation present in the sentence. A more thorough investigation and theoretical embedding will have to await future research.

Table 4.19: Vorfeld occupation per argument, per number of direct modifiers.

	subject			direct object			indirect object		
#MODs	Vf	¬Vf	%	Vf	¬Vf	%	Vf	¬Vf	%
0	6932	1401	83.2	1009	7381	12.0	8	195	3.9
1	6018	4046	59.8	1637	8464	16.2	9	183	4.7
2	1908	2619	42.1	670	3875	14.7	3	66	4.3
3	388	743	34.3	124	1010	10.9	0	10	0
4	46	117	28.2	12	153	7.3	0	2	0
5	7	18	28.0	2	23	8.0			
6	1	4	20.0	0	5	0			
7	1	1	50.0	0	2	0			

in the sentence. The reason for this is that sentence adverbs may occupy the Vorfeld themselves. Adding a sentence adverb to a sentence should therefore reduce the chance of Vorfeld occupation for each of the already present constituents. Any positive effect on object fronting will be moderated by this negative effect on fronting.

To investigate the effect of MODs on object fronting, I will use a slightly different dataset from the one I have used until now. As pointed out in Section 2.4, the CGN annotates examples like (59) as containing two Vorfeld occupants: a subject and a modifier. The finite verb is in boldface.

(59) [ook] [een soort glinsterende regen] **daalde** neer
also a sort glistening rain descended VPART
'Also a kind of glistening rain fell.' (Vl-o 800455:26)

Sentences like (59) have not been part of the dataset in the previous sections, because I only looked at sentences with exactly one Vorfeld occupant. In this section, however, this would interfere too much with the results, because MODs like *ook* can be expected to appear rather often together with another constituent in the Vorfeld.

Table 4.19 shows the effect of adding one or more MODs to a sentence on subject, direct object and indirect object fronting. In the subject data, we clearly see the negative effect of adding an extra modifier: The chance of the subject itself taking up the Vorfeld goes down. In the direct object data, we see the same for the rows where #MOD$\geq$1: The proportion of direct objects in the Vorfeld decreases. However, there is an interesting 'bump' in Vorfeld occupation when we add the first modifier. The probability of direct object fronting starts fairly low at 12%. However, 16% of the direct objects occurring in a sentence with exactly one MOD appears in the Vorfeld. This rise is highly significant (OR $\frac{\#\text{MODs}=1}{\#\text{MODs}=0} = 1.4$, p<.001, 2-t Fisher's). The indirect object data is too scarce to show an

effect. The rise from 3.9% to 4.7% observed for the first MOD added is not significant (OR $\frac{\#\text{MODs}=1}{\#\text{MODs}=0} = 1.2$, p=.806).

Before concluding that the bump in the direct object data is caused by a positive effect of sentence adverbials, we need to have a closer look at the data in two ways. First, we have to exclude the possibly confounding effect of negation incorporation (see also Section 4.3). Secondly, we should take a look at how uniform the positive effect is in the MODs.

We have seen before that indefinite full NP objects in Dutch tend to incorporate sentence negation. Sentence negation in example (60a) is expressed by the negative quantifier (boldfaced) in the direct object NP. However, when the direct object occupies the Vorfeld, sentence negation is expressed by *niet* (60a) and not as a negative quantifier *geen* (60b).

(60) a. daar hebben ze vast **geen** vlizotrap
there have they probably no folding stairs (Nl-a 303:21)

b. ***Geen** vlizotrap hebben ze daar vast.
no folding stairs have they there probably (constructed)

c. Een vlizotrap hebben ze daar vast **niet.**
a folding stairs have they there probably not
'They probably do not have folding stairs.' (constructed)

In our investigation of the effect of negation on Vorfeld occupation, negation incorporation presents us with a problem. The initial increase in direct object fronting in the presence of a MOD may be due to the fact that only postverbal direct objects incorporate sentence negation. When the object is indefinite, word order influences the expression of negation, and thereby the presence of a MOD. A postverbal object in a negated sentence may appear in the MOD=0 group, whereas all Vorfeld direct objects in a negated sentence appear in MOD≥1. I will circumvent the problem by only considering definite full NP and pronominal direct objects.[15]

Table 4.20 gives details of direct object fronting per number of MODs, for definite full NPs and pronouns alone. We can observe the same pattern in Table 4.20 as we did in Table 4.19. There is an initial rise in direct object fronting, followed by a falling tail. The

[15]These NPs may sometimes incorporate negation, too. However, this is either exceedingly rare in the corpus, or not relevant when studying direct objects. For instance, a used proper name can incorporate negation (i), as can a mentioned one (ii).

i overigens nog altijd **geen** Henke Larsson
by the way still always no Henke Larsson
'Henke Larsson is still not playing, by the way.' (Vl-i 600749:39)

ii die heet **geen** Esther
DEM is called no Esther
'She is not called Esther.' (Nl-a 535:84)

Table 4.20: Vorfeld occupation by non-indefinite direct objects, per number of MODs

	Vorfeld		Prop est (%)		
#MODs	yes	no	lo	pt	hi
0	839	2348	24.8	**26.3**	27.9
1	1478	3561	31.3	**32.5**	33.7
2	610	1849	22.4	**24.1**	25.8
3	119	473	15.3	**18.4**	21.8
4	11	73	7.5	**14.1**	23.3
5	1	6	0.3	**14.3**	57.9
Total	3058	8310	26.1	**26.9**	27.7

percentages are consistently higher than in Table 4.19. This is to be expected, because the excluded group of indefinite full NPs has the lowest tendency to front. The initial rise in Table 4.19 is significant (OR $\frac{\#\text{MODs}=1}{\#\text{MODs}=0} = 1.2$, p=.003, 2-t Fisher's), as also indicated by the non-overlapping confidence intervals.

We can conclude that the initial rise is not due to, or at least not completely due to negation incorporation. There is therefore a positive correlation between the presence of a MOD in a sentence and Vorfeld occupation by a direct object. Now, let us look inside the MODs to see whether the observed positive relation between the presence of a MOD and direct object topicalization is limited to, or more pronounced with, certain adverbs.

Table 4.21 lists the 20 most frequent MODs in the sample of sentences with a non-indefinite direct object. Frequency is based on simple string identity. Note that a sentence can have more than one direct modifier, so the presence of one modifier does not exclude the presence of another.

We have no systematic way of deciding when a Vorfeld proportion should count as relatively high. Table 4.20 tells us that on average, 26.9% of the non-indefinite direct objects appears in the Vorfeld. Let us therefore look at the MODs for which the proportion of Vorfeld non-indefinite direct objects is 30% or higher. Of the 20 most frequent MODs, 4 are clearly above this ad hoc cut-off point, because their lower confidence limits are above 30%. These are *niet* 'not', *ook* 'too', *wel* AFF and *dus* 'so'. There are 4 further MODS that have their point estimate above 30%: *al* 'already/yet' *toch* 'still' *niet meer* 'no longer' and *altijd* 'always'. There are also MODs that are associated with reduced direct object fronting: *dan* 'then' (implication/future), *toen* 'then' (past), and *daar* 'there'. The reason for their negative effect on direct object Vorfeld occupation is that they show a strong tendency to appear in the Vorfeld themselves.

Of all the adverbs that go together with high proportions of direct object fronting, *niet* has the greatest impact: It shows a very high percentage of Vorfeld objects and *niet* tops

Table 4.21: Vorfeld occupation of non-indefinite direct objects, in the presence of 20 most frequent MODs.

	Vorfeld		Prop est (%)		
MOD	yes	no	lo	pt	hi
niet 'not'	504	822	35.3	**38.0**	40.7
dan 'then' (implication/future)	132	899	10.8	**12.8**	15.0
ook 'too'	372	620	34.5	**37.5**	40.6
wel (affirmation)	273	547	30.1	**33.3**	36.6
dus 'so'	103	178	31.0	**36.7**	42.6
gewoon 'just'	57	204	17.0	**21.8**	27.3
al 'already/yet'	84	169	27.4	**33.2**	39.4
nu 'now'	55	179	18.2	**23.5**	29.5
toch 'still' (contrast)	71	161	24.7	**30.6**	37.0
nog 'still' (time)	59	165	20.7	**26.3**	32.6
eigenlijk 'actually'	47	141	19.0	**25.0**	31.8
zo 'thus/soon'	28	126	12.4	**18.2**	25.2
toen 'then' (past)	11	146	3.5	**7.0**	12.2
echt 'really'	43	109	21.3	**28.3**	36.2
niet meer 'no more'	47	83	27.9	**36.2**	45.0
natuurlijk 'of course'	35	94	19.7	**27.1**	35.7
weer 'again'	25	97	12.7	**20.5**	28.7
altijd 'always'	39	77	25.1	**33.6**	43.0
even 'just/briefly'	19	86	11.3	**18.1**	26.8
daar 'there'	10	86	5.1	**10.4**	18.3

the frequency list in this subset of the direct object data. Note again that this effect cannot be explained by incorporation of negation, because all NPs are definite. The prominence of *niet* in Table 4.20 aligns well with the observation of the positive effect of sentence negation on predicate topicalization I started this section with. Even though this parallel does not explain the effect, it does indicate that an explanation should draw on a wider range of topicalization data than just predicate topicalization.

However, we can also conclude that the effect is clearly not limited to negation alone. From the 20 most frequent adverbs in Table 4.21, we selected 7 more adverbs that go together with increased direct object topicalization. There are some interesting commonalities between the selected adverbs: They can be argued to be focus-sensitive in at least one of their uses. This goes for negation *niet*, as well as its positive counterpart *wel*, and *toch* when it is used as a contrastive positive polarity marker. The particles *niet meer* 'no more'

and *al* 'already' are temporal/aspectual particles, which have been classified as special instances of focus particles (Haeseryn et al., 1997; Van der Wouden, 2002). The additive particle *ook* 'too' is the classic example of a focus particle, and the temporal adverb *altijd* 'always' has also long been considered to be a focus-sensitive item (Kadmon, 2001, for overview and references).

The odd one out in this group appears to be *dus*, which I have translated in the table with 'so'. However, although it can be used as a connective adverb (61a), inspection of the data show that use as a modal particle is much more common (61b,c).

(61) a. (Speaker explains he is from a Frisian-speaking family and his wife is, too...)
dus spreek je samen Fries
so speak you together Frisian
'...so we speak Frisian together.' (Nl-f 7189:41)

b. dat bedoel ik **dus**
that mean I DPART
'Exactly!' (Nl-a 372:13)

c. want dat hebben we bij ons **dus** niet hè
for that have we at us DPART not TAG
'Because, you know, we don't have that here, do we.' (Vl-h 400225:112)

This *dus* often appears in the presence of *ook* and *niet*, although the proportion of Vorfeld direct objects is still high when these two are not present.[16]

The observation that (all but one of) the adverbs have somehow been associated with focus in the past does not immediately make clear what the explanation of the effect should be. However, it does suggest that a closer study of the relation between these adverbs and information structure (focus) may point towards an answer. An explanation in terms of focus may give us problems as well, though. For instance, the group of temporal/aspectual particles also includes *nog* 'still'. Table 4.21 shows that *nog* is not associated with an increase in object fronting at all. On the assumption that temporal/aspectual particles behave like any other focus particle, this exception would be surprising.[17].

Informal inspection of the instances of the selected particles in the sample shows a) a high number of demonstrative direct objects, and b) a salient amount of verbs of doing, knowing and allowing. The role of verb semantics in word order will not be considered at all in this dissertation, although it is conceivable that it plays a significant role (for the influence on double object constructions, see, for instance, Bresnan et al., 2007 for English and Van der Beek, 2005 for Dutch). With respect to point a), we have already concluded that demonstrative pronouns are good candidates for fronting in this chapter.

[16]As an aside, particles like the adverbs mentioned above are known to frequently appear in groups (Van der Wouden, 2002). In this respect *dus* is no exception.

[17]This was pointed out to me in by Ton van der Wouden in personal communication

However, even if we could explain the direct object fronting proportions in Table 4.21 from properties of the verb and of the direct objects, we would not be able to explain why the eight adverbs given above like to occur with initial direct objects. Neither would we be able to explain the 'bump' in Table 4.20.

Finally, one might try to offer a construction explanation for the data in Table 4.21. It might be that the particles selected happen to be frequent in object-initial constructions or nearly fixed expressions. Examples with *toch* en *wel* are in (62).

(62) a. dat zei ik **toch**
that said I PART
'I told you so.' (Nl-c 8058:142)

b. dat weet ik **wel**
that know I AFF
'I know.' (Vl-a 400404:11)

If these constructions and other like them were very frequent, we would observe a correlation between the adverbs and direct object fronting simply because they are part of such a construction. It should be possible to get an idea of the role that such constructions play in the data using statistical methods. However, ultimately an account in terms of fixed constructions would not explain why we can see the effect of negation in both object and predicate topicalization.

In this section, I have briefly investigated the relation between non-canonical word order and adverbs. The investigation was prompted by the observation that sentence negation facilitates topicalization of predicates and the speculative claim that this is also true for direct objects.

We found that a small group of (focus-sensitive) adverbs is associated with an increase of direct object fronting. The effect was not restricted to negation. The combined positive effect that these adverbs have on direct object fronting, or, alternatively, the positive effect of what underlies the occurrence of these adverbs, is strong enough for an clear increase in direct object fronting to be observed in the overall data. I have briefly pointed in three (overlapping) directions where we may begin to look for an explanation of this effect (information structure, properties of the direct objects themselves, fixed constructions). At this point, I have no explanation to offer myself.

4.7 Conclusion

The central aim of this dissertation is to find out what drives the selection of a Vorfeld occupant in Dutch. In this chapter, I have used evidence from the spoken Dutch corpus to answer part of this question, namely: How do a constituent's properties influence the

chance of that constituent appearing in the Vorfeld? The constituent properties that were investigated were grammatical function, definiteness and grammatical complexity. By investigating these three properties and their relation to Vorfeld occupation, we have gained insight into Vorfeld occupation.

Let me start by briefly summarizing the raw findings of the chapter. Of the three investigated arguments – subjects, indirect object and direct objects – subjects clearly have the highest chance of appearing in the Vorfeld. We also found that indirect objects front slightly more often than direct objects, but this relies on the assumption that indirect and direct objects react equally to the other factors (definiteness, complexity). There is too little indirect object data to decide whether this assumption is really warranted or not. Moreover, indirect objects form a particularly heterogeneous group in the CGN. Further research on a larger corpus is needed to investigate whether this heterogeneity means that there is differentiation in Vorfeld behaviour, too. For now, we can draw up a fronting scale, where an element on the scale fronts more often than an element to its right.

(63) *Positive relation between grammatical function and Vorfeld occupation:*
subject $\prec$ indirect object $\preceq$ direct object

The relation between definiteness and Vorfeld occupation is complicated by the fact that there appear to be two trends. First, there is the trend that constituents that are higher in definiteness front more often, summarized in (64).

(64) *Positive relation between grammatical function and Vorfeld occupation:*
pronoun $\prec$ definite full NP $\prec$ indefinite full NP

The pronoun data is highly differentiated, however. Demonstrative pronouns show very high frequencies of fronting. Yet, full and reduced personal pronouns do not fit in the scale of (64) – they front less often than definite full NPs (in the case of subjects) or even less often than indefinite full NPs (in the case of objects). There is therefore a second scale related to definiteness, (65).

(65) *Negative relation between pronominal form and Vorfeld occupation*
reduced personal pronoun $\succ$ full personal pronoun $\succ$ demonstrative pronoun.

The trend in (65) is stronger than the trend in (64).

Finally, I showed that grammatical complexity has in an indirect influence on Vorfeld occupation. Because grammatically complex constituents are drawn to the right periphery of the clause, they are less likely to appear in the Vorfeld.

An important issue in studying the three constituents properties was whether their influences on Vorfeld occupation can be explained in terms of global word order trends or whether they have to be considered to be special, position-related effects. With this question in mind, the effects described above can be divided into three categories: global

factors, factors targeting the Vorfeld, and factors targeting other positions. The effect of grammatical function and the definiteness scale can be related to global word order trends. The trends found in the Vorfeld mirror word order trends in the Mittelfeld. We can therefore assume that the same forces underly these trends, independent of sentence position or domain. The negative relation between pronominal form and Vorfeld occupation sketched in (65) is of the second category – it is due to special properties of the Vorfeld. According to Gundel's *first-things-first* principle, important (unexpected, new, unpredictable) material should be realized first. Vorfeld occupation can be considered a grammaticalization of this principle. Demonstrative pronouns realize material that is relatively unpredictable compared to (reduced) personal pronouns. Thus, a demonstrative pronoun in the Vorfeld agrees with the first-things-first principle, but a personal pronoun in the Vorfeld does not. The Vorfeld, as a place for important information, prefers not to be occupied by personal pronouns. This local factor also explains why Vorfeld occupation can differ from Mittelfeld word order, even when both are subject to the same global factors. In contrast to the Vorfeld, reduced personal pronouns in the Mittelfeld are preferably realized early because there is nothing that prevents unimportant material to be realized early in the Mittelfeld. Finally, the effect of complexity on Vorfeld occupation is of the third category. There is no sense in which non-complex material is drawn to the Vorfeld or to the left in general. Rather, complex material has a tendency to appear directly at the right periphery. This has the side effect that the Vorfeld contains relatively little complex material. We see that we need to consider the combination of global and local trends to begin to understand Vorfeld occupation.

In the last section of this chapter, I presented evidence that direct object fronting is correlated with the presence of a sentence adverbial in a clause. Especially certain adverbials – most prominently negation, but also other focus-particle-like adverbs – seem to have a positive influence on direct object fronting. These novel findings are in line with prior claims that negation facilitates predicate fronting.

Chapter 5

Word Order Freezing

In the preceding chapters, we have seen which properties of a constituent affect the chance that the constituent is placed into the Vorfeld, and how. The choice of Vorfeld occupant is influenced by global word order trends with respect to grammatical function and definiteness, by local properties of the Vorfeld as a position for important material, and by the preference to put complex material at the right periphery. These results offer a partial answer to the question of what drives Vorfeld occupation by making clear what makes a constituent an attractive candidate for Vorfeld occupancy.

The freedom the speaker has in choosing a Vorfeld occupant comes at a price: It may be that the person interpreting the utterance assigns grammatical function incorrectly and thus misinterprets the utterance. For instance, if the speaker puts a direct object that is not clearly recognizable as a direct object in the Vorfeld, it may be mistaken for a subject by a hearer whose default assumption is that canonical argument order is adhered to. The result would be communicative failure.

It has been observed that some free word order languages deal with comparable problems by suspending word order freedom if communicative success is not guaranteed. This contingent loss of variation has been referred to as *word order freezing*. In these languages, communicative success can be considered to be a factor in word order variation. In this chapter and the next, I will investigate whether variation in Vorfeld occupancy is influenced by communicative success and how this influence is best modelled. The current chapter approaches these questions from a theoretical perspective. A theoretical model of word order freezing will be proposed. In the next chapter, I will test predictions derived from this theoretical model against the spoken Dutch corpus CGN.

The theoretical model of word order that is capable of predicting freezing will be formulated in the framework of *Optimality Theory*. Optimality Theory is a suitable framework to model word order variation in, because it naturally accommodates possibly

conflicting preferences like the different trends in Vorfeld occupation we have seen in the previous chapters. In addition, word order freezing has received quite some attention in the Optimality Theory literature. The influence that a requirement of communicative success may have on speaker choices can be precisely modelled in a variant of Optimality Theory known as *bidirectional Optimality Theory*.

This chapter is organized as follows. In Section 5.1 I will begin by briefly arguing that Dutch shows signs of word order freezing in the choice of Vorfeld occupant, and that word order freezing in general should be treated as a grammatical phenomenon. I will also discuss how the theoretical claims and the discrete model of this chapter should be understood in the context of the empirical studies presented in the rest of the dissertation. Section 5.2 introduces Optimality Theory. On the basis of existing results for the word order freezing language Hindi, Section 5.3 describes how bidirectional Optimality Theory can be used to model word order freezing. Section 5.4 discusses disadvantages of a bidirectional model, and looks at two alternatives that have recently been put forward. In Section 5.5, I will show that the bidirectional model of word order can be extended to successfully address all the problems brought forward against bidirectional models in the previous section. Applied to Vorfeld occupation in Dutch, the resulting theoretical model of word order is a formally precise answer to the question how the speaker's choice for a Vorfeld occupant depends on communicative success.

The last two sections discuss alternative formalizations and future work. Section 5.6 includes a discussion of various alternative definitions of bidirectional Optimality Theory, and their advantages and disadvantages in a model of word order. Section 5.7 considers directions for further investigation, and summarizes and concludes the chapter.

5.1 Introduction

In this section I will introduce word order freezing by showing that it can be observed as a tendency in Vorfeld occupation under certain circumstances. I will also discuss some of the previous approaches to word order freezing and I will briefly explain why word order freezing should be treated as a grammatical phenomenon. The grammatical model capable of capturing word order freezing that I will defend in this chapter is based on data that is rather abstract compared to the corpus data that I have relied on thus far. Furthermore, the model makes discrete predictions about the grammaticality of a certain word order, an approach that seems to be at odds with the gradient nature of corpus data. At the end of this section, I will therefore say a few words about the relation between the theoretical work presented in this chapter and the empirical investigations in the next.

5.1.1 Word order freezing in Dutch

In previous chapters, we saw that the Dutch Vorfeld could be occupied by a range of constituents. The subject occupies the Vorfeld in seven out of ten sentences, but other constituents front quite regularly, too. For instance, almost a fifth of the direct objects appears in the Vorfeld. As a result, a fifth of the direct objects appears in a position that is in the majority of sentences taken up by the subject. What is more, in those sentences, the subject may appear in a position that would have been taken up by the direct object if the sentences had adhered the canonical argument order subject-before-object. Subject and object in Dutch can appear in 'each other's' position. A (nearly) minimal pair taken from the corpus illustrates this.

(1) a. **ik** snap **dat** nooit
I understand.1SG that never
'I don't understand that.' (SVOAdv) (Nl-a 250:66)

b. **dat** bedoel **ik** nou
that mean.1SG I now
'That's what I mean.' (OVSAdv) (Nl-a 422:252)

In (1a), the subject sits in the Vorfeld and the object appears in the Mittelfeld directly after the verb, before the adverbial. In (1b), the roles are reversed: The object occupies the Vorfeld and the subject is in the Mittelfeld between the verb and the adverb.

A consequence of the fact that the subject and object can appear in the same positions is that word order itself does not unambiguously encode grammatical function. Both sentences in (1) are of the form NP–V–NP–Adv, so this form is compatible with two argument orders. Compare Dutch in this respect to a language like English. In as good as all cases, a subject in English can be distinguished from the object by means of word order. Even when the object precedes the subject in a topicalized sentence, the position of the subject and object with respect to the verb will tell us what the correct grammatical function assignment is.[1]

[1]For example, in the sentences (i) and in (iia), the object precedes the subject, but grammatical function can be assigned on the basis of the position of the NPs relative to the verbs.

(i) a. This cat Jennifer likes.
b. Jennifer this cat likes.

(ii) a. What doesn't Jennifer like.
b. What doesn't like Jennifer.

Possible exceptions may be found in 'V2-like' inversion constructions. The quotative inversion is an example.

(iii) a. "A Cheshire cat," said Alice.
b. A Cheshire cat said, "Alice."

Sentences (iiia) and (iiib) have the identical strings, but different grammatical function assignments.

In the sentences in (1), there is other information that will tell us which NP is the subject and which NP is the object. There is morphological information in the form of case (*ik* is subject form) and agreement (*bedoel* and *snap* bear first person singular agreement). Morphologically, the sentences are unambiguously SVO (1a) and OVS (1b). In addition, if I had supplied a context, we would have seen that the *dat*s refer to abstract entities (propositions, situations). The Dutch verbs *bedoelen* and *snappen* cannot take these as their subjects. Selection information of the verb therefore also encodes grammatical function unambiguously in these sentences.

What happens when there is no morphological or selection information present? The fact that NP–V–NP is compatible with SVO as well as OVS may lead us to expect that such a sentence is ambiguous. This expectation is not completely borne out. Out of context, and without indication of a marked intonation, the example in (2) is not perceived to be ambiguous. It is interpreted as SVO.

(2) De jongens zien de meisjes.
the boys see the girls
'The boys see the girls.' (SVO)
Not, or strongly dispreferred: 'The girls see the boys.' (OVS)

The resort to canonical word order, and the resulting lack of ambiguity, in a sentence like (2) is known as *word order freezing*.[2] Word order freezing has been observed for a wide range of free word order languages, although some languages may show freezing in more circumstances than others. It is common for researchers to focus on the lack of morphological information as a trigger for word order freezing, but this is certainly not the only type of information that plays a role. I will talk more about other languages below. First, however, I would like to make clear how strong I consider the effect to be with respect to Dutch Vorfeld occupation.

It might be objected that (2) is not evidence of word order freezing in Dutch at all, because the SVO word order is only preferred and not absolute. It is possible to produce sentences in Dutch similar to (2) that are ambiguous between SVO and OVS, or that are predominantly non-canonical OVS. This is demonstrated by the examples in (3) and (4). The sentence in (3) is pronounced as a focus topicalization (Section 2.5.1) – nuclear stress falls in the Vorfeld. The sentence is genuinely ambiguous between SVO and OVS. Sentence (4b) is pronounced as a contrastive topic topicalization (Section 2.5.2) – a prominent rise on the Vorfeld occupant, and nuclear stress in the postverbal domain. In

[2]The term is due to Tara Mohanan, who supposedly coined it in a paper presented at the Stanford Syntax Workshop in 1992. An early use of the term is found in Mohanan and Mohanan (1994), albeit applied to data that is not directly relevant for the current discussion. The term *freezing* used in this dissertation should not be confused with the term occurring in Chomskyan syntax traditions, where it refers to extraction being prohibited under circumstances very different from ours.

the context of (4a), (4b) is interpreted as OVS. Capitals indicate nuclear stress, slashes indicate rising and falling accents.

(3) \de JONGens\ zien de meisjes.
the boys see the girls
'The BOYS see the girls.' (SVO) *Or:* 'The girls see the BOYS.' (OVS).

(4) a. /Fitz/ valt maar bij \WEInigen\ in de smaak...
'Fitz is liked by a few people only ...'
b. ...maar /Gerald/ vindt \iederEEN\ knap.
but Gerald finds everybody handsome
'...but everybody thinks Gerald is handsome.' (OVSPred)

Examples (3) and (4) show that intonation and context may also bring about a non-canonical reading – a sentence can even be ambiguous. However, in the absence of these extra factors, canonical SVO is preferred to the point that non-canonical readings are not available. The reason that I consider Dutch to show word order freezing is that the SVO reading of (2) does not need any help. There are clear parallels with (my reformulation of) Lenerz' concept of unmarked or canonical word order (Section 2.6.1): The unmarked argument order is that word order a speaker may use without any other factors (such as definiteness) promoting it. The SVO reading of (2) shows that SVO is the unmarked word order in interpretation, too. A hearer does not need other factors (such as case or agreement) to understand a sentence as SVO. It is obvious, however, that the 'other factors' that influence interpretation in Dutch, go beyond morphological information and include selectional information, context and intonation.

5.1.2 Approaches to word order freezing

For many free word order languages, word order freezing has been claimed to exist. The classic claim is made by Jakobson (1936). Russian is a case language that normally shows great word order freedom driven by information structure. All three combinations of subject, object, and verb are in principle possible. However, there is a class of words that shows syncretism of nominative and accusative case. Jakobson observed that grammatical function assignment in a sentence is not free when both S and O show syncretism of case, as in (5).

(5) Mat' ljubit doč'
mother.NOM/ACC loves daughter.NOM/ACC
'Mother loves her daughter.' *Not:* 'The daughter loves her mother.'

If we were to revers the order of the NPs, and start with *doč'*, the sentence would change meaning.

Word order freezing has been observed in languages as different as Hindi, Korean, German, Bulgarian, Russian and Papuan languages (see Lee, 2001a, for a list of references), Haida, Swedish (Morimoto, ms) and Japanese (Tonoike, 1980; Kuno, 1980; Flack, 2007). The circumstances in which word order freezes are often characterized by a lack of morphological information about grammatical function assignment, as in the case of Russian, above. However, languages may vary in this respect. We have seen that a lack of morphological information alone is not enough to cause word order freezing in Dutch.

Word order freezing constitutes a challenge for a theory of word order. On the one hand, one and the same word order has to be compatible with several grammatical function assignments. On the other hand, when other sources of information, say case, in the sentence are also compatible with several grammatical function assignments, the result in a word order freezing language is not ambiguity, but strict word order. We can distinguish three types of analyses of word order freezing. Freezing can be treated as a performance effect, freezing can be treated as a language internal, construction-related effect, or we can try to give a universal grammatical account of freezing.

The first way of dealing with word order freezing is to consider it to be a performance effect. Word order freezing would be related to processing rather than to the grammar of a language. Note that in this case there is no need for a model of word order to be bothered about freezing at all. The reason that canonical word order surfaces in a frozen sentence would be that it is easiest to process or the most frequently observed word order, or both. Such a dismissal of word order freezing as a task for a grammatical model of word order can be found in Müller (2002). In his paper on the typological relation between case and word order variation (scrambling), Müller explicitly considers word order freezing as non-grammatical and cites parsing difficulties as the reason for freezing. For Russian, King (1995) denies the relevance of data like (5) for a grammatical model of Russian word order. In a footnote she observes that, put in the right context, an OVS reading of the Jakobson sentence *is* available. An example to demonstrate this is provided by Bloom (1999, s2.3.2), presented in (6).

(6) S̶ina ljubjat vse ne doč' ljubit tol'ko mat'
son.ACC loves everyone NEG daughter.N/A loves only mother.N/A
'Everyone loves the son, but only the mother loves the daughter.'

In (6), the two nominative-accusative words *doč'* and *mat'* appear in an order that corresponds with an OVS reading of the sentence. According to King, this means that word order never (grammatically) fixes grammatical function assignment in Russian. Note that we are not forced to accept a datum like (6) as evidence against freezing. We saw similar effects of context in (4). However, it does mean that case syncretism alone is not enough to trigger freezing in Russian.

A 'mere processing' approach is not completely unreasonable. The strength of the freezing effect is somewhat unclear: It seems to vary between speakers and situations and it may be sensitive to factors like context and intonation, that are difficult to capture in a grammar model. Furthermore, freezing does not lead to ungrammatical sentences, but rather to the ungrammaticality of a sentence as expressing a certain meaning.[3] The latter is harder to recognize and pin down than the former. There is also a good positive motivation for a processing approach: The assumption that freezing is a processing effect might offer an explanation for why freezing is observed in such a diverse group of languages.

In spite of the difficult nature of part of the freezing-related data, there are also empirical reasons that make a processing-only account seem unlikely. Flack (2007) deals with scrambling in Japanese. Generally, in Japanese, an object is allowed to scramble over a subject, given the right information structure. However, in double nominative sentences, when both arguments are marked with *-ga*, word order freezes to SOV. Flack presents amongst other examples the following datum to demonstrate the strength of the Japanese freezing effect.

(7) *tenisu-ga Taroo-ga zyoozu-da
tennis.NOM Taroo.NOM is good at
Not: 'Taroo is good at tennis.' (OSV)[4]

The sentence in (7) does not allow an OSV reading, even though it is a very plausible one. Similar sentences with distinctive case marking do allow for an OSV reading. The unavailable, plausible interpretation of (7) would have to be expressed as SOV, that is, with *Taroo* initial. As Flack points out, if freezing is a processing effect, data like (7) are highly unexpected.[5]

Other studies of word order freezing have therefore attempted a more grammatical analysis, for instance by treating constructions that are subject to freezing as syntactically different from those that are not (for instance, Bloom, 1999, for Russian; and Tonoike, 1980,

[3]An example from another domain may help make this distinction between absolute and relative grammaticality a bit clearer. Consider the following pair of sentences.

(i) a. *Himself sees Findus.
b. *Findus$_i$ sees himself$_j$. ($i \neq j$)

Sentence (ia) is an example of *absolute ungrammaticality*, it cannot be used to express any meaning in English. Sentence (ib) is only ungrammatical relative to the meaning that I have indicated, in which Findus sees someone else. The form of (ib) can be used to express another meaning in English.

[4]It is unclear to which extent the completely absurd SOV reading, in which *tennis* is the subject and *taroo* the object, is available. The '*' is taken from the cited paper.

[5]Of course, the discussion only touches upon a much greater point, which is that one has to be willing to support a sharp distinction between grammatical facts and processing facts to in order to be selective about one's data in the way described in the text. Here I would just like to argue that even if one supports this distinction, freezing has properties that make it a candidate for grammatical modelling.

for Japanese). For a given language, one can create a fine-tuned construction-based account of freezing that covers exactly the right cases. The disadvantage of this highly grammatical approach is that freezing is treated as a completely language internal fact. The fact that freezing can be observed as a cross-linguistic tendency cannot be explained. Furthermore, it is not clear how a construction-based approach would deal with an example like (6), where information coming from the context prevents freezing.

In this dissertation I will therefore pursue the third type of approach, in which freezing receives a universal explanation. This approach sits in some sense in the middle of the processing approach and the construction approach. Freezing is explained by allowing universal, functional factors to enter the grammatical model. Freezing is treated in the grammar (like in the construction approach), but the grammar relies on universal principles to explain freezing (like in the processing approach). This does require that the grammatical framework that is used allows one to formulate and integrate these universal principles. Previous research has shown that Optimality Theory (OT) allows this. Recall from the discussion of word order freezing in Dutch that the expected ambiguity of grammatical function assignment did not in fact surface. A sentence like (2), p156, only receives an SVO interpretation. Zeevat (2006) and Flack (2007) thus consider ambiguity avoidance to be the cause of freezing. They propose special OT constraints that ban specific ambiguities, such as the ambiguity of grammatical function assignment. These constraints can be used in a model of word order alongside the constraints that capture the 'normal' cases of word order variation. Since OT constraints are assumed to be universal, this approach predicts that word order freezing is in principle possible in all free word order languages. In addition, because the anti-ambiguity constraints are violable, this approach also predicts that it is possible that some free word order languages do not show freezing or that languages show freezing in certain circumstances and in not in others.

A second group of OT researchers has taken a very different approach (Lee, 2001b; Kuhn, 2003; Vogel, 2004; Morimoto, ms). They argue that word order variation in general is contingent on communicative success. A certain word order variant is only grammatical when the speaker of the sentence can be certain that the sentence will be understood correctly. According to this second line of OT research, assessing whether a sentence will be understood correctly involves taking a hearer's perspective. The resulting models are referred to a *bidirectional models* because they model grammaticality by combining preferences of the speaker for a certain realization with preferences of the hearer for a certain interpretation. The cross-linguistic nature of word order freezing is explained by the assumption that bidirectionality is a universal property of natural language grammars.

The *unidirectional* approach of Zeevat and Flack, and the bidirectional approach will be extensively compared in this chapter. We will see that the unidirectional approach deals very well with the effects of the absence of distinguishing morphological marking on word order freedom, but that it cannot deal with the significance of other information about

grammatical function assignment such as animacy and definiteness. I will therefore follow the bidirectional approach, and in particular the work of Lee (2001b), since it is the most elaborate proposal. The comparison of the two approaches in this chapter will proceed as follows: I will start by discussing a (simplified version of) Lee's bidirectional model of word order in Section 5.3. After that, I discuss Zeevat's and Flack's unidirectional proposals and their criticism of bidirectional models, in Section 5.4. We will see that the unidirectional models are insufficient as models of word order in the context of freezing. Section 5.5 then extends Lee's bidirectional proposal, so that a range of word order freezing data can be correctly captured.

5.1.3 Relation with corpus study

As described in the preceding subsection, I will treat word order freezing as a grammatical phenomenon. In the model that I will develop in this chapter, word order variation is contingent on successful communication. A word order variant that does not meet this requirement is ungrammatical (in a relative sense). The assessment of successful communication will involve taking the perspective of the hearer. An important part of the research in this chapter will be to identify which sources of information a hearer has to help him in interpretation besides word order. When there is not enough word order independent information about the intended grammatical function assignment, word order is predicted to freeze to canonical word order.

In this dissertation, I have thus far looked at patterns in subject- and object-fronting. We may hypothesize that the effects of word order freezing are visible in the object fronting data of the spoken Dutch corpus, as well. After all, object fronting leads to a non-canonical order of arguments and this is only grammatical when there is enough additional information to help the hearer in interpretation. Hence, we should expect to see a lack of object topicalization when there is not enough additional information. However, this is not a very good hypothesis to be tested against corpus data. First, we have seen that word order freezing in Dutch is a weak effect, that disappears in the presence of information from morphology, from the context, from selection information, from knowledge, from intonation, etcetera. A speaker of Dutch may be confident that the hearer can use many types of information to retrieve the intended interpretation of a sentence. The problem with this is that we cannot be sure that we can identify all sources of information and the way that they steer interpretation. This problem arises in theoretical investigation as well as in a corpus study. Moreover, even if we could exhaustively list all factors that prevent word order freezing on the basis of theoretical contemplation, chances are that we cannot unambiguously recognize half of these factors in a corpus, automatically or manually.

But suppose we have identified all the relevant factors that prevent freezing and we know how to reliably get information about these factors from the corpus? Surely we

should be able to test our freezing hypothesis? Now we run into a second problem. In spoken language the information in terms of intonation, linguistic context, and extra-linguistic context about what a speaker is trying to say is often abundant. One might argue that we have no reason to expect to see any freezing at all, as the speaker can assume that there is always enough information for a hearer to figure out the correct message. The conditions under which freezing theoretically occurs in Dutch would never be met in real spoken discourse and we would not be able to measure freezing in our spoken Dutch corpus.

Therefore, I propose a gradient interpretation of word order freezing. The hypothesis that I will test against the corpus in the next chapter is that word order freezing is related to the *amount* of disambiguating information. For instance, in Section 5.5, I will argue that a Dutch hearer relies on animacy to distinguish subjects from objects. Subjects tend to be animate, whereas (direct) objects tend to be inanimate. A specific hypothesis that can be tested against a corpus is that topicalization is less frequent when animacy does not allow a hearer to correctly identify the subject – because the object is animate and/or the subject is inanimate. If word order freedom is related to the amount of disambiguating information, and not to the presence of disambiguating information *per se*, it is possible to investigate each source of information that we identify in this chapter independently from other sources.

The bidirectional OT model that I will develop is not capable of making such gradual predictions. In Section 5.6, I will discuss some of the obstacles that have to be overcome before the model can make such gradual predictions. However, this does not mean that the model in the current chapter is irrelevant for the empirical study of Chapter 6. On the contrary, the theoretical work is crucial: It allows us to identify the sources of information that a hearer in Dutch relies on, and it allows us to give a very precise account of how this information enables word order freedom by preventing freezing. Formulation of empirical hypotheses would not be possible without this work.

The cross-linguistically observed phenomenon of word order freezing surfaces in Dutch as a tendency. Without information that the first NP in an NP–V–NP sentence is an object, the sentence is interpreted as SVO. In general, this information can be of a morphological nature, but in Dutch it may also come in the form of selection restrictions, context or intonation. In the rest of the chapter, I propose a grammatical model of word order freezing in the framework of bidirectional Optimality Theory. Some of the predictions of this grammatical model will lead to frequency predictions that can be tested on a corpus in Chapter 6.

5.2 A brief introduction to Optimality Theory

Optimality Theory (OT, Prince and Smolensky, 1993/2004[6]) offers a constraint-based, competition view of grammar. *Constraint-based* means that grammaticality is expressed in terms of static constraints on what is grammatical, rather than by derivation rules that tell us how construct grammatical forms. What characterizes OT most is its conception of how constraints interact with each other. In classic constraint-based frameworks (say, HPSG, Pollard and Sag, 1994), a form has to satisfy all constraints in order to be grammatical. In contrast, in OT, constraints are ranked after importance and satisfaction of an important constraint can happen at the expense of any of the less important constraints. Thus, in OT grammaticality is defined as satisfying the constraints *best*, rather than satisfying them *all*. Determining which form satisfies the constraints best requires comparison of several forms that satisfy and violate the constraints in different ways, which puts *competition* at the core of OT.

Ranking constraints after importance has the advantage that constraints can be made very general. A central methodological assumption in OT is that constraints apply universally, that is, to all languages. Such a methodological principle would be hard to maintain if the constraints were not general in nature. The degree to which a constraint can actually be observed to hold in a particular language depends on its position in the ranking, which is language specific.

The OT framework is often described by deconstructing it into three abstract components: *generation*, *constraints* and *evaluation* (Prince and Smolensky, 1993/2004). In order to give the reader enough background in OT to read the rest of this chapter, I will discuss each of these components in turn below. To illustrate each component, I will use an example OT analysis which I will lay out as I introduce the components. The running example is a simplification of the work in Grimshaw and Samek-Lodovici (1998) and Samek-Lodovici (1996) on cross-linguistic differences in subject realization. Some languages will use expletive subjects to express zero-valent predicates like weather verbs, whereas others will use a subjectless clause. English is an example of the former kind: One says *it rains*, not **rains*. Italian is an example of the latter kind: *piove*, lit. 'rains', is the grammatical way of putting it, and not, for instance, **lei piove*, lit. 'she rains'. The analysis of these facts is based on the identification of two cross-linguistic trends. First, there is a tendency for clauses to have syntactic subjects. Secondly, uninterpreted material in a sentence is avoided. The difference between English and Italian is that when it comes to zero-valent predicates, English follows the first trend, whereas Italian follows the latter.

[6]The document was first published as a tech report at Rutgers University in 1993. In 2002, an updated version was made public on the Rutgers Optimality Archive (`roa.rutgers.edu`). The book was commercially published by Blackwell in 2004.

Generation (GEN)

Canonical OT assumes that optimality is a relation between an input and an output.[7] The output is selected from a set of candidate outputs (henceforth: candidates). This selection is based on a fitness measure called *harmony*, which is explained further in the subsection on the evaluation-component, below. The abstract component GEN refers to the mapping of an input to a set of candidates. The candidate set can be of any size and need not be finite in size. In the literature it is common to concentrate on a small, relevant subset of the candidates. Because the output comes from the set of candidates, the make-up of the candidate set is of great importance to the predictions an OT model makes.

Input and output are abstract notions that are used to refer to very different things in the different subfields OT is used in. OT originated in phonology. In OT phonology, in its simplest form, an input may be a string of phonemes, and the output a string of phones. Optimization proceeds from an underlying representation (the phonemes) to a surface form (the phones). The candidate set is, in this case, a set of possible realizations of the underlying representation that is given as input. Much work in OT morphology and OT syntax assumes the same relation between input and output. The input is a representation of underlying structure (morphemes, predicate-argument structure, a semantic meaning representation), the output is a, possibly structured, surface form (a word, a sentence, a syntactic tree). We will refer to approaches that optimize from an underlying representation to a surface form as taking a *production perspective*.

OT Semantics (Hendriks and de Hoop, 2001) takes the opposite approach. Here, the input is a form (surface form) and the output is a meaning (underlying representation). We will refer to this as the *comprehension perspective*.

Although it is in some sense a natural choice, taking a production perspective in syntax and a comprehension perspective in semantics is not a necessity. For instance, Anttila and Fong (2000) and Zeevat (2006) (to be discussed in detail in Section 5.4.2) argue that one should assign meaning to a sentence using a production model, in which the input to optimization is a meaning, and the output of optimization is a form. In such a production model of semantics, a sentence receives exactly those interpretations for which it is the optimal realization. Since what we have as a given in such an approach (that is, form) does not correspond to the input of optimization (meaning) and what we want to find out (meaning) does not correspond to the optimization (form),this approach seems counterintuitive, but it is formally unproblematic, and in some cases even preferable. In the OT Syntax literature, another dissociation between what is given and what is the input to optimization can be found. There are various proposals that add comprehension optimization to a syntactic production model, which results in a so-called *bidirectional* model. As announced earlier, in this chapter I will also assume a bidirectional model. To

[7]For an input-less proposal, see Heck et al. (2002).

avoid confusion, and because I do not see the need to draw a principled line between syntax and semantics in this chapter, I will consistently use the terms production and comprehension perspective when I need to refer to the direction of optimization. The terms OT Syntax and OT Semantics will only be used to refer informally to the literature.[8]

Irrespective of the perspective chosen, a common methodological principle in OT is that the mapping from the input to candidate set is as universal as possible. Differences between languages should be explained by differences in constraint ranking and not by 'parameterizing' GEN. For further discussion, I refer the reader to Kuhn (2003), who gives an explicit statement of this principle, and to Van der Beek and Bouma (2004), who argue that this methodological ideal cannot be fully upheld when one considers the consequence of lexical differences between languages. A related methodological principle is known as *Richness of the base* (Prince and Smolensky, 1993/2004). According to this principle, GEN should not reject any inputs, but provide a candidate set for any input. Again, the underlying idea is to shift as much of the explanatory burden as possible to the constraints and their interaction.

Now let us turn to the running example. We want to explain the fact that a meaning that does not involve a subject argument is expressed differently in English (with an expletive subject) and Italian (without a subject). In the model we take the production perspective. The input is a predicate-argument structure, the output a sentence expressing the input meaning. Given the methodological assumptions of OT, the difference between English and Italian cannot be due to GEN. English GEN and Italian GEN are one and the same. However, in order to make it clear which language we are modelling, I will use English lexical items for the English model, and Italian lexical items for the Italian one.

Given the input rain(), GEN provides two realizations. One is a clause containing only the appropriate weather verb (*rains* or *piove*). The other realization is a clause with an expletive subject (EXPL).[9]

(8) Input: rain()
Candidates: { [$_{IP}$VERB], [$_{IP}$ EXPL VERB] }

Constraints (CON)

The component CON describes the set of constraints and their ranking. Just like GEN, OT constraints are assumed to be universal: They apply to all possible languages. However,

[8]In the introduction to this chapter, and elsewhere in the dissertation, I use the terms speaker and hearer task. These should be understood as referring to the production and comprehension optimization, respectively. I will continue to use speaker and hearer in informal contexts outside of this chapter.

[9]In English, the word used for this is *it*. Italian does not have a lexical item for expletive subjects. Note that under the example model, the fact that Italian does not have an expletive in its lexicon is the result of the fact that it does not have to use an expletive subject in a grammatical sentence. If it would need to do so, it would have recruited some form for it.

the ranking of the constraints may differ from language to language. In classic OT, a grammar of a language is a complete ordering of all constraints.

As mentioned in the introduction, OT constraints are violable, which means that it is possible that optimal output candidates violate one or more constraints. This, in turn, means that violating a constraint does not equal ungrammaticality, since grammaticality is defined as optimality. Whether a constraint can be violated by a grammatical form depends first and foremost on its place in the ranking. A lower ranked constraint can be broken by a grammatical form, if this means that a higher ranked constraint can be satisfied.

There are two type of constraints in classic OT. The types are defined by the information that we need to assess constraint satisfaction. Depending on whether we need to have information about the input to evaluate an output candidate, we can discern *(output) markedness constraints* and *faithfulness constraints*. I will discuss the two in turn.

Markedness constraints A markedness constraint is a constraint that is defined solely with respect to the output. Satisfaction can be assessed by looking only at an output candidate. Depending on the perspective taken in the model, markedness constraints can constrain form (production) or meaning (comprehension).[10]

One of the constraints we will use in the running example is a markedness constraint. Recall from the introduction that the difference between English and Italian would be explained by appealing to two trends. The first of these revealed a preference for clauses with subjects. This preference is captured by the following markedness constraint.

(9) SUBJECT: Clauses have subjects (Grimshaw, 1997).

This constraint would be violated by a candidate output that contains a clause, but not a subject (that is, $[_{\text{IP}}\text{VERB}]$).

Faithfulness constraints A faithfulness constraint is defined with respect to both input and output. Using faithfulness constraints, one can state correspondences between the input and the output. In phonology, this correspondence can be very direct: The input and output can contain the same kind of features like nasality or voicing. Relative to the input, an output candidate may change such a feature, drop it altogether, or add features that were not part of the input. Faithfulness constraints punish such changes. In OT syntax and semantics, input and output are typically not of the same type. In that case, input and

[10]Defining constraints with respect to the *input* alone is pointless. These constraints cannot influence which candidate is optimal because the input is fixed. However, input markedness constraints may arise 'by accident' if one changes perspective, but keeps the same constraints. For instance, a output markedness constraint in a production grammar is a constraint on form. If we now change to a comprehension perspective, the constraint on form will be an input markedness constraint. Such a constraint does not differentiate between candidates and is therefore harmless.

output do not have features in common. Faithfulness constraints in syntax and semantics therefore also have to indicate in which way an input feature relates to an output feature (Kuhn, 2003, for an overview of existing work and a formal proposal on the basis of LFG). For instance, one could formulate a constraint that says that the agent in the meaning representation should be correspond to the grammatical subject in the form representation. There is no identity between the input and output features in this case; which input feature corresponds to which output feature is a matter of stipulation

An important difference between markedness and faithfulness constraints is that markedness constraints prefer the same output irrespective of the input, whereas faithfulness constraints may prefer different candidates depending on the input.

In our running example, the desire to not have uninterpreted material in a sentence is captured by a faithfulness constraint.

(10) FULL-INTERPRETATION: All lexical material is interpreted. Violated by words that do not correspond to a part of the input meaning.
(based on Grimshaw, 1997)

This constraint states that there should be a relation between the words in a sentence, and the meaning of the complete sentence given as the input. FULL-INT is violated by the use of an expletive subject because it does not correspond to anything in the input.

There is no candidate that would satisfy both constraints when the input is rain(). This is because the constraints put conflicting demands on the output. SUBJECT always prefers a sentence with a subject and, given the input rain(), FULL-INT prefers a sentence without a subject. Constraint ranking gives us a means of resolving this conflict. Ranking two constraints means saying that it is more important to obey the higher constraint than it is to obey the lower. Each possible ranking represents a grammar of a possible language. As a consequence, a constraint set of n constraints predicts that there are $n!$ (types of) languages, one for each ordering.

With the two constraints FULL-INT and SUBJECT we predict two types of languages, those that rank FULL-INT highest (we write FULL-INT $\gg$ SUBJECT) and those that rank SUBJECT highest (SUBJECT $\gg$ FULL-INT). The linguistic consequences of these rankings are spelled out below.

Evaluation (EVAL)

In EVAL, the optimal output is selected for the given input. This is done on the basis of a candidate set and a constraint ranking. To each candidate, we can assign a *constraint profile*, a record of which constraints are violated or satisfied by a candidate. A constraint profile codes therefore how well a candidate satisfies a set of ranked constraints – a property also referred to as *harmony*. The candidate with highest harmony is

selected as output. Relative harmony can be defined on candidates as follows (Prince and Smolensky, 1993/2004):

(11) $c \succ c'$ (one candidate is more harmonic than another)
iff
c has less violations of the highest constraint on which c and c' receive a different number of violations.

It is common in OT to compare candidates and their constraint profiles in so called *tableaux*. In a tableau one can see the input, the candidates (or a relevant subset of them), the constraint ranking, the constraint profiles, and which of the candidates is the optimal candidate. The tableau for our rain example, using the ranking SUBJECT $\gg$ FULL-INT, is in (12):

(12)

		rain()	SUBJECT	FULL-INT
a.		[$_{IP}$ rains]	*!	
b.	☞	[$_{IP}$ EXPL rains]		*

In the top-left corner we find the input. Along the top are the constraints, in ranking order from left to right. The candidates are down the left. The '*'s indicate constraint violations. A '!' indicates that a violation takes a competitor out of the competition, because it will never become the most harmonic candidate.

As discussed in the subsection on CON, there is no candidate that violates neither of the constraints. Still, candidate (b) is more harmonic than (a) because (b) satisfies the highest ranked constraint SUBJECT, and (a) does not. As there are only two candidates this means that (b) is the optimal candidate and the output. This is indicated by the '☞' in front of candidate (b). SUBJECT $\gg$ FULL-INT predicts that rain() will be realized with an expletive subject, as it is in English: *it rains*.

The other ranking, FULL-INT $\gg$ SUBJECT, predicts a language that prefers a subjectless clause for the input rain(), like Italian *piove*. This is shown in (13).

(13)

		rain()	FULL-INT	SUBJECT
a.	☞	[$_{IP}$ piove]		*
b.		[$_{IP}$ EXPL piove]	*!	

The grammar FULL-INT $\gg$ SUBJECT only predicts subjectless sentences when there is nothing that corresponds to the subject in the input. If the input would contain a subject, say run(forest), FULL-INT would not be violated by a clause that contains a subject.

As we have just seen, the optimal candidate is the most harmonic. By modelling the task of GEN as a function from an input to a set of candidates, and using the definition of harmony $\succ$, we can now also define optimality.

(14) Given an input i, an output candidate $o \in Gen(i)$ is *optimal* iff
there is no $o' \in Gen(i)$ such that $o' \succ o$

The definition in (14) is agnostic about what is form and what is meaning, so the production and comprehension perspectives are both captured in this definition. The production perspective involves taking a meaning as i, and a form as o, and comprehension involves taking a form as i and a meaning as o. Either way, we can conceive of a language as a set of form-meaning pairs such that one of the elements is the optimal output when the other one is used as input. The forms that are in these form-meaning pairs are the grammatical strings of the language, the associated meanings are the interpretations of these grammatical strings.

We can also define optimality directly in terms of form-meaning pairs. In that case GEN is not modelled as a function from an input to a set of candidates, but rather as the set *Gen* of all possible form-meaning pairs. A language is a subset of *Gen* such that the form-meaning pairs are optimal. As before, it depends on the ranking as well as on the optimization perspective which of these pairs are optimal. The two perspectives are now be characterized by two different definitions of optimality, as given in (15) and (16), after Blutner (2000, pp199–200).

(15) A form-meaning pair $\langle f,m\rangle$ is a *production optimal* iff
a. $\langle f,m\rangle \in Gen$
b. and there is no $\langle f',m\rangle \in Gen$ such that $\langle f',m\rangle \succ \langle f,m\rangle$

(16) A form-meaning pair $\langle f,m\rangle$ is a *comprehension optimal* iff
a. $\langle f,m\rangle \in Gen$
b. and there is no $\langle f,m'\rangle \in Gen$ such that $\langle f,m'\rangle \succ \langle f,m\rangle$

This separation is very convenient when it comes to formulating bidirectional models, as we will see in the next section.

Let us return to our example one last time. Under ranking FULL-INT $\gg$ SUBJECT, $\langle$*piove*, rain()$\rangle$ is a production optimal, whereas $\langle$it rains, rain()$\rangle$ is a production optimal under SUBJECT $\gg$ FULL-INT. Whenever, in the remaining part of this chapter, I refer to such pairs, they will represent form-meaning pairs, independent of the direction of optimization.

At this point I have introduced most of the architecture and concepts of classic OT. In the rest of the chapter, we will see several extensions of the concepts introduced in this section. Some of these extensions are bidirectional models, models with a third type of constraint (constraints that use more information than just input and output) and alternative views of what a language particular grammar is (variable ranking). However, these concepts are best introduced in the context of their application. The first of these,

bidirectional optimization, will be introduced and used in the next section, when I discuss a bidirectional account of word order freezing.

5.3 A bidirectional account of word order freezing

Lee (2001a; 2001b; 2002; 2004) studies the interaction of grammatical function and information structure and its effect on word order and case in Hindi and Korean. I will here confine my discussion of her analysis to the parts that are relevant to word order freezing as explained in the introduction. The discussion is primarily based on the work in Lee (2001b), which contains the most elaborate exposition of the subject.

5.3.1 Word order freezing in Hindi

Hindi is a free word order language. In general, all six permutations of subject, object, and verb are allowed in the right contexts. For instance, the SOV and OSV sentences in (17) are both grammatical. The subject in (17) bears ergative case. The object bears nominative case (no marking).

(17) a. Ilaa-ne yah k^{h}at lik^{h}aa.
Ila.ERG this.NOM letter.NOM wrote
b. Yah k^{h}at Ilaa-ne lik^{h}aa.
this.NOM letter.NOM Ila.ERG wrote
'Ila wrote this letter.'

SOV word order is considered to be neutral and basic. Other word orders, like OSV in (17b), can be used "to mark a special information structure and [are] generally associated with shifts in prominence, emphasis and semantic effects (e.g., definiteness effects)" (Lee, 2001b, p35).

For certain verb-argument combinations, both arguments bear nominative case. In these circumstances, and when there is no information from context or world knowledge available, Hindi shows word order freezing to subject- or agent-initial word order.[11] For instance, presented out of context, the double nominative sentences in (18) are both SOV.

(18) a. Patt^{h}ar t^{h}elaa todegaa.
stone.NOM cart.NOM break.FUT
'The stone will break the cart.' *Not:* 'The cart will break the stone.'

[11]Lee argues that canonical word order for Hindi should be phrased in terms of thematic roles, and not grammatical function. However, given the large overlap between the two, and in order to keep terminology consistent, I will refer to canonical word order as subject-initial or SOV. The dissociation between grammatical function and thematic role is visible in the passive in (21b).

b. T^{h}elaa patt^{h}ar todegaa.
cart.NOM stone.NOM break.FUT
'The cart will break the stone.' *Not:* 'The stone will break the cart.'

Because grammatical function assignment differs with word order between (18a) and (18b), they are interpreted as meaning different things.

Lee shows that word order freezing also occurs in Hindi with the morphologically ambiguous accusative and dative case markers, the multiply ambiguous *-se* marker (see below), and in Korean with double nominative constructions. In each case, morphology does not give us enough information to decide which NP should be assigned which grammatical function. Lee offers the following generalization:

(19) Generalization: Canonical word order [...] becomes fixed if the case markings on two nominal arguments of a single predicate are identical under two alternative thematic role interpretations of the nominals. (Lee, 2001b, p104)

The generalization in (19) does not cover cases in which it is clear how the NPs should be interpreted. Thus, the example in (20) does not show word order freezing in spite of the fact that there is no distinguishing case.

(20) a. Raam aam k^{h}aayegaa.
Ram.NOM mango.NOM eat.FUT
b. aam Raam k^{h}aayegaa.
mango.NOM Ram.NOM eat.FUT
'Ram will eat the mango.' (Lee's 8, p105)

Sentences (20a) and (20b) are interpreted as Ram eating the mango, although this means that (b) is O[NOM]S[NOM]V. Lee does not explore how strong a pragmatic bias has to be to prevent freezing. However, Lee also presents the example of freezing to agent-initial word order in (21). The marker *-se* is used amongst other things to mark a source and to mark a demoted agent in a passive (glossed as AG). The morphological marking is therefore ambiguous.

(21) a. Coor-se kal Ravii-se paise curaae gae.
thief.AG/SRC yesterday Ravi.AG/SRC money.NOM steal.PERF go.PERF
'Money was stolen from Ravi yesterday by the/a thief.'
Not: '...from the/a thief by Ravi.'
b. Ravii-se kal coor-se paise curaae gae.
Ravi.AG/SRC yesterday thief.AG/SRC money.NOM steal.PERF go.PERF
'Money was stolen from the/a thief yesterday by Ravi'
Not: '...from Ravi by the/a thief.' (Lee's 4, p103)

Example (21b) resists the natural interpretation where the thief is the doing the stealing in favour of the less natural, but far from absurd, interpretation where the thief is being stolen from. So, although pragmatic information in (20) is strong enough to prevent freezing, the pragmatic bias in (21) is not. An obvious difference between the two minimal pairs is that (20) involves an NP with an animate referent and an NP with an inanimate referent, whereas in (21) both *-se*-marked NPs are animate. In Section 5.5.5 we will see that this difference can be used to explain the lack of freezing in sentences like (20). The empirical investigation of whether it indeed is the difference in relative animacy of the NPs that causes the difference in freezing of (21) and (20) in Hindi lies beyond the scope of this dissertation.

5.3.2 Analysis

Lee presents an OT analysis of the Hindi data presented in the section above. I will simplify the account considerably for presentational purposes. The account is set in Lexical-Functional Grammar (LFG) and also deals to some extent with the mapping from proto-roles to grammatical functions and the assignment of case. These parts of the analysis are left out in my presentation. I refer the interested reader to Lee (2001b).

Lee starts out with a production oriented analysis of word order variation in Hindi. As mentioned subject-initial (agent-initial) word order is taken to be canonical, which is captured by a constraint that prefers subjects to be left-aligned in the clauses. The influence of information structure on word order is captured by constraints that require left-alignment of, for instance, topics.

(22) SUBJECT-LEFT: Subject aligns left in the clause
TOPIC-LEFT: Topic aligns left in the clause

Lee defines topic as material that is both given and prominent. Topicality is part of the input, but is unclear whether the topicality of material in the input follows from the context, or whether it is something that we freely choose to express. I will say more about the influence of the context at the end of this section, and in later sections. I will assume that the input is of the form PRED(ARG1,ARG2). Underlining in the input indicates topic status. Also, I will make the simplifying assumption that the mapping of parts of the input to grammatical roles is taken care of implicitly, so that the inputs can be reliably read as VERB(SUB,OBJ). Finally, correct case assignment is also implicit in the exposition below.

The SOV and OSV word orders in (17) are the result of different inputs. In (17a), SOV is the realization of an input where the subject is topic.[12] In (17b), OSV is the realization

[12] Alternatively, since SOV is the canonical word order, one might expect SOV to be the result of an input that does not specify a topic at all. However, I will concentrate on the simple two-way contrast between a topical subject and a topical object.

of an object-topic input. Lee predicts this input-driven variation in a production model with the ranking in (23).

(23) TOPIC-LEFT ≫ SUBJECT-LEFT

The tableaux in (24) show that the variation in (17) is captured under the constraint ranking in (23), by varying the topic in the input.

(24) a.

	write(Ila, this letter)	TOP-L	SUB-L
☞	Ilaa-ne yah k^{h}at lik^{h}aa		
	yah k^{h}at Ilaa-ne lik^{h}aa	*!	*

b.

	write(Ila, this letter)	TOP-L	SUB-L
	Ilaa-ne yah k^{h}at lik^{h}aa	*!	
☞	yah k^{h}at Ilaa-ne lik^{h}aa		*

There is no visible effect of SUBJECT-LEFT in the tableaux in (24). Because TOPIC-LEFT is the highest constraint, word order is completely driven by topicality as specified in the input.

Now let us turn to the case of word order freezing in (18). It turns out that the production model that we have so far does not distinguish between the cases in (17) and (18). Analogous to the prediction in (24b), the model predicts that an input with a topical object will be realized with an object-initial double nominative construction. This is demonstrated in (25).

(25)

	break(cart, stone)	TOP-L	SUB-L
	T^{h}elaa patt^{h}ar todegaa	*!	
☞	Patt^{h}ar t^{h}elaa todegaa		*

Thus we see that the production model predicts that the intended meaning ‘the cart will break the stone’ can be mapped onto the sentence *patt^{h}ar t^{h}elaa todegaa*, lit. ‘stone cart will break’. This mapping would mean that the sentence is OSV. However, from (18), we know that this is incorrect. The word order ‘stone cart will break’ only has the interpretation ‘the stone will break the cart’ (SOV). To put it differently: The model fails to predict word order freezing in double nominative cases.

In the generalization in (19), we could already see that Lee blames word order freezing on the lack of distinguishing case marking in the double nominative cases. Because case does not tell us that the first NP is not the subject, it is taken to be the subject. However, the production model that we have assumed so far does not refer to the presence or

absence of distinguishing case marking at all. It therefore cannot not capture the fact that in the absence of information to the contrary, subject-initial word order surfaces.

Lee proposes that we think of the lack of 'information to the contrary' in terms of *recoverability*. If the NPs are case marked in such a way that only one grammatical function assignment is possible, their grammatical function is recoverable from the case marking alone. When case marking does not make grammatical function assignment clear, identifying grammatical function has to rely on other, additional information. The extra information that is used for this purpose is word order. However, to be useful information, word order needs to be fixed: The first NP is always the subject. Word order freezes when case cannot guarantee recoverability.

Lee therefore adds a recoverability requirement to the grammar, which she formalizes with the help of comprehension optimization (see the definition in 16, p169 of this thesis). 'A meaning m is recoverable from a form f' is defined as 'm is a comprehension optimal for f'. Adding a recoverability requirement to a production grammar thus amounts to defining a *bidirectional* grammar. Bidirectional grammars combine the two optimization perspectives. The particular combination that Lee (2001b) uses is *strong bidirectional OT*.[13,14] The definition of grammaticality under strong bidirectional OT is given in (26), based on Blutner (2000, pp199–200).

(26) A form-meaning pair $\langle f, m\rangle$ is grammatical, iff

a. $\langle f, m\rangle \in Gen$

b. and there is no $\langle f', m\rangle \in Gen$ such that $\langle f', m\rangle \succ \langle f, m\rangle$ (production)

c. and there is no $\langle f, m'\rangle \in Gen$ such that $\langle f, m'\rangle \succ \langle f, m\rangle$ (comprehension)

In prose, the definition in (26) is a follows: A form-meaning pair is grammatical if the form is the optimal realization of the meaning, and the meaning is the optimal interpretation of the form.

The tableaux below show strong bidirectional OT in action. We start by demonstrating that bidirectional OT correctly handles the data in (17), p170: In the presence of distinguishing case marking, the mapping between OSV and topicality of the object is grammatical. In the first tableau in (27), we see that OSV is a production optimal if the input contains a topical object. In the second tableau, we see that we arrive at the original meaning in comprehension, given OSV as input. The fact that we after comprehension know that we have a bidirectional optimal, and not just any optimal, is indicated by the victory symbol '☙'.

[13]There are many ways to define grammaticality in OT that combine the two directions of optimization. Such approaches are collectively referred to as bidirectional OT. I will continue to talk about and use strong bidirectional OT in the rest of this chapter, only to return to other setups in the end of the chapter.

[14]In Lee (2001b), a so called *production/comprehension chain* is used (attributed to Smolensky). A meaning m is recoverable when we arrive at m after taking the optimal form f for m, and then taking the optimal meaning for f. However, this is just a procedural version of the static definition in (26).

(27) Production:

write(Ila, <u>this letter</u>)	TOP-L	SUB-L
Ilaa-ne yah k^{h}at lik^{h}aa	*!	
☞ yah k^{h}at Ilaa-ne lik^{h}aa		*

Comprehension:

yah k^{h}at Ilaa-ne lik^{h}aa	TOP-L	SUB-L
write(<u>Ila</u>, this letter)	*!	
✌ write(Ila, <u>this letter</u>)		*

The meaning candidates in the comprehension step in (27) both have the correct grammatical function assignment. This is due to case, which is handled implicitly. However, topic assignment is variable and because of TOPIC-LEFT, the first NP is selected as topic. This is the correct choice, for it leads us back to the meaning that was the input of production.

The situation is different, however, when case does not fix grammatical function assignment. In (28), we have the same kind of meaning (topical object) as in (27), but both arguments will bear nominative case. Apart from the actual strings, the production steps of (27) and (28) are identical. In comprehension, (28) has more possibilities because the case marking is compatible with SOV as well as OSV. The comprehension optimal candidate is not a bidirectionally optimal candidate, so it is indicated with a normal '☞'. The candidate that would have led to a bidirectionally optimal candidate is marked with '✌'.

(28)

break(cart, <u>stone</u>)	TOP-L	SUB-L
T^{h}elaa patt^{h}ar todegaa	*!	
☞ Patt^{h}ar t^{h}elaa todegaa		*

Patt^{h}ar t^{h}elaa todegaa	TOP-L	SUB-L
break(<u>cart</u>, stone)	*!	*
✌ break(cart, <u>stone</u>)		*!
☞ break(<u>stone</u>, cart)		
break(stone, <u>cart</u>)	*!	

As we can see in the comprehension tableau in (28), comprehension selects the candidate that assigns topic *and* subject to the first NP, since this candidate satisfies both constraints. However, the optimal meaning in comprehension was not the input of production. Therefore, the form meaning pair is ungrammatical. The bidirectional model correctly predicts the missing OSV interpretation of the double nominative cases in Hindi.

What remains is the SOV reading of double nominative cases. This reading is available, and the bidirectional model correctly captures this, too (29):

(29)

break(cart, stone)	TOP-L	SUB-L
☞ T^{h}elaa patt^{h}ar todegaa		
Patt^{h}ar t^{h}elaa todegaa	*!	*

T^{h}elaa patt^{h}ar todegaa	TOP-L	SUB-L
✌ break(cart, stone)		
break(cart, stone)	*!	
break(stone, cart)	*!	*
break(stone, cart)		*!

Lee points out that we can observe a phenomenon known as *emergence of the unmarked* (McCarthy and Prince, 1994) in comprehension optimization. Emergence of the unmarked refers to the situation in which the effect of a low ranked constraint becomes visible, because higher ranked constraints do not select an optimal candidate. In all cases of production that we have seen, and in the cases of comprehension where case marking took care of grammatical function assignment, the constraint SUBJECT-LEFT has no effect. But when the input of comprehension optimization is ambiguously case marked, SUBJECT-LEFT leaves its mark by causing the subject-initial interpretation to be optimal.

The situation in (28) also demonstrates something that is not easily achieved under unidirectional OT: An input is not assigned an output. Because one candidate always wins, we are also certain that anything that serves as input is actually mapped on to an output.[15] That is, an input is never ungrammatical. However, in the bidirectional case, some input (be it form or meaning) may not be mapped onto an output, because the output fails to map back onto the input.

Lee (2001b) goes on to define constraints that allow the context to play a role in recovering the correct grammatical function assignment. As she points out, comprehension optimization is well suited to allow all kinds of information to help identify the correct reading. I will not discuss this part of her proposal here, because I want to concentrate on the basics of a bidirectional analysis of freezing. However, in Sections 5.5.3 and 5.5.4, I will return to the importance of context in a proper analysis of freezing.

Lee shows that bidirectional OT offers an elegant explanation of word order freezing in Hindi (and Korean). Freezing occurs when non-canonical argument order is not recoverable from the case marking on the argument NPs. Comprehension optimization

[15]There are proposals to allow a so called Null Parse as a candidate: nothing of the input is realized. This way, one can have a input receive no output. See Ackema and Neeleman (2000) for a syntactic application.

provides a formalization of a recoverability requirement, that can be added to an existing production analysis. As such, one can take word order freezing as an argument in favour of a bidirectional conception of grammar. Other authors have also argued for the bidirectionality of grammar on the basis of freezing (Kuhn, 2003; Vogel, 2004; Morimoto, ms). Although sometimes differing in emphasis, detail and formal framework, these authors all note that a straightforward bidirectional model of word order would rule out incorrect predictions that would be made by a unidirectional production model. In the next section, I will review recent criticism of bidirectional accounts of freezing, and discus two alternative extensions to unidirectional OT that have been proposed in Zeevat (2006) and Flack (2007) to account for freezing. I will proceed to show that these alternative unidirectional accounts do not predict the data correctly, and that a bidirectional model is called for after all.

In the formal concepts of production and comprehension, we can recognize what I informally have called the speaker's task and the perspective of the hearer. In the introduction, I pointed out that the speaker's word order freedom may result in a failure of communication because the hearer may misinterpret an utterance. In a bidirectional model, a form-meaning pair is only grammatical if the pair is optimal in both directions of optimization. The word order preferred by a speaker is the production optimal candidate. However, this word order is only allowed when it will be understood correctly. Whether a word order will be understood correctly can in turn be assessed by taking the hearer's perspective in comprehension optimality. The requirement on word order variation that communication be successful is thus embodied in the combination of production and comprehension itself.

The claim that grammar is bidirectional is very strong, because bidirectional OT grammars behave very differently from unidirectional ones. In general one cannot assume that existing results from the OT literature carry over bidirectional models. Therefore, if we want to use word order freezing as an argument in favour of bidirectional OT grammars, we should be certain that bidirectional OT actually makes the right predictions about word order variation and freezing. In the next section, we will see that the (simple) model as presented so far, runs into problems.

5.4 Against a bidirectional account?

Recently, the bidirectional accounts of word order freezing as presented in the previous section have been criticized for being too strict. These objections come from two unrelated sources, but interestingly, the proposed solutions have a lot in common (Zeevat, 2006; Flack, 2007). In this section I will lay out the problems put forward in these two papers and I will also discuss the proposed solutions. At the end of the section, I will show that the criticism is only partly justified. I will argue that the criticism shows that we have to

take the interpretation part of the bidirectional model more seriously, rather than discard the whole architecture. The gist of the objections in both papers is that bidirectional OT systematically undergenerates: Too many form-meaning pairs are not bidirectionally optimal because the optimization directions fail to converge.

5.4.1 Problems with a bidirectional account of freezing

Zeevat (2006) argues against a bidirectional account of word order freezing on two grounds. As a first point of critique, Zeevat puts forward a well known problem with many kinds of bidirectional optimization, the so called *Rat/Rad* problem (Hale and Reiss, 1998). Consider a language with final devoicing, like German or Dutch. In these languages, underlying forms that differ only in voicing of a coda consonant should come out the same. So both /ra:t/ (German *Rat*, 'council/advice', pl [rɛ:tə]) and /ra:d/ (German *Rad*, 'wheel/bicycle', pl [rɛ:dɐ]) are pronounced [ra:t].

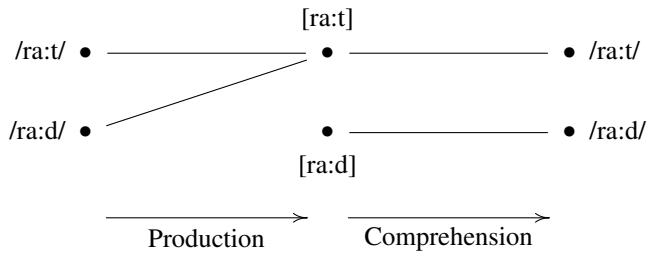
The loss of an underlying (semantic) distinction in the surface form is referred to as *neutralization*. Neutralization is straightforwardly modelled in production OT. For instance, final devoicing can be modelled by assuming that a faithfulness constraint FAITH[VOICE] is outranked by a markedness constraint that bans final voicing *VOICEDCODA. The markedness constraint prefers the final consonants in both *Rat* and *Rad* to be voiceless. Since it outranks the faithfulness constraint that prefers a voiced final consonant in the case of *Rad*, both *Rat* and *Rad* are produced with a voiceless final vowel. The underlying voicing distinction is thus lost or neutralized.

However, we run into a problem if we try to retrieve the original underlying distinction through comprehension optimization. In comprehension, only /ra:t/ is optimal for the input [ra:t], because of FAITH[VOICE]. As a result, only ⟨[ra:t],/ra:t/⟩ is bidirectionally optimal. In Figure 5.1, I have drawn a production-comprehension sequence showing the non-recoverability of the underlying voicing contrast. Going from left to right, we can move from /ra:t/ to [ra:t] (production), and from [ra:t] to /ra:t/ again (comprehension). The pair ⟨[ra:t],/ra:t/⟩ is grammatical in bidirectional OT. There is no such route that starts at /ra:d/. The underlying representation /ra:d/ is not part of a bidirectional optimal pair.

In Section 5.3, we saw that the loss of a reading in the case of freezing was exactly the effect we wanted to achieve by moving from uni- to bidirectional OT. But in the Rat/Rad problem, the meaning loss is at odds with the facts: [ra:t] *is* ambiguous between wheel and council in German, and the model should therefore predict both pairs ⟨[ra:t],/ra:t/⟩ and ⟨[ra:t],/ra:d/⟩ to be grammatical. Since ambiguity due to neutralization is common in the phonological domain, one can question the suitability of bidirectional OT for phonology.

Zeevat claims (2000; 2006) that in the domain of word order we can find counterparts to Rat/Rad. Wh-fronting in a language like German or Dutch may lead to a semantic/syntactic version of the problem. Zeevat gives (30) for German:

Figure 5.1: Production-comprehension sequences showing the Rat/Rad problem



(30) Welches Mädchen liebt Peter?
which.NEUT girl loves Peter
'Which girl loves Peter?' *Or:* 'Which girl does Peter love?' (his example 7)

Zeevat considers this sentence to be completely ambiguous between a subject question and object question.[16] In production optimization, this ambiguity could easily be modelled by assuming a constraint on wh-fronting WH-LEFT, that outranks constraints linking arguments to word order like SUBJECT-LEFT. The ambiguity in (30) is then a matter of neutralization, since the same wh-initial form will be the production optimal candidate for both interpretations of (30). In comprehension, however, wh-initial sentences will lead to S[wh+]VO interpretations when, as in (30), there is no case or agreement that says otherwise. WH-LEFT does not differentiate between SVO and OVS meanings, and thus SUBJECT-LEFT decides in favour of S[wh+]VO. As in Rat/Rad, an observed ambiguity that can be modelled in production optimization is lost in the bidirectional model. So, it might seem that a production-based, unidirectional account would be preferable after all.

Zeevat's second point of critique concerns the interpretation of information structure in frozen sentences. The simple bidirectional model introduced in Section 5.3, as well as the model cited by Zeevat (that is, Lee, 2001a), predict that a frozen sentence is only compatible with a topic-initial reading. This is because comprehension not only fixes the assignment of grammatical function, but also the assignment of topichood. The model predicts that information structure freezes, too. For instance, consider the Hindi sentence in (31). The model correctly predicts that only SOV is possible as a grammatical function assignment. However, the model also predicts that there is only one topic assignment possible.

[16] In his discussion, Zeevat even goes so far as to say that the "psychological prediction[s] [...] that Rad is more difficult to recognize than Rat, that it is more difficult to recognize *Welches Mädchen liebt Peter* (subj) than *Welches Mädchen liebt Peter* (obj) [...] are not plausible." (Zeevat, 2006, p1100). Although I have no reason to disagree on the Rat/Rad part, we shall see later on that there *is* good evidence that O[wh+]VS is harder to parse than S[wh+]VO.

(31) botal patt^{h}ar todegaa
bottle.NOM stone.NOM break.FUT
break(bottle, stone)
Not: break(stone, bottle)
Also predicted not: break(bottle, stone) (Zeevat's example 2, due to Lee, 2001a)

In comprehension, SUBJECT-LEFT and TOPIC-LEFT together prefer the first NP to be topic and subject. However, Zeevat considers this to be undesirable in the case of freezing. After all, he reasons, freezing is interesting exactly because it concerns the exceptional circumstance that argumenthood governs word order in spite of information structure, where normally information structure governs word order. Just like the neutralization of grammatical function in wh-questions leads to ambiguity with respect to grammatical function, one could say that information structure is neutralized in frozen word orders and that this leads to ambiguity of information structure. The sentence in (31), according to Zeevat, should also have the object-topic reading that is represented in the bottom line of (31).

I will take Zeevat's arguments 'as is' for now. However, I would like to point out that the question of whether (31) has an object-topic reading is hard to answer abstractly, without considering what it means to have a subject- or object-topic reading. I will return to this issue in Section 5.5.3 and argue that, if one takes context into account, Zeevat's claim does not stand up. With respect to the OVS reading of wh-questions, we shall see in Section 5.5.2 that we can exploit linguistic knowledge about what counts as a good subject to get the OVS readings in a bidirectional model.

Flack (2007) looks at scrambling in Japanese and she too discusses a case in which freezing is 'violated' as in Zeevat's wh-questions. Let us start by looking at the Japanese freezing data that Flack presents. In Japanese, two nominal arguments can scramble to yield OSV sentences in addition to the canonical SOV. Scrambling of an NP is triggered by information structural properties like topic and focus. In (32), we see scrambling with the verb *osore* 'to fear', which takes a nominative and an accusative argument.

(32) a. Taroo-ga Hanako-o osore-ru
Taroo.NOM Hanako.ACC fears
'Taroo fears Hanako.' (SOV)
b. Hanako-o Taroo-ga osore-ru
Hanako.ACC Taroo.NOM fears
'Taroo fears Hanako. ' (OSV)
c. Jishin-o Taroo-ga osore-ru
earthquake.ACC Taroo.NOM fears
'Taroo fears earthquakes.' (OSV)

However, there is a class of stative verbs that take two nominatives. The verbs in this class do not allow the scrambling illustrated in (32). The data in (33) demonstrates this with the stative verb *kowa* ('to be afraid of').

(33) a. Taroo-ga Hanako-ga kowa-i
Taroo.NOM Hanako.NOM is afraid of
'Taroo is afraid of Hanako.' (SOV)
Not: 'Hanako is afraid of Taroo' (OSV).

b. Jishin-ga Taroo-ga kowa-i
earthquake.NOM Taroo.NOM is afraid of
'Earthquakes are afraid of Taroo.' (SOV)
Not: 'Taroo is afraid of earthquakes.' (OSV)

Flack claims that Japanese double nominative constructions exhibit word order freezing, *even* in cases like (33b), where animacy information could disambiguate towards OSV (see also the discussion between Tonoike, 1980, and Kuno, 1980, and references therein). Japanese shows a very strong freezing effect. However, freezing only affects topic scrambling. Japanese focus scrambling, in which the fronted element is a focus rather than a topic, is not subject to freezing:

(34) TAROO-GA Hanako-ga kowa-i
Taroo.NOM Hanako.NOM is afraid of
'TAROO is afraid of Hanako.' (SOV, subject focus)
Or: 'Hanako is afraid of TAROO.' (OSV, object focus)

Focus scrambling is a case in which the strong freezing tendency of Japanese is overridden. In Zeevat's terms, we might call focus scrambling in Japanese a syntactic Rat/Rad. We have seen that Rat/Rads are problematic in a bidirectional model. A more general point that Flack makes is that basically *all* ambiguity is banned in a bidirectional model. This is because in comprehension optimization, one of the meanings of a would-be ambiguous expression is bound to be more harmonic than another. Banning all ambiguity indeed seems to be far too strong a claim, and if this is a property of all bidirectional models, this would cast serious doubt on the suitability of bidirectional OT as a grammatical framework. I will show in Section 5.5, that a strong bidirectional OT model can successfully model ambiguity (and its counterpart optionality) if we allow underspecified constraint rankings (for instance Anttila, 1997; Boersma and Hayes, 2001) and pay close attention to the *dis*ambiguating information we have available in comprehension.

Zeevat and Flack's points against the bidirectional accounts of freezing amount to the same thing: A fully bidirectional account is too strict. Before I address these points in Section 5.5, I will discuss the counterproposals Zeevat and Flack offer. I will present the

models more or less at the same time, because they are very similar in spirit and suffer from much of the same problems.

5.4.2 Extending unidirectional OT to model freezing

Consider the Dutch sentence in (35) as an example of word order freezing. The combination of the facts that the word order in (35) is in principle compatible with SVO and OVS in Dutch, and that the two NPs in (35) are compatible with either argument role might lead us to expect that (35) is fully ambiguous. This expected ambiguity is not observed, and the sentence only has SVO as a clear interpretation.

(35) De jongens zoeken de meisjes.
the boys search the girls
'The boys look for the girls.'

The bidirectional model introduced in Section 5.3 ascribes this to the fact that a hearer makes the default assumption that a sentence is in canonical word order. There is no information to the contrary, so the default assumption leads to an SVO interpretation of (35). Because grammaticality in a bidirectional model is defined in terms of speaker and hearer preferences (production and comprehension optimality), sentence (35) is only assigned an SVO reading in a bidirectional model.

In this subsection, I will discuss alternative ways to model word order freezing that have been put forward by Zeevat and Flack. Zeevat and Flack identify the desire to avoid certain ambiguities in language as the source of freezing. Under this approach, sentence (35) is not ambiguous between SVO and OVS, because a language like Dutch does not tolerate grammatical function ambiguities. Of course, in Zeevat's and Flack's criticism of the bidirectional model, we came across several examples of grammatical function ambiguity. In an OT setting, it is natural to solve this paradoxical situation by modelling the dislike of grammatical function ambiguity as a violable constraint. This constraint can then be used in a regular production model of word order to ban certain types of word order variation, whilst permitting others.

The lion's share of Zeevat's and Flack's efforts lies in making it possible to formulate such a special anti-ambiguity constraint in a production model. This is not trivial, which becomes clear if we consider what ambiguity is from a production perspective: A form is ambiguous when it is the winner of several production competitions with different input meanings. Hence, a constraint that refers to ambiguity must be able to refer to the optimization of more than one input. This is radically different from what the classic faithfulness and markedness constraints are allowed to do: A classic constraint may refer to one candidate and the input. A constraint that refers to several alternative inputs requires an adjustment of the system.

Zeevat and Flack propose different modifications to OT to allow for such unusual constraints. However, suppose we have a constraint that is violated by ambiguous candidate forms. And suppose that we have a way to model word order freezing with such a constraint. Why would modelling freezing using a constraint in a production model be a solution to the problematic cases that were presented in Section 5.4.1, above? Let me answer this question per case.

1. Dutch/German wh-questions are not subject to freezing. Since the mechanism that causes word order freezing is captured in a constraint, its effect can be overridden by higher constraints. As mentioned, wh-fronting can be captured by a constraint WH-LEFT. If this constraint outranks the freezing-constraint, word order freezing will not stop a wh-constituent from being fronted. Under a unidirectional production model, a form is ambiguous if two distinct meanings select it as their optimal candidate (neutralization). A subject question and an object question will both be mapped onto a wh-initial form because of the highly ranked WH-LEFT. The wh-initial form is therefore ambiguous between SVO and OVS.

2. Frozen word order is ambiguous as regards topicality. If we can model word order freezing in a unidirectional production model, we almost get the ambiguity of topic assignment for free. As with the previous point, we need two meanings to be mapped onto the same form. Consider a model in which object topicality in the input leads to OVS, except in the circumstances under which word order freezing occurs. In the case of word order freezing, an object-topic input yields canonical word order SVO. Since subject-topicality also leads to canonical SVO, we have two different topic assignments mapped onto SVO. As a result, frozen SVO is ambiguous with respect to topic assignment.

3. Japanese focus scrambling defies freezing. Japanese focus scrambling is modelled in the same way as wh-fronting in Dutch is modelled. The constraint that demands focus to be left aligned outranks the constraint causing word order freezing. A focussed subject will scramble under this ranking, as will a focussed object. If this causes two different input meanings to be mapped onto one output form, we have the desired ambiguity in Japanese.

What remains is to discuss how the two authors modify classic OT to model freezing. I mentioned above that this requires that constraints are able to refer to different input-output pairs. Zeevat's solution is to allow special constraints that directly refer to alternative competitions. Flack's solution is to change the conception of what a competition is: In a classic production model, the input is one meaning and each candidate is one form, in Flack's proposal the input is a cluster of meanings, and each candidate is a cluster of forms.

Alternative competitions

Zeevat proposes a new family of constraints he calls MAX-constraints. The name derives from a subfamily of faithfulness constraints in classic OT, that require a feature in the input to be present in the output. For instance, MAX(VOICE) in phonology would require an output segment to have [+voice] when the corresponding segment in the input does. MAX(VOICE) is not violated by a voiced output segment that corresponds to a voiceless input segment.

The idea is that Zeevat's MAX-family also requires a feature in the input to be expressed in the output. However, this correspondence in syntax/semantics is not as straightforward as in phonology. Apart from the fact that in syntax/semantics we are dealing with two radically different types of objects (meaning and form in syntax/semantics do not relate to each other like phonology's underlying form and surface form), Zeevat notes that there are many ways in which a meaning feature like subjecthood can be expressed in form, even in one and the same language. Two ways that have figured prominently in the discussion so far are case and word order. Zeevat also argues that the *grammar* should ultimately determine how a meaning feature is expressed in the output, and that the range of possible correspondences cannot be established *a priori*. On this assumption, a syntactic MAX constraint cannot be expressed simply as: 'if the input has meaning feature *x*, the output should have form feature *y*', because we do not know what the form feature will be.

Instead, Zeevat suggests that a meaning feature is expressed in a form, when that form is different from all other forms that are paired with meanings that do not carry this feature. Zeevat dubs this *marking* a feature. The definition of marking is in (36). It already shows the characteristics of an OT constraint.

(36) "A form marks dimension X for an input I if and only if the form lacks interpretations which are exactly as I but have a different value for the dimension X. A [violation – gjb] mark is given for every variation of the input for the dimension for which the form is also optimal." (Zeevat, 2006, pp1002–3)

Two dimensions of meaning we are interested in are information structure and grammatical function. Zeevat uses the following two constraints:

(37) MAX(TOPIC): Topicality should be marked
MAX(θ): Thematic information should be marked

In order to evaluate these two constraints in one competition, we will have to consider other competitions whose input differs only in thematic information (for MAX(θ)), and competitions whose input differs only in what is topic (for MAX(TOPIC)). To calculate the outcome of optimizing the expression of hit(piet, jan), Zeevat includes competitions for three other inputs: two resulting from changing either what is subject or what is topic, and one from changing both these features.

In addition to the marking constraints, we need a constraint on word order. Zeevat defines the constraint PROMINENCE, given in (38).

(38) PROMINENCE: Prominent things come first.
Violated once for every prominent meaning feature that is not expressed in the first constituent. In effect: subjects and topics come first.

It is not clear what Zeevat means by prominence, although it is clear that it is a property that both topics and subjects have. It is important to understand the nature of the PROMINENCE constraint. Putting constituents first is not inherently a means of marking topicality or subjecthood for Zeevat. Fronting a constituent might not even inherently be a means of marking at all. According to Zeevat, it is because of their prominence that subjects and topics want to come first. In a language particular grammar, it might turn out that word order is used as a distinguishing feature for subjects or topics, but this is a mere side effect of the interaction of PROMINENCE with the MAX constraints. The constraints are ranked as in (39).

(39) PROMINENCE $\gg$ MAX(θ) $\gg$ MAX(TOPIC)

Note that ranking MAX(θ) $\gg$ MAX(TOPIC) will lead to freezing, because it expresses that we will prioritize marking thematic information over marking topicality. If word order is needed to mark thematic information, it may do so at the expense of marking topicality.

In (40), I give the tableaux for the four relevant competitions. Competitions that are in the same row differ only in which participant is topical, competitions in the same column differ only in thematic role assignment. This grouping is functional because MAX(θ) can now be understood as 'be different from the winner downstairs/upstairs', and MAX(TOPIC) as 'be different from the winner on your left/right.' As before, topicality is marked by underlining, and the predicate-argument structures in the input are consistently read as VERB(SUBJECT,OBJECT).

(40)

hit(jan, piet)	PRM	M(θ)	M(T)
☞ Jan slaat Piet			*
Piet slaat Jan	*!*	*	

hit(jan, piet)	PRM	M(θ)	M(T)
☞ Jan slaat Piet	*		*
Piet slaat Jan	*	*!	

hit(piet, jan)	PRM	M(θ)	M(T)
Jan slaat Piet	*	*!	
☞ Piet slaat Jan	*		*

hit(piet, jan)	PRM	M(θ)	M(T)
Jan slaat Piet	*!*	*	
☞ Piet slaat Jan			*

PROMINENCE can be evaluated for each competition quite easily, and in fact picks a

winner when subject and topic coincide (top-left and bottom-right). Topical subjects have to be placed first. This makes evaluating the MAX constraints in the other competitions easier. For instance, we know that candidate ⟨*Piet slaat Jan*, hit(jan, piet)⟩ (top-right) violates MAX(θ): Using this form would lead to same form being associated with two meanings that are alike except for their thematic roles because ⟨*Piet slaat Jan*, hit(piet, jan)⟩ (bottom-right) is already established as an optimal. In the same fashion, all the other violation marks can be filled in in (40).

The model predicts freezing: The object-topic inputs (top-right and bottom-left) are realized as subject-initial. We can also see that ambiguity of information structure is no problem, since competitions that are in the same row have the same winner. In fact, we see that all winning forms are ambiguous with respect to topichood, and violate constraint MAX(TOPIC).

The tableaux in (40) demonstrate that Zeevat's extended unidirectional proposal can model word order freezing. Word order variation can be achieved in two ways. When the subject or object is case marked, MAX(θ) is satisfied in all candidates since subjecthood is 'Zeevat-marked', in the sense of (36). When MAX(θ) is satisfied, MAX(TOPIC) will selects topic-initial word order as the output. Alternatively, non-canonical word order can occur when it is selected by a constraint that outranks MAX(θ) and PROMINENCE, for instance WH-LEFT.

The way in which a language particular marking device follows as a side effect is attractive, because it precludes the need for postulated correspondences between meaning and form. However, there are some problems with the account. Two of them have to do with the lack of a calculation procedure and with the formulation of PROMINENCE, and I will them discuss here. A third problem is a general problem with a purely production view of affairs, and is discussed in Section 5.4.3, as it addresses a problem common to Zeevat's and Flack's solutions.

The first problem with Zeevat's account concerns the evaluation of MAX constraints. Zeevat (2006) does not give us a calculation method and it is not obvious that we can find a suitable one. To evaluate a MAX constraint on a candidate in one competition, we have to rely on the complete outcome of another competition. This other competition however contains MAX constraints, too – *its* outcome may rely on the outcome of the first competition or on a third competition, and so forth.

The success of the semi-parallel procedure we have followed in the tableaux in (40) above, in which the constraint profiles in each competition were constructed at roughly the same time, relied on the fact that PROMINENCE determined the winner of two of the competitions (top-left and bottom-right in 40). This allowed us to calculate the outcome of the other two competitions (top-right and bottom-left), by evaluating MAX(θ). With all four competitions decided, scoring the rest of the candidates with respect to the MAX-constraints is trivial. But if the relevant alternative competitions are not decided

when we want to evaluate a MAX constraint, we have a problem. Hypothetical reasoning about whether assigning a MAX violation will lead to a contradiction-free set of violation assignments across competitions might be a way to solve this problem. But even using that strategy it is possible to arrive at more than one stable configuration of violations.[17]

The second problem with Zeevat's account is that MAX only has an effect because PROMINENCE does not decide all of the competitions. The constraint equally favours subject- and topic-fronting. Although I appreciate the way in which PROMINENCE might explain subject- and topic-fronting instead of just positing it directly, I do think it problematic that one cannot refine the relation between prominence and subjecthood or topichood without ruining the effect of MAX. For instance, we might want to be able to say that subjects are more prominent than topics (or vice versa), or that it is more important to front a subject than a topic (or vice versa). After all, there are languages in which subject-first word order is nearly categorical, so even if we do not want to differentiate topics and subjects in our freezing languages, we want to be able to do so with our set of *universal* constraints. However, if we were to replace PROMINENCE with different constraints for subject- and topic-fronting, MAX would have no effect because subject fronting and topic fronting would no longer ever be equally good under the classic ranking assumptions. As a way out, we could have some redundancy in the system by having PROMINENCE alongside more specialized constraints that we rank low in freezing languages. Another option would be to use a non-standard method of leaving the subject-fronting constraint and the topic-fronting constraint unranked, so that in one competition the candidates that front the subject and the candidates that front the topic are equally good (so-called *local constraint ties*, Müller, 2001). Either way, the model calls for further assumptions and mechanisms before it can be used in a cross-linguistic setting.

Cluster-based optimization

We have seen that Zeevat modified OT by proposing a class of constraints that could refer to other competitions. This meant he was able to formulate a (family of) anti-ambiguity constraints MAX, with which he could model freezing. Flack (2007) explores a second option. Instead of allowing special constraints to refer to information in places unreachable in classic OT, she moves all the information needed into a place that classic constraints already have access to. Flack builds an OT model capable of capturing freezing that is inspired by Dispersion Theory (Flemming, 1995). In Dispersion Theory

[17]Zeevat (p.c.) suggests that first the highest constraint is evaluated for all candidates in all competitions, and then the next, etc. The special MAX constraints can be marked as violated if their form is still a *possible* winner at that stage in an alternative competition. I will not investigate this calculation method here. As I have indicated in the main text, there are more serious problems associated with the production perspective, which will be discussed later in the section.

(phonology), inputs consist of *clusters* of underlying forms and outputs consist of clusters of corresponding surface forms. As a result, normal constraints can compare more than one pair of surface form and underlying form. Thus, in a cluster-based approach, all the information we need to evaluate anti-ambiguity constraints is present in the input (different meanings) and each candidate (different forms).

In a cluster setup, we can define constraints on individual members of a candidate cluster or on corresponding pairs of members of the input cluster and members of a candidate cluster. These constraints correspond to the classic markedness and faithfulness constraints. It is also possible to define 'special' markedness constraints that say something about the relation between members of a candidate cluster – for instance that they be the same or different in some aspect. Similarly, we can define special faithfulness constraints that compare members of the candidate clusters and their corresponding input cluster members. An example of the latter type of constraint is a constraint that requires that a contrast between members of the input cluster is preserved in the output cluster. Flack (2007, s2.3) defines the constraint in (41), which demands preservation of s contrast in subjecthood.

(41) PRESERVECONTRAST(SUBJECT): "Inputs with different subjects must map to separate outputs."
"Given two pairs of input-output correspondents I, O and I', O' where O and O' are in cluster C, if $\text{Subject}(I) \neq \text{Subject}(I')$ and $O = O'$, assign one violation to C."

Flack thus defines an anti-ambiguity constraint that targets subject ambiguity alone. This is exactly what Zeevat accomplished with his MAX(θ) constraint.

Flack's definition of inputs and outputs also differs in another way from what we have seen until now. Flack assumes a minimalist setup, in which the mapping to syntactic structure and the mapping to PF are separated into two optimization steps. I will only consider the second step in detail because this is where Flack places word order freezing.

In the two-step model, object-initial word order in Japanese is achieved by creating a chain of copies of a topical object in the mapping to syntactic structure (step 1), and pronouncing only the first of these in the mapping to PF (step 2). This is represented in (42). The striked-out members of the chain in (42c) are not pronounced.

(42) a. *input to step 1:*
fear(hanako, <u>taroo</u>)
b. *output of step 1/input to step 2:*
Taroo-ACC Hanako-NOM Taroo-ACC osore-ru
c. *output of step 2:*
i. ~~Taroo-o~~ Hanako-ga Taroo-o osore-ru
ii. Taroo-o Hanako-ga ~~Taroo-o~~ osore-ru

There are several stages at which a topical object may fail to be pronounced in the initial position. First, we can see in (42c) that the mapping to PF can choose not to pronounce the initial copy, as in (42c i). Secondly, creating a chain of copies is optional in the mapping to syntactic structure. This is illustrated in (43).

(43) a. fear(hanako, taroo)
b. Hanako-NOM Taroo-ACC osore-ru
c. Hanako-ga Taroo-o osore-ru

There is only one output at PF in (43c), because there is no material to delete from a chain of copies.

The constraint PR[ESERVE]CONTR[AST](SUBJECT) is used at PF, alongside a constraint that prefers to pronounce the first element in a chain of copies (44).

(44) MAX(HEAD): Do not delete the head of a chain.

The constraints on PF are ranked PRCONTR(SUBJECT) $\gg$ MAX(HEAD). Flack creates input clusters for the mapping to PF by combining several outputs of the mapping to syntactic structure. I refer the reader to Flack (2007) for details on the derivation of these input clusters.

Let me briefly repeat the discussion of freezing in Japanese on page 181. Freezing in Japanese is observed with stative verbs like *kowa* 'to be afraid', that take two nominative arguments. An example is the sentence in (45). The sentence only has an SVO reading, even though Japanese otherwise allows for OSV.

(45) Hanako-ga Taroo-ga kowa-i
Hanako.NOM Taroo.NOM is afraid
'Hanako is afraid of Taroo.' (SOV)
Not: 'Taroo is afraid of Hanako.' (OSV)

In a production OT model, a form is assigned those meanings for which the form is the optimal output (Section 5.2). To show that Flack's model correctly predicts that (45) has the SVO reading, but not the OSV reading, we have look at the optimization of both inputs.

The syntactic structures corresponding to the SVO and OVS readings of (45) occur in the same input cluster in two forms each: one with a chain of copies of the object, and one without – recall that creating copies of the object was optional in the mapping to syntactic structure. There are thus four syntactic structures in the input cluster in the tableau (46). I have boldfaced the subject in each member of the input cluster.

(46)			PRC(SU)	M(HD)
		i. **Hanako-NOM** Taroo-NOM kowa-i ii. Taroo-NOM **Hanako-NOM** Taroo-NOM kowa-i iii. **Taroo-NOM** Hanako-NOM kowa-i iv. Hanako-NOM **Taroo-NOM** Hanako-NOM kowa-i		
a.	☞	i. Hanako-ga Taroo-ga kowa-i ii. ~~Taroo-ga~~ Hanako-ga Taroo-ga kowa-i iii. Taroo-ga Hanako-ga kowa-i iv. ~~Hanako-ga~~ Taroo-ga Hanako-ga kowa-i		$*^{\text{ii}}*^{\text{iv}}$
b.		i. Hanako-ga Taroo-ga kowa-i ii. Taroo-ga Hanako-ga ~~Taroo-ga~~ kowa-i iii. Taroo-ga Hanako-ga kowa-i iv. ~~Hanako-ga~~ Taroo-ga Hanako-ga kowa-i	$*^{\text{ii,iii}}$!	$*^{\text{iv}}$
c.		i. Hanako-ga Taroo-ga kowa-i ii. ~~Taroo-ga~~ Hanako-ga Taroo-ga kowa-i iii. Taroo-ga Hanako-ga kowa-i iv. Hanako-ga Taroo-ga ~~Hanako-ga~~ kowa-i	$*^{\text{i,iv}}$!	$*^{\text{ii}}$
d.		i. Hanako-ga Taroo-ga kowa-i ii. Taroo-ga Hanako-ga ~~Taroo-ga~~ kowa-i iii. Taroo-ga Hanako-ga kowa-i iv. Hanako-ga Taroo-ga ~~Hanako-ga~~ kowa-i	$*^{\text{ii,iii}}$! $*^{\text{i,iv}}$	

The violations of the constraints are indexed by numbers that indicate which members of the cluster cause the violation. MAX(HEAD) can be evaluated on one member of a cluster alone. It is violated whenever the left-most copy in a chain is deleted. PRCONTR(SUBJECT) is evaluated on pairs of members – both members are given as superscripts. PRCONTR(SUBJECT) is violated whenever two identical forms in a cluster have different subjects. *Hanako* is the subject of members (i) and (ii), *Taroo* is the subject of members (iii) and (iv). So, PRCONTR(SUBJECT) is violated whenever (i) and (iii) sound alike, etcetera.

The winner of the competition is cluster (a), because this cluster does not violate PRCONTR(SUBJECT). The cluster does not contain any sentences that have different subjects but sound the same. Note that the initial copy of the object is never pronounced in this cluster, which means that there are no OSV sentences, and only SOV sentences. The model thus correctly captures the freezing of (45). The non-optimal cluster (d) shows what would have happened if we would have scrambled the double nominative sentences. Scrambling the object over the subject would lead to ambiguity of subjecthood: Members (ii) and (iii) are pronounced the same, as are members (i) and (iv).

As in the Zeevat model, there are two ways to achieve non-canonical word order in Flack's model. When there is morphological marking that distinguishes the subject from

the object (for instance, case), PRCONTR(SUBJECT) is not violated by pronouncing initial objects, and MAX(HD) selects the winning cluster: the case-marked counterpart to (46d). Alternatively, a constraint may prefer scrambling even though it leads to ambiguity of subjecthood. Flack proposes a special version of MAX(HEAD): MAX(HEAD)/FOCUS. The constraint is ranked above PRCONTR(SUBJECT), so that initial copies of focussed objects are pronounced even when it leads to subject-ambiguities.

We have seen two different implementations of in essence quite similar approaches to the problem of modelling freezing from a production perspective. The two approaches have in common that they place ambiguity avoidance at the heart of the analysis of word order freezing. Constraints against specific ambiguities are special constraints that can compare form-meaning pairs. Both models correctly predict freezing. Moreover, it is straightforward to account for the problematic cases that Zeevat and Flack put forward.

This is not to say that the two production models are without problems. I have already indicated that I have doubts about the computational feasibility of Zeevat's (2006) proposal. Both models, however, have fundamental empirical shortcomings. In the next section, I will discuss a range of data that the models fail to explain. Also, I will argue that the reason for this failure is exactly the production nature of the models.

5.4.3 Problems for unidirectional production models

An important feature of the Flack/Zeevat models is that it depends on the comparison of identical forms whether freezing is triggered or not. The anti-ambiguity constraints responsible for freezing – respectively PRCONTR(SUBJECT) and MAX(θ) – are violated when two identical forms correspond to different grammatical function/thematic role assignments. As a result, all and only information present in the form is used as disambiguating information in the Flack/Zeevat models. Below, I will discuss data that cannot be modelled by such an approach. The terms *form information* and *non-form information* are meant to refer to information that is reflected in word order or morphology and information that is not visible in form but that relates to meaning or reference. I will begin by discussing data that shows that language may differ in whether non-form information prevents freezing. After that, I will present data of a more tentative nature that suggests that freezing can occur in spite of form information that is present.

Flack (2007) is very clear in claiming that non-form information does not influence word order freezing in Japanese[18]. To back this claim up, she presents cases like (33b), repeated below in (47).[19]

[18]Zeevat (2006) does not consider the issue of non-form information at all, but his model behaves in exactly the same way as Flack's with regard to this type of information.

(47) Jishin-ga Taroo-ga kowa-i
earthquake.NOM Taroo.NOM is afraid of
'Earthquakes are afraid of Taroo.' (SOV)
Not: 'Taroo is afraid of earthquakes.' (OSV) (her example 8)

On the basis of animacy information, selection restrictions and/or knowledge of the world, one can easily understand that *jishin-ga* is an unlikely subject. Still, this sentence and others like it reportedly show freezing (see Tonoike, 1980, and Kuno, 1980). We can conclude that non-form information like the animacy of the argument NPs or plausibility of a reading does not prevent freezing in Japanese.

A similar claim for Russian can be found in Bloom (1999). Consider the example in (48).

(48) *Kofe da'ot mat' pap'e
coffee.NOM/ACC gives mother.NOM/ACC father.DAT
Not: 'The mother gives the father coffee.' (DO V S IO) (his example 18)

Again, *kofe* is a highly unlikely subject, and a good object, but freezing is still observed.

The Japanese and Russian data are as predicted by the Flack/Zeevat models. For instance, the sentence in which *jishin-ga* is the object and has scrambled over *Taroo-ga* is string-identical with the canonical sentence in which *jishin-ga* is the subject. Flack's PRCONTR(SUBJECT) and Zeevat's MAX(θ) constraints are violated, and the scrambled construction that corresponds to the plausible reading is not optimal.

Even if the Japanese and Russian data above is as solid as it is presented, it is certainly the case that there are freezing languages that will allow non-form information to prevent freezing. In fact, we have already seen one example in this chapter that demonstrates this. Lee's (2001b) model does not cover cases like (20), p171, here repeated as (49).

(49) aam Raam k^{h}aayegaa.
mango.NOM Ram.NOM eat.FUT
'Ram will eat the mango.' (OSV)

The fact that 'mango' is a perfect object of 'to eat', and 'Ram' a good subject, makes the OSV interpretation available. The OSV interpretation is available even though case marking is ambiguous between SO and OS.

[19] All judgements in this subsection are from the cited papers, except for the Dutch examples. For the Russian example in (48) it should be noted that Bloom assumes a 'non-emotive' intonation, which is canonical in the sense that focus/nuclear stress falls towards the end of the sentence. My two Russian informants disagreed with the alleged freezing of (48), one of the informants denied the existence of freezing to begin with. A Swedish informant, on the other hand, gave SVO as the predominant reading for (50) when presented in isolation, contrary to Morimoto's claim. Since this chapter is not intended as a cross-linguistic empirical study of freezing, I will stick to the judgments as supplied in the cited papers. What is important, however, is that Flack/Zeevat would have trouble integrating sensitivity for non-form information, as I will show in this subsection.

Similar examples have been reported for Swedish. Morimoto (ms) claims Swedish shows word order freezing to SVO when both arguments are animate, and no form information is present. However, when the object is inanimate, and the subject is animate, OVS is possible, as in (50):

(50) Boken såg Anna
book.DEF saw Anna
'The book saw Anna.' (SVO)
Or: 'Anna saw the book.' (OVS) (OVS is Morimoto's example 13)

Finally, it seems that in Dutch, too, freezing is readily prevented by animacy information (my judgement).

(51) Het koekje eet Jan op.
the biscuit eats Jan VPART
'The biscuit eats Jan.' (SVO)
Or, possibly preferred: 'Jan eats the biscuit.' (OVS)

Both *Jan* and *het koekje* are morphologically possible subjects for *eet op*, but unlike other, similar, cases in Dutch, word order does not freeze to SVO.

The final example is of a slightly different nature. The Dutch sentence in (52) allows an OVS reading although there is no distinguishing case or agreement present. The difference with the previous examples is that both NPs are equal in animacy. The OVS reading appears to be based purely on plausibility.

(52) Zo'n klein jongetje slaat alleen een heel grote bruut.
such a small boy beats only a very big brute
'Such a small boy would only beat a very big brute.' (SVO)
Or: 'Only a very big brute would beat such a small boy.' (OVS)

To be fair, I have to mention that I have received mixed judgments on this sentence. It was provided to me by Henk Zeevat in personal communication, without the word *heel* 'very'. Zeevat himself does not accept (52) as a proper realization of the most plausible reading, and prefers SVO or 'brute-initial' word order, instead. However, informally asking several other native speakers yielded both SVO/OVS and even OVS-only judgments. None of my informants found the sentence ungrammatical or expressed the desire to reorder the arguments. For the OVS reading there seems to be a strong intonational preference, which has a rise on *zo'n* and a fall on *heel*. This intonation renders the sentence rather emphatic about the relative size and innocence of the two characters. To my ears, the presence of *alleen* also makes a big difference in the acceptability of OVS. This, of course, fits well with the corpus findings about the relation between the presence of certain particles and the frequency of OVS in Section 4.6 (Chapter 4).

Under the Flack/Zeevat models, the data from Hindi, Swedish, and Dutch come as a surprise. In each case, there is no form indication of what the subject and object are. Therefore the OVS sentences are string-identical with SVO sentences. The Flack/Zeevat models predict freezing, but no such freezing is observed.

Non-form information presents a problem for Flack/Zeevat models, and it seems that this is closely related to the fact that the models are production oriented. Consider three ways in which one might attempt to make Flack/Zeevat models to predict the Hindi, Swedish and Dutch data.

1. Posit constraints on meaning. The solution that comes to mind first is the one that is suggested by the discussion of the examples. The unfrozen OVS readings in Hindi, Swedish, and Dutch are available because they are plausible readings. We know that Hans eating a biscuit is a more likely state of affairs than the biscuit eating Hans. For three of the four examples we could also offer a more grammatical analysis: We know about subjects that they are often animate. The animate NP is therefore better suited as a subject than the inanimate one.

It is easy to implement preferences for animate subjects, or dispreferences for animate objects in a constraint (see for instance Øvrelid, 2004 for a comprehension model of argument recognition in Norwegian). However, such constraints only have an effect in a comprehension model. A constraint against, say, animate objects is an input-markedness constraint in a production model where grammatical function is fixed by the input. As explained in the introduction to OT (Section 5.2), input-markedness constraints have no effect on unidirectional optimization. Similarly, a hypothetical constraint that meanings should be 'plausible' is an input-markedness constraint and would not improve the Zeevat/Flack models.

2. Force inanimate NPs to front. A second strategy would be to argue that there is a high ranked constraint requiring *inanimates* to front. That is, we treat the fact that we can front inanimates in Swedish, Hindi, and Dutch in the same way as we treat the fact that we can front a wh-constituent in German or a focussed constituent in Japanese in spite of freezing. Unfortunately, this would not fit well with what we have come to expect on the basis of cross-linguistic data – a much less controversial assumption would be the opposite: *animates* front. Also, a constraint that promotes fronting inanimates would not explain the example in (52), since the fronted object is animate.

3. Make the anti-ambiguity constraint selective. A third way out would be to somehow prevent the anti-ambiguity constraint from comparing *boken såg Anna* (SVO) and *boken såg Anna* (OVS). If the selective anti-ambiguity constraint is not violated, object fronting can go through as usual, and OVS is available. However, I fail to see a principled and systematic way of achieving this.

I conclude that there is little hope of capturing the exceptions to freezing presented above in a Zeevat/Flack-like model. In fact, the most obvious analysis of the data, proposed as the first solution above, suggests that a production model will not suffice at all, and that a model that includes comprehension optimization is called for. This means that a bidirectional model, as proposed by amongst others Lee (2001b), may have been on the right track after all.

Finally, it has been claimed in the literature that freezing may occur even when formal evidence of non-canonical word order is present, although this case is not nearly as clear as the opposite situation we have just seen. Bloom (1999) gives the following example. In Russian, verbs in the past tense show gender agreement (boldfaced in gloss) with the subject.

(53) Ditja videlo myš'.
child.**NEUT**.NOM/ACC sees.PAST.**NEUT** mouse.FEM.NOM/ACC
'The child sees the mouse.' (his example 28)

Interestingly, Bloom notes that OVS in this case is ungrammatical (54) and claims that this is due to freezing in spite of the disambiguating gender agreement on the verb.

(54) *Myš' videlo ditja. (his example 29d)

Freezing would be responsible for this ungrammaticality by the following reasoning. Word order freezing means the sentence is interpreted as SVO. But if *myš'* is the subject, the verb should show feminine agreement *videla* in a grammatical sentence. Hence, the sentence in (54) is ungrammatical because the verb does not agree with its subject.

If the data in (54) is correct and the reasoning is, too, it would pose a severe problem for Flack/Zeevat: *myš' videlo ditja* is not the same string as *myš' videla ditja*, and would not trigger the contrast preserving/marking constraint needed to rule out the former.

This example is problematic, however. Grammaticality is involved, rather than the resulting reading, which makes comparison with other freezing cases less obvious. It might be that there is something else about this sentence that causes the ungrammaticality. Also, one of my Russian informants has pointed out that *videlo* and *videla* are normally pronounced identically. If this has interfered with the judgements of Bloom's informants, despite written presentation, Flack/Zeevat may simply claim that there is string identity, and that they therefore predict this data correctly.

5.4.4 Summary

We started this section with a presentation of drawbacks of the simple bidirectional model of word order freezing introduced in the previous section. Concretely, we have seen that

the bidirectional model incorrectly predicts that wh-questions in Dutch and German are subject to freezing, that frozen sentences do not receive an object-topic reading in a bidirectional model (both objections from Zeevat, 2006), and that a bidirectional model incorrectly predicts Japanese focus fronting to be subject to freezing (Flack, 2007). The more general tenet of the criticism was that bidirectional OT is too strict because it cannot deal with ambiguity or optionality. The exceptions to freezing in German, Dutch, and Japanese lead to ambiguity: Object fronting yields a string that could also be analyzed as subject-initial. To this we can add that Flack (2007) considers topic-fronting in Japanese to be optional. Bidirectional OT has as much of a problem with optionality as it does with ambiguity.

Ambiguity is no problem whatsoever in a unidirectional production model. A form is ambiguous when two different meanings are mapped onto it, that is, when it is the optimal candidate in two different production competitions. Zeevat and Flack independently propose similarly spirited production models of freezing. The Zeevat/Flack models feature constraints that require differences in grammatical function assignment to be visible in form. Word order freezing occurs when word order is the only way of showing this difference. Zeevat and Flack have to go beyond classic OT in order to be able to formulate the needed constraints, but the resulting models do successfully predict word order freezing as well as some exceptions to freezing. Ambiguity follows for free under a production model.

Unfortunately, unidirectional production models have their own set of problems. Most importantly, some freezing languages allow word order variation in the absence of case and agreement, when non-form information like animacy can be used to distinguish the subject from the object. This is quite easily explained in a comprehension perspective, but a pure production approach struggles with these facts. Generally put, the analysis proposed by Zeevat/Flack overestimates the role that form information plays in word order freezing. We have seen data that shows that form information is not the only type of information that can prevent freezing.

To summarize, we have yet to see a model that can capture all of the freezing data that we have considered so far: the ‘standard’ cases of freezing, the exceptions to freezing presented by Zeevat and Flack, and the exceptions to freezing due to animacy/plausibility. However, if we were able to extend the bidirectional model in such a way that ambiguity is possible, we might be in a very good position to capture all of the data. In the next section, we will have a detailed look at what stands in the way of ambiguity in a bidirectional model. We will see that we can get past this obstruction and capture almost all of the data we have seen thus far in a bidirectional model.

5.5 A bidirectional account of freezing, revisited

In the previous section we saw that the bidirectional model of word order freezing fails to account for certain data. A fundamental shortcoming of the strong bidirectional setup is that ambiguity can not be modelled. However, we have also seen data that strongly suggested that the comprehension perspective is important in word order freezing – or rather, the absence of word order freezing – and that therefore a bidirectional model is a more likely candidate for a comprehensive account of word order than a unidirectional production model. In this section, I will begin by extending the bidirectional model so that ambiguity is possible under a strong bidirectional regime. After that, I will give analyses of almost all data presented in this chapter, mostly using constraints that have been independently motivated in the OT literature.

The solution that I will lay out below is a natural extension of Lee's proposal. Lee herself has suggested it as future work in several places (Lee, 2001a; Lee, 2001b). It has been worked out to some extent in an unpublished manuscript on ambiguity and word order in Kikuyu and Sesotho (Lee, ms). Lee (ms) allows constraint ties in the grammar as in Anttila (1997), and shows that under such a ranking regime, the bidirectional model is flexible enough to deal with ambiguity. Lee (ms) assumes a so-called *weak bidirectional* model, using a recursive definition of optimality, where candidates have to be optimal in order to be allowed in the competition in the first place. In this dissertation, I will also use Anttila's ranking regime to achieve ambiguity in bidirectional OT, but I will continue to use strong bidirectional OT. I put off discussing other types of bidirectional OT, as well as other ranking regimes, until the end of the chapter (Section 5.6).[20]

5.5.1 Ambiguity and optionality in bidirectional OT

It has long been pointed out that ambiguity and optionality are hard in bidirectional OT (see for instance Hale and Reiss, 1998, who introduce the Rat/Rad problem). Let us start by considering why this is the case. We speak of ambiguity when a form f appears in at least two grammatical form-meaning pairs $\langle f, m_1 \rangle$ and $\langle f, m_2 \rangle$. Optionality can be considered the opposite case, where one meaning m appears in two grammatical pairs $\langle f_1, m \rangle$ and $\langle f_2, m \rangle$ (Asudeh, 2001). To understand why ambiguity and optionality are hard in bidirectional OT, one has to understand how they relate to unidirectional OT.

[20]Much of the work presented below, especially the spelling out of properties of the proposed architecture, was done following Lee's (sometimes detailed) suggestions in published work, but *independent* of Lee's manuscript. Nevertheless, the model in the manuscript and the model presented in this dissertation have much in common because of the way variable ranking interacts with bidirectional OT in general. The constraints are mostly different since the constructions addressed are different. Lee's manuscript, which dates from 2000, is not currently on-line, and can only be retrieved through the Internet archives at `archive.org`. The fact that it is hard to find could explain why papers like those of Zeevat and Flack do not consider the model presented in it. See the bibliography for a download location.

Beaver and Lee (2004) observed that classic unidirectional OT deals naturally with either ambiguity or optionality depending on the optimization perspective. Ambiguity is easy in production, but hard in comprehension. Optionality is easy in comprehension, but hard in production. The reason for this is that mapping two inputs to one output is easy – it is just neutralization. Neutralization of a meaning contrast in unidirectional production leads to ambiguity, whereas neutralization of a form contrast in unidirectional comprehension leads to optionality.

Mapping one input to two outputs, which I will refer to as *variation*, is harder to achieve in classic OT. Classic OT guarantees that there is at least one winning candidate, because there is exactly one winning constraint profile. If we want two candidates to be optimal in a competition, we would have to make sure that they have exactly the same, winning constraint profile. This can be often be done for smaller grammars, but with larger grammars, this becomes cumbersome or even impossible. Also, each time such an analysis is extended with more constraints, variation is at risk because the constraint profiles change (see Müller, 2001, for extensive discussion of approaches to optionality in production OT, and Bíró, 2006, for discussion and a performance-based approach to variation). It is better to use a more controlled way of achieving variation, like Anttila (1997), introduced below.

Figure 5.2, p199, shows (amongst other things) the relation between neutralization, ambiguity/optionality and optimization perspective. Optimization proceeds from left to right. Converging lines '$>$' indicate neutralization. Diverging lines '$<$' indicate variation. An optimal form-meaning pair under unidirectional production optimization is a pair such that we can start with a meaning at the left, and travel to a form in the middle. Starting at either m_1 or m_2 at the left, we can travel to f in the middle. The lines are converging, so both pairs $\langle f, m_1 \rangle$ and $\langle f, m_2 \rangle$ are optimal in a production model. Starting at m at the left, we only have diverging lines. Diverging lines represent variation, which we cannot model yet. So at this point we can only really travel to one of f_1 and f_2 at a time. Either $\langle f_1, m \rangle$ or $\langle f_2, m \rangle$ is optimal in the production model, but not both. Optimal form-meaning pairs for a unidirectional comprehension model can be read from Figure 5.2 in the same way, but now one has to start from a form in the middle and end at a meaning on the right. In comprehension both $\langle f_1, m \rangle$ and $\langle f_2, m \rangle$ are optimal, but normally only one of $\langle f, m_2 \rangle$ and $\langle f, m_2 \rangle$ is optimal.

Before I introduce Anttila (1997) to capture variation, let me connect these facts about unidirectional OT to strong bidirectional OT. The definition for strong bidirectional OT given in Section 5.2 is repeated below in (55):

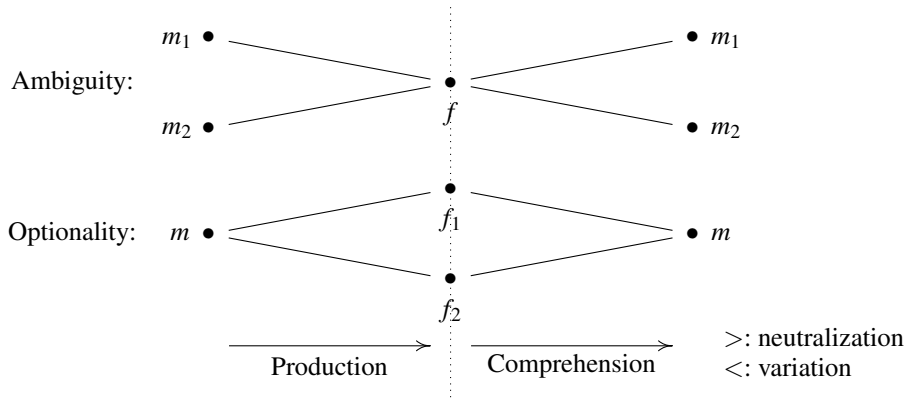
(55) A form-meaning pair $\langle f, m \rangle$ is grammatical, iff

a. $\langle f, m \rangle \in Gen$

b. and there is no $\langle f', m \rangle \in Gen$ such that $\langle f', m \rangle \succ \langle f, m \rangle$ (production)

c. and there is no $\langle f, m' \rangle \in Gen$ such that $\langle f, m' \rangle \succ \langle f, m \rangle$ (comprehension)

Figure 5.2: Production (left) and comprehension (right) and their relation to ambiguity (top) and optionality (bottom)



A strong bidirectionally optimal pair has to be a production optimal as well as a comprehension optimal. The set of such pairs is thus the intersection of the set of production optimal pairs and the set of comprehension optimal pairs. The first of these sets does not contain different pairs with the same meaning (optionality), the second does not contain different pairs with the same form (ambiguity). The intersection therefore only contains pairs that have different meanings and different forms, that is, the set contains neither ambiguity nor optionality (Beaver and Lee, 2004). With respect to ambiguity and optionality, strong bidirectional OT gets the worst of both worlds.

A bidirectionally optimal form-meaning pair in Figure 5.2 is a form-meaning pair such that, starting at a meaning on the left, we can travel through a form in the middle back to the original meaning, but now on the right. Note that we always meet diverging lines in this way. Therefore, only one of $\langle m_1, f\rangle$ and $\langle m_2, f\rangle$, and one of $\langle m, f_1\rangle$ and $\langle m, f_2\rangle$, can be bidirectionally optimal.

As mentioned in the introduction to OT in Section 5.2, a language-particular grammar is a complete ranking of all universal constraints in classic OT. The language described by the grammar is the set of all optimal form-meaning pairs under that ranking. Anttila (1997) captures variation in production by taking a different view of what a language-particular grammar is. He proposes to use underspecified, partial rankings that can be described by putting constraints in so-called *strata*. A language-particular grammar is a distribution of the universal constraints over a number of strata. Constraints in different strata are ranked with respect to each other, constraints within strata are unranked with respect to each other. The language described by a stratified grammar is the set of form-meaning pairs that are optimal under any of the complete rankings compatible with the stratified grammar.

When two constraints in one stratum conflict, we predict variation. For instance, take the constraints we used earlier SUBJECT-LEFT and TOPIC-LEFT. Also assume we have a language in which SUBJECT-LEFT and TOPIC-LEFT are in the same stratum. We write: {SUBJECT-LEFT, TOPIC-LEFT}. Given an object-topic input, the two constraints conflict: SUBJECT-LEFT prefers SO, but TOPIC-LEFT prefers OS. The language described by our stratified grammar will therefore have both: SO from SUBJECT-LEFT $\gg$ TOPIC-LEFT, and OS from TOPIC-LEFT $\gg$ SUBJECT-LEFT.

Anttila (1997) also presents a way of deriving approximate corpus frequencies from stratified grammars. However, in this section I will only be interested in the fact that we can have variation, and not in the relative frequencies of the variants. See Section 5.6 for discussion.

The example I used to illustrate how a stratified ranking can capture variation was a case of production optionality (of topic fronting). We already had comprehension optionality. In principle, we are therefore able to have optionality in bidirectional OT, too. In terms of Figure 5.2, we can now travel over both diverging lines at the same time. So, we can now travel from m through both of f_1 and f_2 (variation), to m again (neutralization).

Since ambiguity in comprehension is just the opposite of optionality in production, we can use stratified grammars to achieve ambiguity in comprehension (Asudeh, 2001). And as a result of this, we can in principle model bidirectional ambiguity, too.

We can adapt the definition of strong bidirectional optimality to include the new ranking scheme. To do this, we need to make explicit that the harmony order $\succ$ depends on a complete constraint ranking (see also Prince and Smolensky, 1993/2004; Kuhn, 2003). I propose the definition of *stratified strong bidirectional OT* as in (56):

(56) A form-meaning pair $\langle f,m\rangle$ is grammatical in a stratified grammar S (a set of compatible full rankings), iff
  a. $\langle f,m\rangle \in Gen$
  b. and there is an $s \in S$ such that
    i. there is no $\langle f',m\rangle \in Gen$ such that $\langle f',m\rangle \succ_s \langle f,m\rangle$ (prod.)
    ii. and there is no $\langle f,m'\rangle \in Gen$ such that $\langle f,m'\rangle \succ_s \langle f,m\rangle$ (comp.)

Note that by letting the selection of the full ranking s from the set of possible full rankings described by a stratified ranking S out-scope the production and comprehension optimality clauses, I am assuming that production and comprehension happen under one and the same ranking. I have no motivation for this choice other than that it is restrictive and that it suffices for the grammars in this section. I will briefly return to the alternative (production and comprehension can each select a full ranking) in Section 5.6

By modelling variation with a stratified grammar, ambiguity and optionality lies within reach of the strong bidirectional OT. This does not mean that we have resolved the Zeevat

and Flack counterexamples. Modelling these examples will involve finding constraints that cause neutralization and variation in exactly the right way. In fact, we shall see that cases of ambiguity/optionality require a *double explanation*. We have to find one or more constraints that cause neutralization, and we have to find two or more conflicting constraints that cause variation when they are in one stratum.

In the next subsection, I will begin with the ambiguity of wh-questions. After that, Subsection 5.5.3 looks in detail at Zeevat's claim that frozen canonical sentences are ambiguous as regards topichood. Subsection 5.5.4 deals with the interaction between focus scrambling and freezing in Japanese (Flack), and Dutch. After I have analyzed the three Zeevat/Flack counterexamples and have shown that real cases of ambiguity and optionality can indeed be handled by strong bidirectional OT, I will proceed to demonstrate that the influence of animacy on word order freedom can be elegantly captured by the bidirectional model, as well as the (tentative) data on gender agreement in Russian (Subsection 5.5.5).

5.5.2 Wh-questions

According to Zeevat, wh-questions in German and Dutch are not susceptible to freezing (Section 5.4.1). The German sentence in (57a) is morphologically compatible with SVO as well as OVS because the case marking is ambiguous. The result is an ambiguous sentence, and not word order freezing. Similarly, the Dutch question in (57b) is also ambiguous, and not frozen.

(57) a. Welches Mädchen liebt Peter?
which.NOM/ACC girl loves Peter
'Which girl loves Peter?' (SVO) *Or:* Which girl does Peter love?' (OVS)
b. Welk meisje belt Frank?
which girl calls Frank
'Which girl calls Frank?' (SVO) *Or:* 'Which girl does Frank call?' (OVS)

Lee's bidirectional model of word order freezing, if applied to these examples, would predict that only the SVO reading is available. I propose we start modelling wh-questions by using the following two constraints. They are ranked as in (59).

(58) WH-LEFT: Wh-constituents are initial (e.g., Legendre et al., 1995).
SUBJECT-LEFT: Subjects are initial.

(59) WH-LEFT $\gg$ SUBJECT-LEFT

We have already seen the constraint SUBJECT-LEFT in action in this chapter. In the context of word order in Dutch, it is worth pointing out that the first half of this dissertation can serve as support for SUBJECT-LEFT. We have seen that there is a global tendency in

the Dutch sentence for subjects to be realized early (Section 2.6.1 and Chapter 4). The ranking in (59) is justified by the fact that wh-fronting is obligatory in Dutch, irrespective of grammatical function.[21]

The grammar in (59) forces subject-questions and object-questions to come out as NP[WH]–V–NP in production. Let's assume there is no morphological information as to which NP is the subject. The underlying semantic difference is therefore not visible in form, which means we have a case of neutralization. However, in comprehension, WH-LEFT does not influence optimization because it is an input-markedness constraint. SUBJECT-LEFT therefore selects SVO as the optimal interpretation. Thus, in comprehension, a wh-question is always interpreted as a subject-question. When production and comprehension are combined, only subject-questions are bidirectionally optimal when there is no morphological disambiguating information. This is clearly at odds with the facts in (57).

The object-question reading is not recovered in comprehension, because in the model word order is the only information available that pertains to grammatical function assignment. According to Zeevat, this is correct: There is no other information contained in an ambiguous question. If this is truly the case, bidirectional OT is in trouble because nothing in comprehension will help recover the interpretation that is less harmonic with respect to word order. However, I think there is information in the questions in (57) that can be used in our model. The Dutch data in (60) show that morphologically ambiguous questions are not simply ambiguous. The preference for SVO or OVS is influenced by the second NP.

(60) a. Welke jongen belt u?
which boy calls you.FORMAL
'Which boy is calling you?' (SVO)
Preferred: 'Which boy are you calling?' (OVS)

b. Welke jongen belt Frank?
which boy calls Frank
'Which boy is calling Frank?' (SVO)
Or: 'Which boy is Frank calling?' (OVS)

c. Welke jongen belt een meisje?
which boy calls a girl
'Which boy is calling a girl?' (SVO)
Less preferred (or even unavailable): 'Which boy is a girl calling?' (OVS)

[21] Like English, Dutch allows for emphatic/pragmatically marked cases like echo questions, utterances of surprise, etcetera, where the wh-constituent is left in situ. A sentence with multiple wh-constituents will also have one (or more) wh-constituents in situ, since only one is fronted. Neither of these constructions are considered in the current analysis.

The sentences are all of the form NP[WH]–V–NP, but the second NP is a (local) pronoun in (60a), a proper name in (60b) and an indefinite NP in (60c). None of the NPs contains disambiguating case or agreement information.[22] The data show that properties of the second NP are important in interpretation. If we move from pronominal to indefinite second NP, the preference for this second NP to be interpreted as the subject goes down. Apparently, the definiteness level of the second NP influences interpretation, through the association between high definiteness and subjecthood. This influence can be exploited to explain the ambiguity of wh-questions in comprehension.

There is psycholinguistic evidence regarding the interpretation of Dutch questions that supports the strategy sketched above. Kaan (1997; 1999; 2001) presents reading time experiments showing that there is a general, but slight subject-first preference in wh-questions (main and embedded) in Dutch. In embedded questions, which show much of the same kind of ambiguity as the main clause questions, but place the finite verb last (NP[WH]–NP–V), the subject-first preference is subdued when the second NP is pronominal, and stronger when it is an indefinite. This is in line with the main clause question intuition data presented in (60). A corpus study revealed that the proportion of (di-)transitive subject-questions, and therefore subject-initial interrogatives, is conditional upon the type of the second NP (Kaan, 1997). If the second NP is pronominal, less than 10% ($\frac{7}{127}$) of the questions is subject initial. If it was a definite NP about 30% ($\frac{33}{105}$) is subject initial, and with an indefinite NP about 40% ($\frac{17}{43}$). Overall, O[WH]VS was more frequent than S[WH]VO. All interrogatives had complex wh-constituents formed with *welke* 'which'. Finally, Kaan compares the subject-first preference in questions to declarative main clauses with a definite initial NP in a reading time experiment. She shows that the subject-first preference is much stronger in the declarative sentences. Because she considers wh-constituents to be indefinite, she concludes that the subject-first preference is also dependent on the definiteness of the initial NP.

The work cited above supports the following two generalizations: First, there is an SVO interpretation preference across declarative and interrogative sentences. Secondly, taking wh-constituents to be indefinite, higher definiteness levels increase the preference for an NP to be interpreted as a subject, irrespective of its position. The first generalization motivates the constraint SUBJECT-LEFT, which applies to interrogative as well as declarative sentences. The second generalization is captured by constraints that link subjecthood to syntactic/semantic properties of the NP. These constraints can be found in the work on differential object marking of Aissen (1999; 2003). The constraints express that certain NPs are better suited as subjects or objects than other NPs. It is generally assumed that a good way to derive and motivate families of such constraints and their universally fixed ranking is through a technique called *harmonic alignment* (Prince and

[22]The formal second person pronoun *u* does not show case or number. Its agreeing morphology on regular verbs is identical with third person singular.

Smolensky, 1993/2004). A (supposedly) universally valid scale is aligned with a binary opposition. If we take the definiteness scale familiar from Chapters 2 and 4 pronoun, definite full NP, and indefinite full NP,[23] and align it with the opposition subject versus object, the result is the two sub-hierarchies of constraints in (61).

(61) *SUBJECT/INDEFINITE ≫ *SUBJECT/DEFINITE ≫ *SUBJECT/PRONOUN
*OBJECT/PRONOUN ≫ *OBJECT/DEFINITE ≫ *OBJECT/INDEFINITE[24]

These ranked constraints say that the worst offence for a subject is to be an indefinite full NP and that it is least bad for a subject to be a pronoun. For objects, the scale is reversed. In Section 4.3, we could see that this reversal of the association between definiteness and grammatical function is nicely reflected in the CGN data. In Table 4.9, p107, I showed the association between grammatical form and definiteness level in terms of so-called *pointwise mutual information* (PMI). Negative PMIs indicate a disfavoured combination, positive PMIs a favoured one. If we focus on the association between subjects and direct objects on the one side and definiteness level on the other, we can see the constraints of (61) and their universal ranking in the PMIs. In the CGN, subject indefinite full NPs have a PMI of -1.42, subject definite full NPs of -0.17, and subject pronouns of +0.16. Direct object pronouns have an PMI of -0.79, direct object definite full NPs of +0.48, and direct object indefinite full NPs of +1.67. In each case, a combination that is forbidden by a certain constraint has a higher PMI than the combination forbidden by a constraint that is higher in the ranking of (61). Also see Zeevat and Jäger (2002) for a discussion on the relation between corpus statistics and harmonic alignment.

The ambiguity in of the wh-questions in (57)/(60b) can now be analyzed as follows. The ambiguous sentences are of the form NP[WH]–V–NP[DEF]. Following Kaan (1997, etc.), we assume that [WH] implies [IND]. Consequently, the constraint *SUBJECT/INDEFINITE prefers an OVS reading of these sentences. The SVO reading is preferred by SUBJECT-LEFT. Therefore, putting these two conflicting constraints in the same stratum brings the variation in comprehension that we are trying to model. That is, *SUBJECT/INDEFINITE should be inserted into the grammar in (59) as shown in (62). The compatible full rankings are given in (62a,b).

(62) WH-LEFT ≫ { SUBJECT-LEFT, *SUBJECT-INDEFINITE }
a. WH-LEFT ≫ SUBJECT-LEFT ≫ *SUBJECT/INDEFINITE
b. WH-LEFT ≫ *SUBJECT/INDEFINITE ≫ SUBJECT-LEFT

[23] The definiteness scale used in the cited papers by Aissen shows more distinctions: pronoun, proper name, definite description, specific indefinite, and non-specific indefinite. This difference in granularity does not matter here.

[24] A constraint $*F/T$ is read as: An NP should not have function F and be of type T at the same time. The type *definite* in the constraint names refers to definite full NPs, not pronouns. The function *object* refers to direct object.

The tableaux in (63) show that object-questions with definite subjects are bidirectionally optimal under full ranking (62b). Recall from Section 5.3 that bidirectional optimality was indicated with a '✌'. The questioned argument is preceded by a '?' in the meaning representation.

(63)

call(frank, ?human)	WH-LEFT	*SUBJECT/IND	SUBJECT-LEFT
☞ wie belt Frank			*
Frank belt wie	*!		

wie belt Frank	WH-LEFT	*SUBJECT/IND	SUBJECT-LEFT
call(?human, frank)		*!	
✌ call(frank, ?human)			*

I will omit tableaux that show that the SVO reading of (60b) is also captured. The combination of the constraints WH-LEFT and SUBJECT-LEFT was already sufficient for that. Subject questions are optimal under the spelled-out ranking (62a). I do wish to focus briefly on another prediction, however. When the second NP in a wh-question is indefinite, the grammar in (62) predicts word order freezing. Such a question is of the form NP[WH]–V–NP[IND], which means that both SVO and OVS violate *SUBJECT/INDEFINITE. In comprehension we therefore select SVO under each spelled out ranking because of SUBJECT-LEFT. The prediction *O[WH]VS[IND] is compatible with the datum in (60c): The object-question reading is strongly dispreferred. Wh-questions with indefinite postverbal NPs really appear to be frozen.

As a result of word order freezing in wh-questions with an indefinite postverbal NP, calls(∃boy, ?human) and comparable meanings are not paired with a form under the current grammar. This is referred to as *ineffability*: a meaning cannot be expressed using a certain construction in the language under consideration. Ineffability is a problem for unidirectional OT. Every input always receives an output, which means that every meaning is always paired with a form in production OT. Examples of (language particular) ineffability are often cited to argue for a bidirectional OT framework (see Fanselow and Féry, 2002 for an extensive overview of ineffability in all domains of grammar).[25] Note that the form-counterpart of ineffability, where a form is not paired to any meaning, is ungrammaticality.

[25]To see that the ineffability of object questions with indefinite subjects is predicted to be language particular under the current set of constraints, consider a grammar in which WH-LEFT is not ranked very high, which could result in wh-constituents in situ. In such a language, object questions with indefinite subjects could be expressed as S[IND]VO[WH].

The analysis presented above addresses the main point raised by Zeevat about the interpretation of questions, but it does raise further questions and issues that need to be investigated in future research. To start with, the assumption that wh-constituents are like any other indefinite, taken from Kaan (1997), needs closer scrutiny. We will probably need to refine our concept of meaning. After all, the current 'meanings' are not much more than syntactic structures written down as if they were semantic constructs, in which syntactic features like IND/DEF and WH are conveniently translated to existential quantifiers, question mark operators, etcetera. One way in which wh-constituents and regular indefinites arguably differ is in their information structural status, a point that will be speculated more upon in Subsection 5.5.4. Special indefinites like specific or generic indefinites also ask for a further refinement of the meaning notion.

Another aspect of questions that I have willfully ignored is the number of different constructions that are available in Dutch to express them. For instance, Dutch has a very strong tendency to use an existential construction (EC) when the subject is indefinite (Section 2.6.2). The EC is characterized by the use of the expletive element *er*, which often, but not always, occurs in the Vorfeld. Subject questions can also be realized as ECs. In this case, the expletive *er* does not appear in the Vorfeld, since the wh-constituent occupies it.

(64) Wie belt er een jongen / Frank
who calls EXPL a boy Frank
'Who is calling a boy/Frank?'
Not: 'Who is a boy/Frank calling?'

Furthermore, there are often ways to demote one of the arguments to an oblique (PP) argument. For instance, the verb *bellen* 'call' can select for a direct object or for a PP headed by *met* 'with', as in (65). Again, this can be combined with an expletive, provided the subject is indefinite. Because of the preposition, these sentences are not ambiguous with respect to argument order.

(65) a. Wie belt (er) met een jongen / Frank
who calls EXPL with a boy Frank
'Who is calling (with) a boy/Frank?'
b. Met wie belt (er) een jongen / (*er) Frank
with whom calls EXPL a boy EXPL Frank
'Who is a boy / Frank calling (with)?'

The subject can also be demoted to an oblique argument by using a passive. This construction is the natural candidate for expressing the hitherto ineffable object question with an indefinite subject. After passivization, this is a subject question with an indefinite oblique argument, like (66).

(66) Wie wordt (er) door een meisje gebeld?
who AUX EXPL by a girl called
'Who is being called by a girl?'

The reason for mentioning all these alternative constructions is that in a competition framework like OT, the competitors are crucial for determining the grammaticality of a form-meaning pair. Adding the constructions just mentioned to the candidate sets in production might break existing results. Before we can allow the choice between all these constructions to be governed by (an extension of) the grammar proposed here, we need to investigate the properties of these constructions further.

These questions are good starting points for future research. I hope to have shown here, however, that the ambiguity of certain wh-questions can be explained by interpretation preferences. The analysis of preferences in the interpretation of wh-questions is supported by psycholinguistic studies. Moreover, the constraints used can be found in the cross-linguistic literature, and reflect clear trends in the CGN.

The analysis presented in this subsection strongly suggests that wh-questions are not 'syntactic Rat/Rad' cases after all. The argument in Rat/Rad was that there is nothing about [ra:t] that helps us recover the voicing of the coda in comprehension. In the wh-questions we have analyzed here, it is the definiteness of the second NP that helps us recover non-canonical word order. The ambiguity of wh-questions can no longer be used as an argument against bidirectional OT. On the contrary, because the interpretation *is* influenced by preferences as to what makes a good subject, it would seem that unidirectional approaches will have trouble integrating this data, just like unidirectional models have trouble integrating information from animacy and plausibility (Section 5.4.3).

5.5.3 Information structure in frozen sentences

Let us turn to the second problematic case for the bidirectional model. The bidirectional model described in Section 5.3 makes the following prediction (repeated from 31, p180).

(67) botal patt^{h}ar todegaa
bottle.NOM stone.NOM break.FUT
break(<u>bottle</u>, stone)
Not: break(stone, <u>bottle</u>) (word order freezing)
Also predicted not: break(bottle, <u>stone</u>) (information structure unambiguous)

The reason for this prediction is that having *botal* as subject and topic in comprehension is a way of satisfying both SUBJECT-LEFT and TOPIC-LEFT. Recall from Section 5.4.1 that Zeevat (2006) claims that the missing SVO reading where the object is topic should be available as well. He writes: "[...it] is precisely the point of freezing [...] that the

word order is determined by the thematic roles, even if the normal trigger (topic) for optional inversion is present" (2006, p1097).

Zeevat's unidirectional model predicts that the availability of an object-topical reading in a canonical word order sentence is contingent upon freezing. The semantic contrast subject-topic versus object-topic is neutralized in word order only when word order is frozen. It is unclear to me what the situation in Hindi is, but what we have seen so far about Japanese and Dutch suggests that this is incorrect. Canonical word order in Dutch and Japanese may always be associated with an object-topic reading, simply because topic fronting is optional (Flack, 2007, for Japanese; for Dutch see below; I am not claiming the same notion/type of topic applies to both). In fact, the behaviour of Zeevat's model is at this point incompatible with the quote above, since the quote seems to indicate that Zeevat considers subject-object inversion to be optional, too.

Since the bidirectional model of Section 5.3 cannot deal with ambiguity or optionality, it cannot deal with the ambiguity of canonical word order and the related optionality of topic fronting in Japanese and Dutch. Even if we use stratified bidirectional OT, and put the constraints TOPIC-LEFT and SUBJECT-LEFT in one stratum to give us optionality of topic fronting, canonical word order does not become ambiguous in a bidirectional model. The constraints TOPIC-LEFT and SUBJECT-LEFT cannot recover an object-topic reading from canonical word order under any ranking.[26]

However, the inability of the bidirectional model of word order to predict the object-topic reading of a canonical word order sentence is only an argument against bidirectional OT if we have exhausted all the means of identifying topics in the proposed model. It would be an argument against a bidirectional model if word order were the only way to identify topics. But this is highly unlikely. Consider the case of contrastive topics in Dutch. Section 2.5.2 shows that contrastive topics may be topicalized. However, contrastive topics are also accompanied by prominent accenting, and they occur in contexts that contain a contrastive element. The two latter properties are independent of topicalization. Let me illustrate this with an example. The context is set up such that in the answer *hen* 'them', which refers to the Zorgeloos family, is naturally interpreted as the contrastive topic.

(68) A Ik weet dat Frank de familie Niemandsverdriet heeft gebeld, maar heeft hij de familie Zorgeloos ook gebeld?
'I know that Frank called the Niemandsverdriet family, but has he called the Zorgeloos family, too?'

B Nee, hij heeft **hen** niet gebeld.
No he has them not called
'No, he did not call them.'

B′ Nee, **hen** heeft hij niet gebeld.

[26]In fact, in a bidirectional model, the grammars {TOPIC-LEFT,SUBJECT-LEFT} and TOPIC-LEFT ≫ SUBJECT-LEFT predict nearly the same languages: Topics are always initial.

I have not given an indication of prosody in (68B) and (68B′), but a natural pronunciation for the answers is (69). The prominent rise on *hen* in both word order variants confirms the status of *hen* as the contrastive topic in both cases. Slashes indicate rises/falls, capitals mark nuclear accent.

(69) a. hij heeft /hen/ \NIET\ gebeld.
b. /hen/ heeft hij \NIET\ gebeld.

The example in (68) shows that contrastive topic fronting for Dutch is optional, and that therefore, at least some canonical word order sentences receive an object-topic reading. However, this does not mean that all canonical word order sentences have the object-topic reading. It only shows that if there are cues in the context that this meaning is the intended one, the object-topic meaning is available in a canonical word order sentence. Of course, the information coming from the context about which argument is the contrastive topic can be used in comprehension in a bidirectional model.

Before I show how putting information from the context into the system helps us to model the object-topic reading of the canonical sentence (68B), let me discuss how this applies to Zeevat's example. Zeevat claimed that a frozen sentence should have an object-topic reading. What I will attempt to capture in the bidirectional model is that any canonical sentence (frozen or not) in the right context may receive an object-topic reading. One may be worried that I am not addressing Zeevat's objections. However, the point is that I do not know how to understand the claim that a sentence 'has a certain information structural reading' in any other way than 'can be felicitously used in a context that triggers this reading.' The information structural concepts like focus, contrastive topic, etcetera, are so strongly rooted in context that I find it impossible to judge of a sentence in isolation whether it has a certain information structural reading without trying to embed sentences in a context. If Zeevat's claim indeed is that it should be possible to use a frozen sentence in a context that sets up for object topicality, then the analysis that I will describe below addresses most of his criticism. What the bidirectional model will not be able to capture is the prediction of Zeevat's unidirectional model that frozen sentences are unique in this respect and that normal canonical word order sentences do not allow for an object-topic reading, even in an object-topic context. In fact, the bidirectional model will make the strong typological prediction that word order freezing only occurs in languages in which information structurally induced word order is optional, but not obligatory. I have demonstrated above that for Dutch this is a correct prediction, and Flack claims that this is true for Japanese, too. Whether it is the case for Hindi I do not know.

The grammar we have developed thus far is repeated here from (62) as (70).

(70) WH-LEFT $\gg$ { SUBJECT-LEFT, *SUBJECT-INDEFINITE }

We need to add two constraints to this grammar. First, we will need a constraint that causes topicalization. Secondly, we will need a constraint that constrains interpretation depending on the context. In Section 2.5, I discussed the information structural properties of topicalized constituents in Dutch. The conclusion of that discussion was that it is hard to identify a clear common property to all Vorfeld objects, but that they all could be understood as important material in the sense of Givón (1988) and Gundel (1988). Material that is contrastive, new, or unexpected is important material. In the first half of this dissertation, I have argued that the Vorfeld in general prefers to be occupied by important material. The corpus study of Chapter 4 showed that personal pronouns in any function tended to avoid the Vorfeld. This supports the claim that the Vorfeld is a position for important material, under the very reasonable assumption that personal pronouns typically realize unimportant – non-contrastive, highly predictable – material.

Givón (1988) and Gundel (1988) propose a cross-linguistic principle that important material should be fronted. Gundel calls this the *first-things-first* principle. I propose that we use this principle as a constraint to trigger topicalization.

(71) FIRST-THINGS-FIRST: Provide the most important information first.

A similar constraint is found in Legendre (2001), under the name ALIGNNOTEWORTHY. I will make the simplifying assumption that exactly one of the arguments is marked important information in a form-meaning pair. This is indicated by underlining the important argument in the meaning representation. The constraint FIRST-THINGS-FIRST is ranked in the same stratum as SUBJECT-LEFT, to model the fact that topicalizing important material such as a contrastive topics is optional.[27]

(72) WH-LEFT $\gg$ { SUBJECT-LEFT, *SUBJECT-INDEFINITE, FIRST-THINGS-FIRST }

The second constraint to be added to the grammar allows information from the context to enter optimization. I propose that bidirectional optimization as a whole occurs against the backdrop of a given context. All constraints can in principle have access to information, in any optimization direction. A representation of the context in OT is also part of the work of Lee (2001a, and other places),[28] Kuhn (2003) and Beaver (2004). I will assume that the

[27] As an aside, the fact that SUBJECT-LEFT and FIRST-THINGS-FIRST appear in one stratum also explains why reduced pronominal subjects are allowed in the Vorfeld, and reduced pronominal objects are not (Section 2.3). As a (grammaticalized) form for unimportant material, the only way a reduced personal pronoun can appear in the Vorfeld is when it is put there by SUBJECT-LEFT.

[28] Lee proposes that context is part of the input to comprehension. This does appear to be the direction of optimization in which the importance of context is clearest. However, I have tried to emphasize the symmetric and declarative nature of bidirectional optimization in this chapter. The way I have calculated bidirectional optimality by first calculating one optimization direction and then feeding its output to the other direction is just an operationalization of the definition of strong bidirectionality. Adding something to the input of comprehension that is not of the same type as the output of production (or vice versa) is conceptually at odds with the symmetric view, and has too much of a procedural flavour to it. In terms of predictions, it is hard to see differences between the two ways of using context at this point.

context representation is a statement of which referent is important. This is a considerable simplification. The context is crucial in determining what is important and what not, by providing contrastive elements, by determining what is given, or by making reference to certain elements expected. However, importance is ultimately a property of material in an utterance. We cannot establish which material is important in a context until we actually know which material is going to be mentioned in the utterance. Nevertheless, a context like the one in (68) creates the strong expectation that the Zorgeloos family is going to be the contrastive topic, and hence important material, in the answer. A more detailed analysis of the two-way relation between the context and an utterance in determining what is important material will have to await further study. The information about what is important in the context enters optimization through a constraint.

(73) COHERE: The context determines what is important.
Violated when the referent given as important in the context, is not important in a form-meaning pair.

The double representation of information structural importance (in meaning and in context), and the two ways importance can enter optimization (FIRST-THINGS-FIRST and COHERE) represent the dual nature of importance (Givón, 1988). On the one hand importance is heavily influenced by context, on the other hand we are free to decide to present material as important or unimportant, even if it leads to incoherence. I order COHERE in the same stratum as FIRST-THINGS-FIRST. The consequences of this choice will be discussed below.

(74) WH-LEFT $\gg$ { SUBJECT-LEFT, *SUBJECT-IND, 1ST-THINGS-1ST, COHERE }

The stratified ranking in (74) specifies 4! = 24 possible full rankings. The tableaux below – comprehension first, then production – together show that the interpretation of the canonical word order (68B) in which the direct object is contrastive topic, is available when COHERE and SUBJECT-LEFT both outrank FIRST-THINGS-FIRST. In the example, the pronouns bear case, so that only information structure varies in comprehension.

(75) Context: zorgeloos

hij$_{frank}$ heeft hen$_{zorgeloos}$ niet gebeld	WH-L	COH	SU-L	1ST1ST	*SU/I
¬call(frank, zorgeloos)		*!			
☞ ¬call(frank, zorgeloos)				*	

call(frank, zorgeloos)	WH-L	COH	SU-L	1ST1ST	*SU/I
✌ hij$_{frank}$ heeft hen$_{zorgeloos}$ niet gebeld				*	
hen$_{zorgeloos}$ heeft hij$_{frank}$ niet gebeld			*!		

The tableaux show that by making comprehension sensitive to contextual factors, canonical word order can receive an object-topic interpretation in a bidirectional model.

I have put FIRST-THINGS-FIRST in one stratum with COHERE. This predicts that we should be able to use word order to force an incoherent reading of a sentence (1ST1ST ≫ COH). This prediction seems to be correct if we consider cases like (76).

(76) A I heard that Ella called the Zorgeloos-family, but did Frank call them, too?
B Nee, hen heeft hij niet gebeld
no, them has he not called
'No, he did not call THEM.' (but he might have called someone else)

This has to receive the incoherent reading where *hen* is contrastive topic, even though the context sets up for *hij* 'him' to be the contrastive topic. Sentence (76B) receives a rise on *hen*, *hij* is not accented, and *niet* bears nuclear accent. As a result of the unexpected information structure of the answer, the proposition that Frank did call someone else besides the Zorgeloos family becomes salient. See Büring (2003) for a discussion of more subtle examples of incoherence. The tableaux in (77) show that the grammar of (74) can model this case.

(77) Context: frank

$\text{hen}_{\text{zorgeloos}}$ heeft $\text{hij}_{\text{frank}}$ niet gebeld	WH-L	1ST1ST	SU-L	COH	*SU/I
¬call(frank, zorgeloos)		*!	*		
☞ ¬call(frank, zorgeloos)			*	*	

call(frank, zorgeloos)	WH-L	1ST1ST	SU-L	COH	*SU/I
$\text{hij}_{\text{frank}}$ heeft $\text{hen}_{\text{zorgeloos}}$ niet gebeld		*!		*	
✌ $\text{hen}_{\text{zorgeloos}}$ heeft $\text{hij}_{\text{frank}}$ niet gebeld			*	*	

The incoherence of the optimal interpretation in this context is reflected by the fact that COHERE is violated by the winning form-meaning pair.

The coherent reading of (76B), in which *hij* is the contrastive topic and receives a prominent rise in intonation, and *hen* is deaccented, is unavailable. This is correctly predicted by the bidirectional model, too. A high-ranked COHERE would select the coherent reading (frank) as the comprehension-optimal candidate. But in production, no ranking of the constraints will produce an object-initial sentence when the input marks the subject as important. Therefore, there is no bidirectional optimal pair.

All in all, we see that, on the assumption that fronting contrastive topics is optional, and that we can use information from the context in comprehension, stratified bidirectional

OT can capture the information structural ambiguity of a canonical word order sentence. This result carries over to frozen sentences because, for the model, they are canonical sentences like any other. The possibility of information structural ambiguity of canonical sentences does, however, depend on the optionality of topicalization. If we were to fix the ranking 1ST1ST ≫ SU-LEFT, we would end up with predictions similar to those made by the simple bidirectional model of Section 5.3. We would predict that sentences that do not have enough disambiguating information to support non-canonical word order will never express object topicality. In this respect the stratified strong bidirectional model differs from the unidirectional production models of Zeevat and Flack.

In the analysis and the formulation of the constraints that I have used in this section, I have made some simplifying assumptions. Investigating these assumptions is a topic for further work. For instance, I have assumed that there is a unifying principle that explains contrastive topic fronting, non-contrastive topic fronting and focus fronting: the first-things-first principle, which prefers important material to be stated first. Importance is a vague concept and it should therefore be investigated more. But even apart from that, we should scrutinize the assumption that these three types of material behave alike when it comes to fronting. For instance, in Dutch, focus fronting is possible, but relatively rare (Section 2.5.1), whereas non-contrastive topic fronting of demonstrative pronouns is frequent. This difference is not captured by the constraint FIRST-THINGS-FIRST, which treats all important material on a par. Another important issue that needs to be investigated is the representation of contextual information, and the way contextual information enters optimization. In my analysis, I made the assumption that the context directly specifies what is important material. This has allowed me to show that once we have information like that in the system, bidirectional OT can model the mismatch between word order and information structure that Zeevat has pointed out. A deeper investigation of the relation between importance in an utterance and the context will probably involve a deconstruction of importance into the different types of material that researchers like Givon and Gundel consider to be important: new material, contrastive material, and unexpected material. This investigation is beyond the scope of the current work.

Overview of the model predictions and the effect of coherence

At this point the grammar has five constraints in two strata (78), repeated from (74).

(78) WH-LEFT ≫ { SUBJECT-LEFT, *SUBJECT-IND, 1ST-THINGS-1ST, COHERE }

There are 24 full orderings of the constraints. To gain more insight into the predictions of the grammar, we can calculate which form-meaning pairs in a hypothetical linguistic universe are grammatical under (78). Consider a finite universe of form-meaning pairs, Gen. The possible meanings are formed by taking *call* (Dutch: 'bellen') as predicate, one masculine and one feminine argument, and one of four semantic operators for each ar-

gument: existential/indefinite $\exists$, definite ι, pronoun *he/him/she/her* and question *?*. The possible forms are always NP–*belt*–NP 'NP–calls–NP' sentences, where the NPs are *hij/hem/zij/haar* 'he/him/she/her', *een/de/welke jongen* 'a/the/which boy', or *een/het/welk meisje* 'a/the/which girl'. The set of possible form-meaning pairs Gen is the set of those form-meaning pairs that adhere to the correct mapping of semantic argument to case and gender, and of semantic operator to NP-type.

In Figure 5.3, p217, I have given a relevant subset of this universe, and indicated which form-meaning pairs are grammatical under (78) by connecting the elements with a line. Forms are given in the left column, meanings in the right. Any point in the figure can have more than one line connecting to it. A form that has more than one connection is ambiguous, a meaning that has more than one connection can be optionally realized by several forms. Some points are not connected to anything at all. These points represent ungrammatical forms or ineffable meanings. They are drawn with an open circle '$\circ$'. The numbers written in the lines are the number of rankings that predict the connected pair to be grammatical. Some grammatical pairs illustrate the same prediction. For space reasons I have omitted some of the superfluous pairs.

The predictions fall into three groups. In Subfigure (a), the predictions for declarative sentences are given in a context that sets up the feminine argument as important. In this context, COHERE is satisfied by meanings in which the feminine argument is underlined. The effect of the COHERE constraint can be observed by comparing Subfigure (a) to the predictions in a *null*-context, given in Subfigure (b). Subfigure (b) shows the predictions of the bidirectional model criticized by Zeevat: Importance of the object is ineffable when we have two arguments that lack case marking and are equal in definiteness. Subfigure (a) shows that COHERE recovers the desired information structural ambiguity of SVO, provided that the feminine argument is the intended important argument.

There are two interesting cases worth presenting in a bit more detail. First, the following ambiguity is predicted in the context of an important feminine argument as well as in the null-context. The sentence can be found as the fourth form from the top in both Subfigures (a) and (b).[29]

(79) Een jongen belt het meisje.
a boy calls the girl
call($\underline{\exists\text{boy}}$, ιgirl) *Predicted also*: call(ιgirl, $\underline{\exists\text{boy}}$)

The second reading – OVS – is due to *SU/INDEF, the same reason a question with a definite second NP receives an OVS reading. On the basis of introspection, this prediction does not strike me as obviously correct, that is, this sentence should freeze to SVO. The reason for the incorrect prediction may be that the model treats all important material on a par and is not sensitive to other determinants in Vorfeld occupation. Earlier in this

[29] In Subfigure (a) the sentence receives three readings because of COHERE.

dissertation, I have already established that there is a global word order trend in Dutch that discourages early realization of indefinite full NPs. In Section 4.3, for instance, we could see that indefinite full NP objects in the Vorfeld are rare. If this knowledge about other influences on word order in Dutch were to be incorporated in the OT model, it might be able to predict that (79) is not very a very good OVS sentence. Incorporating the results on the influence of definiteness on word order into a bidirectional OT model is a topic for future research. Note, however, that the explanation for the lacking OVS reading of (79) that I just sketched is a production explanation: An indefinite direct object does not have a very good chance of being put into the Vorfeld to begin with. This would also explain the difference between (79) and a wh-question. In production, wh-constituents have to be put into the Vorfeld because of WH-LEFT. What the incorrectness of the prediction in (79) does not necessarily show, however, is that the constraints on grammatical function and definiteness are wrong. In the next chapter, we will see that the relative definiteness of subject and object has an effect on non-canonical word order and that this effect is as predicted by a statistical interpretation of the bidirectional model.

A second prediction in Subfigure (a) worth looking at is the following. As we have seen above, an OVS sentence in a context where the subject is important is always predicted to be incoherent. In Subfigure (a) this can be seen in the sentences *de jongen belt zij* 'the boy she calls' and *hem belt het meisje* 'him the girl calls' – six and two sentences from the bottom, respectively. These sentences are only connected to meanings in which the masculine argument is important. However, the model also predicts that SVO in a context where the object is important is ambiguous between object-important (coherent) and subject-important (incoherent). This ambiguity is for instance predicted for *hij belt het meisje* 'he calls the girl', the last sentence in Subfigure (a). However, even though SVO word order is compatible with an subject-important reading, it seems to me that the reading is not available in an context where the object is important, as long as we are not forced to it by intonation. To illustrate, let us look again at part of example (68), here repeated as (80).

(80) A I heard Frank called the Niemandsverdriet family, but did he call the Zorgeloos family, too?
B Nee, hij heeft hen niet gebeld.
No, he has them not called
'No, he has not called them.'

The context in (80A) is such that the in situ object *hen* in (80B) is the contrastive topic. The reading where the initial *hij* is the contrastive topic does not surface naturally. The predicted ambiguity in what is important material in (80) does not exist. Apparently, canonical word order does not force an incoherent reading in which the important material is realized in first position like non-canonical word order does. At this point, it is not clear

Figure 5.3: Selected predictions of the grammar in (78), p213. Subfigure (a): declarative, the feminine argument is important in the context. Subfigure (b): declarative, no important material in context. Subfigure (c): interrogative. ▶

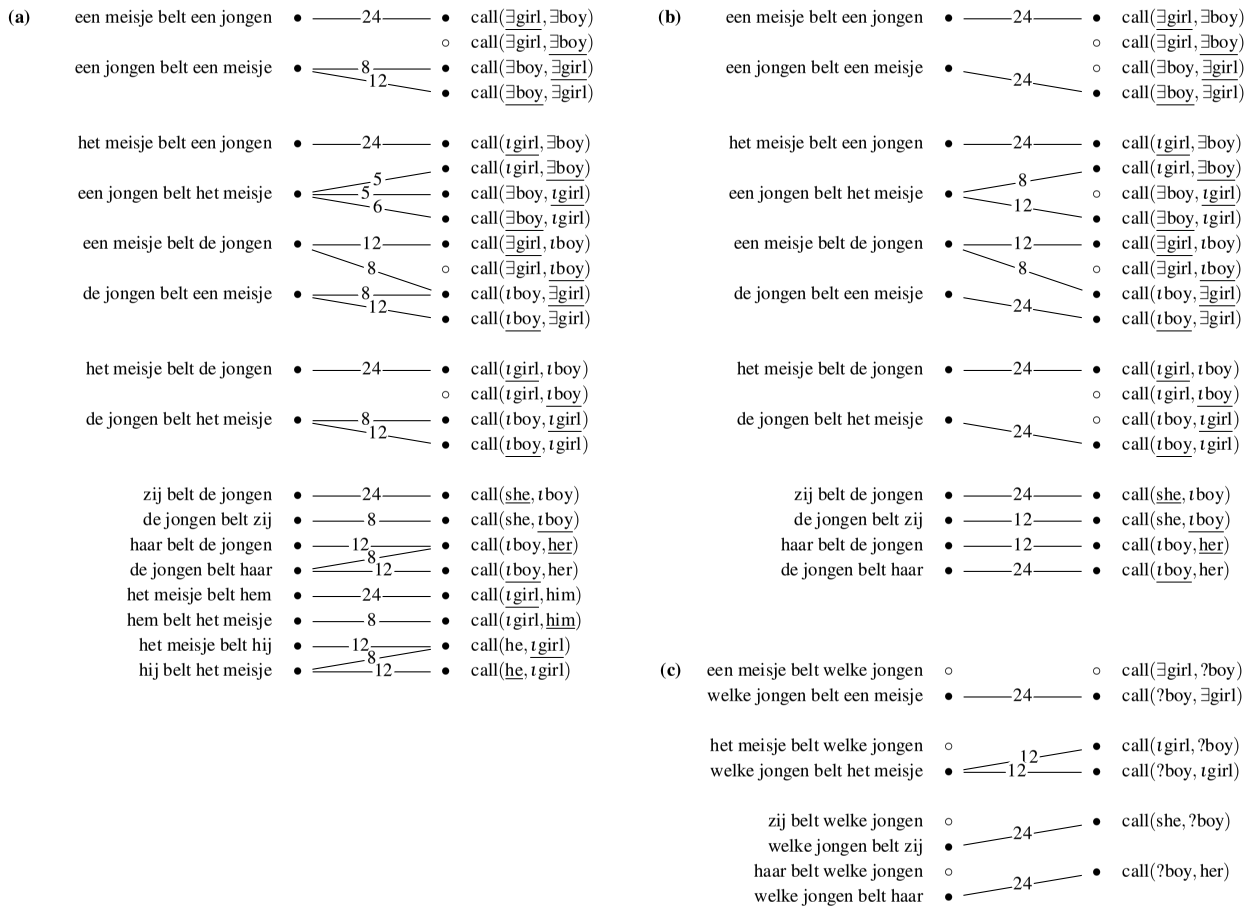
whether the solution is just a matter of understanding information structure better and formulating other constraints, or whether the falsely predicted ambiguity shows a deeper problem with the bidirectional model. To see why the latter might be the case, consider the following. The combined facts that one cannot force a coherent reading onto a non-canonical word order and that one cannot force an incoherent reading onto a canonical word order result in a ranking paradox in comprehension. On the one hand, the behaviour of non-canonical word order means that FIRST-THINGS-FIRST is at least allowed to outrank COHERE. On the other hand, the behaviour of canonical word order suggests that FIRST-THINGS-FIRST should never COHERE. Investigating this paradox in depth is beyond the scope of this dissertation, but in Section 5.6, I will discuss a formalization of bidirectional OT that can deal with this paradox. However, I would like to point out that although the ranking paradox just sketched is particular to a model that includes the comprehension perspective, the false prediction of the data in (80) is not. A unidirectional production model would also wrongly predict (80) to be ambiguous. Canonical word order *should* be production optimal when the subject is important. Therefore a production model predicts that the pairing of SVO and subject-important is grammatical. Since a production model has no way of factoring in context at all, the subject-important reading is predicted to be an available reading in any context, including the one in (80).

Under Subfigure (c), we have the predictions for wh-questions. The prediction that stands out here is the ineffability of an object question with an indefinite subject. I have discussed this case in Section 5.5.2. The predicted ineffability is in my opinion correct. A way to ask such a sentence would be to use a passive.

Overall, the graph shows that the model predicts more ambiguity than optionality. This is due to the fact that there is only one pair of alternative rankings that affects form in production: FIRST-THINGS-FIRST $\gg$ SUBJECT-LEFT vs SUBJECT-LEFT$\gg$FIRST-THINGS-FIRST. In contrast, meaning is possibly influenced by all 24 alternative rankings.

5.5.4 Focus fronting

The last of the Zeevat/Flack problems is the case of focus scrambling in Japanese. Recall that even when topic scrambling is ruled out because of freezing, focus scrambling was still allowed. We therefore have the contrast exemplified in (81a) and (81b).


(a)
een meisje belt een jongen
24
call(∃girl, ∃boy)
call(∃girl, ∃boy)
een jongen belt een meisje
8
12
call(∃boy, ∃girl)
call(∃boy, ∃girl)
het meisje belt een jongen
24
call(ιgirl, ∃boy)
call(ιgirl, ∃boy)
een jongen belt het meisje
5
5
6
call(∃boy, ιgirl)
call(∃boy, ιgirl)
een meisje belt de jongen
12
8
call(∃girl, ιboy)
call(∃girl, ιboy)
de jongen belt een meisje
8
12
call(ιboy, ∃girl)
call(ιboy, ∃girl)
het meisje belt de jongen
24
call(ιgirl, ιboy)
call(ιgirl, ιboy)
de jongen belt het meisje
8
12
call(ιboy, ιgirl)
call(ιboy, ιgirl)
zij belt de jongen
24
call(she, ιboy)
de jongen belt zij
8
call(she, ιboy)
haar belt de jongen
12
8
call(ιboy, her)
de jongen belt haar
12
call(ιboy, her)
het meisje belt hem
24
call(ιgirl, him)
hem belt het meisje
8
call(ιgirl, him)
het meisje belt hij
12
8
call(he, ιgirl)
hij belt het meisje
12
call(he, ιgirl)
(b)
een meisje belt een jongen
24
call(∃girl, ∃boy)
call(∃girl, ∃boy)
een jongen belt een meisje
24
call(∃boy, ∃girl)
call(∃boy, ∃girl)
het meisje belt een jongen
24
call(ιgirl, ∃boy)
call(ιgirl, ∃boy)
een jongen belt het meisje
8
12
call(∃boy, ιgirl)
call(∃boy, ιgirl)
een meisje belt de jongen
12
8
call(∃girl, ιboy)
call(∃girl, ιboy)
de jongen belt een meisje
24
call(ιboy, ∃girl)
call(ιboy, ∃girl)
het meisje belt de jongen
24
call(ιgirl, ιboy)
call(ιgirl, ιboy)
de jongen belt het meisje
24
call(ιboy, ιgirl)
call(ιboy, ιgirl)
zij belt de jongen
24
call(she, ιboy)
de jongen belt zij
12
call(she, ιboy)
haar belt de jongen
12
call(ιboy, her)
de jongen belt haar
24
call(ιboy, her)
(c)
een meisje belt welke jongen
call(∃girl, ?boy)
welke jongen belt een meisje
24
call(?boy, ∃girl)
het meisje belt welke jongen
12
call(ιgirl, ?boy)
welke jongen belt het meisje
12
call(?boy, ιgirl)
zij belt welke jongen
24
call(she, ?boy)
welke jongen belt zij
haar belt welke jongen
24
call(?boy, her)
welke jongen belt haar

(81) a. Taroo-ga Hanako-ga kowa-i
Taroo.NOM Hanako.NOM is afraid of
'Taroo is afraid of Hanako.' (SOV)
Not: 'Hanako is afraid of Taroo.' (OSV)

b. TAROO-GA Hanako-ga kowa-i
Taroo.NOM Hanako.NOM is afraid of
'TAROO is afraid of Hanako.' (SOV, subject focus)
Or: 'Hanako is afraid of TAROO.' (OSV, object focus)

This aligns well with Bloom's (1999) observations that so-called emotive sentences in Russian, with nuclear stress in a non-final position, are exceptions to freezing. It is also compatible with (somewhat anecdotal) evidence from Dutch: When asking lay informants about the interpretation of NP–V–NP, they are more willing to accept an OVS reading for a sentence that has nuclear stress on the preverbal NP. It also happens that informants produce this intonation without prompting when the possibility of OVS is pointed out. An example is in (82).

(82) \ELLA\ belt Frank
Ella sees Frank
'ELLA calls Frank.' *Or:* 'ELLA, Frank calls.'

Nuclear stress on the first NP indicates that this NP is in focus, and the rest of the sentence is in the background (Section 2.5.1). The interpretation of (82) contrasts with (83), which does not show an SVO/OVS ambiguity.

(83) /Ella/ belt \FRANK\
Ella calls Frank
'Ella calls FRANK.' *Dispreferred:* 'Ella, FRANK calls.'

The intonation pattern in (83), a marked rise on the Vorfeld constituent and nuclear accent on the postverbal constituent, indicates that the initial element in (83) is a contrastive topic. In principle, (83) should be compatible with a contrastive topic topicalization (OVS), but we see that this reading is not readily available. What could prompt the difference between (82) and (83)?

I think the difference between the two sentences lies in the information structure that is indicated by the intonation. In (82), the first NP receives nuclear stress, and is thus focussed. In (83), the NP carries a contrastive topic accent, and can therefore also be considered to be focussed (see Section 2.5.2). In (82), the second NP is deaccented, and the material in the NP therefore belongs to the information structural background. In (83), the second NP receives nuclear stress and is focussed. The difference between OVS in (82), and OVS in (83) is that the former does not focus the subject.

In the OT literature, one finds several instances of constraints that prefer subjects not to be focussed, but to be part of the information structural background. Note that such a constraint would prefer OVS in (82) because this interpretation puts the object in focus and the subject in the background.

(84) BACKGROUND(SUBJECT): The subject is part of the information structural background of an utterance.

The constraint BACKGROUND(SUBJECT) resembles constraints found in the work of Lee (2001b), and in Beaver (2004). Zerbian (2007) uses a constraint like BACKGROUND(SUBJECT) in an account of Northern Sotho (Bantu). In Northern Sotho, focussed grammatical subjects are banned altogether, and the expression of the corresponding meaning requires a special construction, such as a cleft construction or a (marked) postverbal realization of the subject. Apparently, Northern Sotho shows an absolute, grammatical version of the interpretation tendencies observed in Dutch and Japanese.

BACKGROUND(SUBJECT) causes OVS in (82). This constraint must therefore be allowed to outrank SUBJECT-LEFT. This ensures that comprehension selects the OVS reading. In production, FIRST-THINGS-FIRST favours focus topicalization. However, the sentence in (82) is ambiguous and can also receive an SVO reading. This ambiguity is captured by putting BACKGROUND(SUBJECT) in the same stratum as SUBJECT-LEFT.

What remains is the sentence in (83). The lack of ambiguity in this sentence is predicted. Since neither contrastive topic nor focus is part of the background, there is no way BACKGROUND(SUBJECT) can be satisfied. SUBJECT-LEFT thus always fixes the interpretation. The result is the observed word order freezing.

Under the assumption that the Japanese topics are background material, and foci are not, BACKGROUND(SUBJECT) explains the ambiguity of focus-initial sentences in double nominative sentences in Japanese, too. Japanese shows a rather extreme freezing pattern: Word order freezing is strong enough to block plausible meanings in favour of absurd ones, but the dislike for focussed subjects readily overrides freezing.

The constraint BACKGROUND(SUBJECT) suggests further research in several directions. By treating both contrastive topics and foci as non-background material, we predict that the sentence in (85) displays the same ambiguity as the sentence in (82).

(85) /Frank/ heeft Ella \geBELD\
Frank has Ella called
'Frank has called Ella.' (SVOV)
Also, predicted by BACKGROUND(SUB)*:* 'Ella has called Frank.' (OVSV)

I cannot decide on the basis of intuition alone whether the sentence in (85) is like (82) (not frozen) or like (83) (frozen). Therefore, I cannot say whether the prediction

that BACKGROUND(SUBJECT) makes is actually correct. An experimental study that investigates the acceptability of sentences that are biased towards OVSV by world knowledge but differ in whether the subject or the verb is focussed might help to answer this question.

A second direction for further investigation of BACKGROUND(SUBJECT) concerns the consequences of adopting a constraint like BACKGROUND(SUBJECT) for wh-questions. The issue of information structure *within* questions is an interesting and as far as I am aware little researched one. Nevertheless, it seems clear that wh-constituents are not, as a complete constituent, part of the background of a sentence. Thus, the constraint BACKGROUND(SUBJECT) also prefers an object-initial interpretation of wh-questions, since it regards the wh-constituent to be unfit as a subject. The subject-definiteness constraints I have employed in Section 5.5.2 and BACKGROUND(SUBJECT) overlap in predictions. Interestingly, Northern Sotho shows a syntactic reflex of the violation of BACKGROUND(SUBJECT) in the case of questions, too. Like focussed subjects, questioned subjects cannot be realized in their unmarked position, but require realization in a cleft construction or in the marked, postverbal position. Other constituents can be questioned in situ (Zerbian, 2007, and references therein). The similarities between a non-backgrounded subject and a questioned subject in a language like Northern Sotho suggest that further study of the relation between the backgrounding and questioning in general may be worthwhile topic for future research.

5.5.5 Animacy, world knowledge and gender agreement

At this point, we have addressed the concerns of Zeevat/Flack by showing that a bidirectional model of freezing can deal with the different cases of word order related ambiguity and optionality pointed out in the respective papers. In this section, I will demonstrate that a bidirectional account can naturally deal with the data that I argued to be problematic for the unidirectional accounts in Section 5.4.3. I will start with the cases in which freezing is prevented by non-form information like animacy. The treatment of these cases also allows me to demonstrate that stratified bidirectional OT is like classic unidirectional OT in that differences between languages can be modelled by differences in constraint ranking. In the second part of this section, I will take on the (tentative) Russian data in which form information, in the form of verbal gender agreement, did not prevent freezing.

Cross-linguistic variation in the role of animacy

Recall that the influence of animacy differed between languages: Russian and Japanese, so it is claimed, show word order freezing even if there is animacy information present to tell us otherwise. The Japanese example is repeated in (86).

(86) Jishin-ga Taroo-ga kowa-i
earthquake.NOM Taroo.NOM is afraid of
'Earthquakes are afraid of Taroo.'
Not: 'Taroo is afraid of earthquakes.'

Being afraid selects a sentient and therefore human or animate subject, so there is enough information to identify *Taroo* as subject of the clause in (87). Still, word order is frozen, presumably because there is no distinct form marking of the subject.

In Dutch, Hindi, and Swedish, such sentences are ambiguous between object-initial and subject-initial, to say the least. They may even only allow an OVS reading. We can model the information used in Dutch, Hindi and Swedish, but ignored in Japanese and Russian, by further constraints on subjects. Like *SUBJECT/INDEFINITE, the constraint in (87) can be derived from Harmonic Alignment.

(87) *SUBJECT/INANIMATE: Do not have subjects with inanimate referents.
(Aissen, 1999)

This constraint is violated by the SOV reading of (87). The missing OSV reading would satisfy it.

A language like Japanese puts this constraint below the word order constraint SUBJECT-LEFT, meaning that word order is a stronger force than animacy in determining what is the subject. As a result, $\langle$*jishinga Tarooga kowai*, afraid(taroo, $\underline{\text{earthquake}}$)$\rangle$ is not a grammatical form-meaning pair. A language like Hindi, however, would tie *SUBJECT/INANIMATE with SUBJECT-LEFT or even put it in a higher stratum, indicating that animacy can be (or always is) used as a factor in deciding what is subject.[30]

The attractiveness of this kind of analysis within OT lies in the fact that a cross-linguistic difference is completely explained by appealing to differences in constraint ranking, which puts us in the classic OT tradition with a framework that is decidedly not classic, as it features both bidirectional optimization and Anttila-style ranking.

An example that we cannot deal with in this way, and which I will not be able to deal with either, is one where only plausibility of reading played a role. For Dutch we have seen such an example, repeated in (88).

(88) Zo'n klein jongetje slaat alleen een heel grote bruut.
Such a small boy hits only a very big brute
'Such a small boy would only hit a very big brute.'
Or: 'Only a very big brute would hit such a small boy.'

[30]If one wishes to give this analysis in terms of selection restrictions instead of animacy, this presents no problem. However, animacy has the advantage of being general. It is also fairly straightforward to annotate for it, so that it can be employed in a corpus investigation (Chapter 6).

Short of positing a high ranked constraint that an interpretation should make sense, or be the most plausible, this example cannot be handled correctly by the model proposed here. Unless we have a sensible way of expressing this plausibility – perhaps in terms of probability or even predicate-argument frequencies for the computationally inclined – I suggest we hold off on such a constraint, because it would be too much of a *deus ex machina*. I would like to note, however, that when we do find a more precise way of modelling plausibility, and moreover, if we do so in the form of a constraint, the direct effect of plausibility will only be noticeable in comprehension optimization. Plausibility here refers to the plausibility of meanings (situations, assertions). A production optimization approach would not be able to use this information in optimization, because meaning is fixed in the input.

Gender agreement in Russian

The data in which Russian ignores form information (gender agreement) poses no problem for a bidirectional model, either (Section 5.4.3). The Russian data is as in (89). The ungrammaticality of (89b) is to be explained by the combination of freezing and the resulting incorrect gender agreement on the verb. The correct agreement would be as in (89c).

(89) a. Ditja videlo myš'.
child.NEUT.NOM/ACC sees.PAST.NEUT mouse.FEM.NOM/ACC
'The child sees the mouse.'
b. *Myš' videlo ditja.
c. Myš' videla ditja.

Bloom argued that the ungrammaticality of (89b) was due to the fact that freezing occurs in spite of the gender agreement on the verb. If word order is indeed frozen, *myš' is* the subject in (90b), but *videlo* does not agree with it.

Bloom's explanation can be phrased almost directly in terms of a bidirectional model. The solution lies in the relative ranking of the constraints on agreement and SUBJECT-LEFT. Since we have no analysis of Russian past tense gender agreement at hand, suppose it is captured by a constraint GENDERAGREEMENT.

The tableaux in (90), with only the relevant constraints, show that ranking GENDERAGREEMENT below SUBJECT-LEFT predicts the data in (89). The crux is that under this ranking, GENDERAGREEMENT does not influence comprehension. SUBJECT-LEFT makes sure that the leftmost NP is the subject in the optimal candidate. However GENDERAGREEMENT still enforces the correct agreement in production. The comprehension optimal meaning for (89b) leads to (89c) in production. Therefore, we do not have a bidirectionally optimal pair.

(90)

myš' videlo ditja.	1ST1ST	SUB-LEFT	GENAGR
see(child, mouse)	*!	*	
see(child, mouse)		*!	
☞ see(mouse, child)			*
see(mouse, child)	*!		*

see(mouse, child)	1ST1ST	SUB-LEFT	GENAGR
ditja videlo myš'	*!	*	*
ditja videla myš'	*!	*	
☠ myš' videlo ditja			*!
☞ myš' videla ditja			

The treatment of the Russian example shows an interesting property of the bidirectional model when compared to a unidirectional model. The logic of ranking constraints in a bidirectional model is very different from the unidirectional case. For instance, when we talk about features like agreement and case, and their influence on grammatical function assignment in a language like German or Dutch, we typically consider this influence to be near absolute in comprehension. We do not expect an accusative NP to be interpreted as a subject, just because its referent happens to be human or because the NP is definite. In a unidirectional comprehension model, this corresponds to ranking case constraints very high. In the models of this chapter, I have not even bothered to define case and agreement constraints explicitly for this very reason. As a result, we would have to predict that NPs that show case, or NPs that have a verb agreeing with them, are always assigned the correct grammatical function – freezing of such NPs would be a mystery.

The Russian example shows that in a bidirectional model, the correspondence between being a constraint that is never broken in a grammatical sentence and being highly ranked does not hold anymore. The constraint GENDERAGREEMENT is not ranked high, which leads to the fact that *comprehension* can ignore the presence of gender agreement. However, this does not mean that the whole model predicts that a sentence with the 'wrong' gender agreement is grammatical – *production* forces satisfaction of the constraint. Ranking a formal constraint below SUBJECT-LEFT does not necessarily mean that we predict that a meaning assignment that violates the formal constraint is possible. It only means that the formal constraint in question does not supply enough information to prevent freezing.

5.5.6 Summary

In this section, we have seen that a strong bidirectional model in combination with stratified ranking is flexible enough to deal with ambiguity and optionality. The counterex-

amples brought forward by Zeevat and Flack can be dealt with elegantly, often by considering comprehension more closely. As I have indicated, the proposed analyses require refinement and suggest areas of further research. However, I have shown that it is possible to model freezing and exceptions to freezing using linguistically motivated, and mostly existing constraints. I have also shown that problems for the extended unidirectional models are easily solved by the bidirectional approach, apart from the issue of plausibility.

The analyses of the cases of ambiguity and optionality always followed the same template: Find a constraint that forces neutralization in one direction, and find two conflicting constraints to bring out variation in the other. Because we were dealing with word order, one of the conflicting constraints was always SUBJECT-LEFT. The case of ambiguity of wh-questions is a clear example: Production neutralization is forced by WH-LEFT and comprehension variation by {SUBJECT-LEFT,*SUBJECT/INDEFINITE}. This template shows that stratified bidirectional OT makes a strong claim about the nature of ambiguity and optionality. Ambiguity and optionality always require a *double explanation*. Many-to-one/one-to-many mappings need to be caused by constraints in production as well as in comprehension. Even instances of ambiguity and optionality completely unrelated to the material discussed in this chapter need to receive such a two-fold explanation. Although I hope to have shown that this double explanation can be given in some cases, it remains to be seen whether a double explanation can be found for all ambiguity or optionality. This suggests a good strategy for generating future research.

Although I have been using the strong bidirectional model, I have not discussed alternative setups. There is a plethora of bidirectional models available, and I will not be able to consider all of them carefully. In the next section, I will however discuss a few important alternatives. I will concentrate on the combination of bidirectional OT with variable rankings. The conclusion of that discussion will be that strong bidirectional OT is a good choice, because it does not run into certain problems, while staying simple. The choice for stratified OT over another approach to variation in OT (that is, Stochastic OT) will also be discussed in the next section.

5.6 Combining variation, production, and comprehension

In this section, we shall take a more detailed look at the combination of variation and bidirectional OT, and the production and comprehension optimization in bidirectional OT. In the literature, there are two salient proposals to model variation. We have seen Anttila's stratified rankings at work in this chapter. I will also introduce the other proposal, Stochastic OT, and explain why I have not used Stochastic OT here, although it may be an interesting alternative for future work.

There have been many proposals in the OT literature on how to combine the two directions of optimization. I shall discuss those proposals here in terms of what kind

of phenomena we can explain with them and what we predict to be impossible in a language. Good overviews of different types of bidirectional OT can be found in Beaver and Lee (2003; 2004). For a brief discussion of different types of bidirectional OT in the context of word order freezing, see Zeevat (2006, s3).

5.6.1 Variation in strong bidirectional OT

To achieve variation I have relied on Anttila's (1997) stratified grammars. An output candidate is optimal in a stratified grammar if it is optimal in any of the fully ranked grammars that are compatible with it. In (91), the definition I gave of stratified strong bidirectionality in (56) is repeated.

(91) A form-meaning pair $\langle f,m\rangle$ is grammatical in a stratified grammar S (a set of compatible full rankings), iff
  a. $\langle f,m\rangle \in Gen$
  b. and there is an $s \in S$ such that
    i. there is no $\langle f',m\rangle \in Gen$ such that $\langle f',m\rangle \succ_s \langle f,m\rangle$ (prod.)
    ii. and there is no $\langle f,m'\rangle \in Gen$ such that $\langle f,m'\rangle \succ_s \langle f,m\rangle$ (comp.)

I will discuss two independent alternatives to the definition in (91). First, we may allow the two directions to use different rankings. Secondly, we may choose to use an alternative to Anttila's stratified rankings, Stochastic OT (Boersma, 1997; Boersma and Hayes, 2001). In the context of the latter, I will also discuss a topic that has been left largely ignored in this chapter: Ambiguity and optionality in combination with a preference for one of the variants.

The definition in (91) requires that the full constraint ranking is the same between production and comprehension. A less restricted alternative would be to allow the ranking to change between directions, defined in (92).

(92) (Alternative)
A form-meaning pair $\langle f,m\rangle$ is grammatical in a stratified grammar S (a set of compatible full rankings), iff
  a. $\langle f,m\rangle \in Gen$
  b. and there is an $s \in S$ such that
    there is no $\langle f',m\rangle \in Gen$ such that $\langle f',m\rangle \succ_s \langle f,m\rangle$ (prod.)
  c. and there is an $s \in S$ such that
    there is no $\langle f,m'\rangle \in Gen$ such that $\langle f,m'\rangle \succ_s \langle f,m\rangle$ (comp.)

The definition in (92) defines languages that are supersets of the languages defined by (91). The form-meaning pairs that are grammatical when both directions use the same full ranking are captured by both. However, there might be form-meaning pairs that are

grammatical only when the rankings in comprehension and production are different. This situation is captured by (92), but not by (91). We have not seen any data that would suggest that we fail to predict grammaticality of form-meaning pairs that rely on different rankings in comprehension and production. Therefore, I will continue with the more conservative (91).[31]

Before I can go into the second topic, Stochastic OT, we will need to discuss an aspect of ambiguity and optionality that I have ignored thus far. Even when a meaning m may optionally realized by form f_1 or by form f_2, it may very well be the case that there is a preference for one of the realizations. Likewise, it may be that a form f is preferably interpreted as m_1, but that m_2 is a possible, only less salient reading.

Preferences for a certain realization or interpretation are only captured in part by the model I have defended in this chapter. In the introduction, I noted that the SVO reading of an NP–V–NP sentence in Dutch could be argued to be 'no more' than a strong preference – a preference that is overridden by the right context, for instance. This preference is captured by the bidirectional model of word order. The model predicts that NP–V–NP is interpreted as SVO unless there is (contextual, intonational, etcetera) evidence that the OVS interpretation is correct.

However, it may also be that there are preferences that cannot be given such an external explanation. The optionality of topic-fronting could be an instance of such a preference. Even though fronting a topic in a language is not obligatory (that is, SUBJECT-LEFT and TOPIC-LEFT are in one stratum), it may be the case that people prefer to do so if they can. Fronting a topic may simply be the more common thing to do. Anttila proposes a quantitative interpretation of stratified rankings to capture these kinds of preferences in a unidirectional production model. A meaning is preferably realized with a certain form, if this form is optimal under most of the full rankings compatible with the stratified grammar. Anttila links these preferences to corpus frequencies. He proposes that observed frequency of a certain realization of a meaning should be approximated by the proportion of competitions in which this realization wins. In the words of Anttila and Fong (2000) it reads as (93).

(93) If a candidate wins in n tableaux and t is the total number of tableaux in the partial order, then the candidate's probability of occurrence is n/t. (their 46b)

We can also apply this quantitative interpretation to our bidirectional grammar. Take the grammar in (74), p211, here repeated as (94).

[31]There are even less restrictive options that I will not consider here. For instance, one might allow the stratification to be different between production and comprehension. A radically different approach would be to use completely different grammars, with different constraints, in each direction. See Zeevat (2000) for a proposal along these lines in the context of *production-filtered comprehension optimality* (explained on page 235 of this dissertation).

(94) WH-LEFT $\gg$ { SUBJECT-LEFT, *SUBJECT/IND, 1ST1ST, COHERE }

An overview of the predictions that this grammar makes was given in Figure 5.3, p217. Take the input call(ιboy, $\underline{\exists\text{girl}}$) in the context of $\underline{\text{girl}}$. Subfigure (a) tells us that the grammar predicts that this form is optionally realized as *de jongen ziet een meisje* (SVO) and *een meisje ziet de jongen* (OVS). At this point, the numbers in the connecting lines in Figure 5.3 become relevant. These numbers indicate the number of full rankings ('tableaux' in Anttila and Fong's terms) under which a form-meaning pair is optimal. In the case of our example, both realizations were optimal under 8 rankings. A total number of $4! = 24$ full rankings are compatible with the grammar in (94). Therefore the quantitative interpretation predicts that the probability of an SVO realization for this meaning is $\frac{8}{24} = 33\%$, and the probability of an OVS realization is $\frac{8}{24} = 33\%$, too.

The quantitative interpretation immediately leads to a problem: The predicted probabilities do not add up to 100%. This is because under 8 of the 24 rankings, no bidirectionally optimal form is found. In a unidirectional model, this does not happen, because there is always a winning candidate and therefore an optimal pair. We could interpret the missing 33% as ineffability. If ineffability means that a meaning cannot be expressed with a certain construction, 33% ineffability can be taken to mean that in a third of the cases a speaker would use a completely different construction to express this meaning (for instance, a passive or a cleft). This reasoning cannot be applied to ambiguity however. If a form is mapped to a meaning in 18 out of 24 rankings, we would have to say that one quarter of the time the form is ungrammatical/uninterpretable. However, to predict that a form is both grammatical and ungrammatical or interpretable and uninterpretable at the same time seems absurd.

Anttila and Fong (2000) offer an alternative quantitative interpretation principle, given in (95).[32]

(95) (Alternative quantitative interpretation:)
If n is the number of rankings under which a point x is connected to a point y, and t is the total number of rankings under which x is connected to any point, then the probability that x is interpreted/realized as y is n/t. (after their 66b).

The alternative quantitative interpretation principle relates probability to the number of rankings that yield an interpretation/realization, rather than the number of possible rankings. Let us return to our example. The grammar in (94) connects call(ιboy, $\underline{\exists\text{girl}}$), in the context of $\underline{\text{girl}}$, to a meaning under 16 rankings. In 8 of these rankings the optimal candidate is SVO (50%), in another 8 the optimal candidate is OVS (50%). The grammar therefore predicts that the alternation of form is fully optional in this case. Of course, this prediction is completely hypothetical. We have no evidence whatsoever that such a claim

[32] I have modified their formulation to be agnostic about form and meaning.

is realistic. One could conduct psycholinguistic experiments, or detailed corpus studies to collect evidence about preferences for one realization or the other. I reserve this topic for further research. In the corpus investigation of freezing in Chapter 6, I will also test predictions of the bidirectional model. However, the hypotheses that are tested in that chapter relate to the bidirectional model in a less direct way.

The quantitative interpretation of variation is central to Stochastic OT (Boersma, 1997; Boersma and Hayes, 2001), and it enters the model already at the stage of constraint ranking. The general method to capture variation in a language in an OT model is to allow a language particular grammar to consist of several full rankings; in optimization each ranking predicts another variant to be grammatical. In Anttila's approach this was achieved by allowing language particular grammars to be underspecified or stratified. Boersma proposes that a language particular grammar is a ranking of constraints on a continuous scale – a kind of meta-ranking. From this continuous scale, we generate the different full rankings to be used in optimization, but each full ranking is generated with a probability. The chance that two constraints have a certain order in a full ranking is determined by their relative position on the continuous scale. Optimization itself proceeds as in classic OT with a full constraint ranking. To illustrate, consider the two continuous scales in (96).

(96) a. WH-LEFT SU-LEFT *SU/IND

b. WH-LEFT SU-LEFT *SU/IND

In both continuous scales, WH-LEFT is placed far above the other two constraints. This means that in optimization, the chance that either one of SU-LEFT or *SU/IND outranks WH-LEFT is negligibly small. Also, in both scales, SU-LEFT is placed above *SU/IND. This means that the chance that SU-LEFT $\gg$ *SU/IND in optimization is greater than the chance that *SU/IND $\gg$ SU-LEFT. However, the closer two constraints are to each other on the continuous scale, the more alike the chances of the two possible rankings in optimization become. Thus, the chances of *SU/IND $\gg$ SU-LEFT are better in (96b) than in (96a). If two constraints are at the same point on the continuous scale, either order can occur with a chance of 50%. Compared to stratified ranking, Stochastic OT gives us additional control over the constraint rankings that are used in optimization. Not only can we specify for a certain language that constraint order can vary, we can also specify how likely each ordering is. In essence, a language particular grammar in Stochastic OT assigns a probability to each of the universally possible full rankings.

The quantitative interpretation of the predictions of a Stochastic OT model is a lot like Anttila's quantitative interpretation, but we have to take the probability of a ranking into account. For example, suppose there are two possible full rankings r_1 and r_2. The probability of r_1 is 80%, and the probability of r_2 is 20%. If candidate f_1 is optimal for m under r_1, and f_2 under r_2, then m is predicted to be realized as f_1 80% of the time, and

as f_2 20%. Stochastic OT also comes with a learning algorithm. The learning algorithm combined with the added control over the quantitative predictions make Stochastic OT an attractive alternative to stratified rankings.

However, recall from end of the discussion of the quantitative interpretation of stratified rankings that at the level of abstraction I have been working at, we have no idea of how quantified preferences and/or corpus frequencies are related to the data. Since ranking constraints on a continuous scale does not make much sense without the quantitative interpretation, I have chosen to use Anttila's approach over Stochastic OT.

There is also a second problem with the combination of Stochastic OT and bidirectional OT. We saw that the alternative quantitative interpretation of Anttila and Fong, defined in (95), is the more attractive way to predict form-meaning frequencies from a bidirectional model because it avoids the problem of part-time ungrammaticality. However, in Stochastic OT, this leads to unwanted results. A language particular grammar in Stochastic OT assigns a probability to each ranking, but it is not possible to assign a zero probability to a certain ranking to indicate that it will never be used in optimization. All one can do is to assign an exceedingly small probability to that ranking (Boersma and Hayes, 2001). So, in (96), the chance that WH-LEFT is outranked by another constraint is tiny, but not non-existent. Suppose we want to see what the predictions are that a *stochastic strong bidirectional OT* model makes using the alternative quantitative interpretation, for the realization of an object-question with an indefinite subject under the grammar of (96a). Let us assume that the chance that WH-LEFT outranks SU-LEFT is 99.99%, and that the chance that SU-LEFT outranks WH-LEFT is 0.01% The input to production is an object question with an indefinite subject. In production, 99.99% of the winners is O[WH]VS, and 0.01% SVO[WH]. Note that the latter, wh-in-situ, is not grammatical in Dutch, and that this is correctly reflected in the production percentages. In comprehension, none of the O[WH]VS forms will lead to a bidirectional optimal pair, because no ranking of WH-LEFT, SU-LEFT and *SU/IND will select an object-question as the winner. However, SVO[WH], which was optimal in a small fraction of the production competitions, will be interpreted as an object-question with an indefinite subject. In fact, each possible ranking of the three constraints selects this interpretation. As a result, there are no bidirectional optimal pairs with O[WH]VS, and there are bidirectional optimal pairs with SVO[WH]. Under the alternative quantitative interpretation of (95), this means that SVO[WH] is predicted to occur 100% of the time, and to be fully grammatical. This is plainly wrong given the fact, established in Section 5.5.2, that object-questions with indefinite subjects are ineffable.

To summarize, there are alternative ways of combining variation and strong bidirectional OT. Stratified strong bidirectional OT as I have used it in this chapter is a very restrictive and conservative approach compared to some other approaches. I have also sketched how

we could give a quantitative interpretation to our model. Considering the fact that we as yet do not know anything quantitative about the data, I have refrained from actually doing so. The primary reason that I am using Anttila's stratified rankings to model variation in stead of Boersma's Stochastic OT, is that the former can also be sensibly used without giving the model a quantitative interpretation. In addition, we have seen that the combination of Stochastic OT and bidirectional OT is problematic. With these conclusions in mind, we can now discuss alternatives to the strong bidirectional model.

5.6.2 Symmetric bidirectionality

We will start the discussion of alternative bidirectional setups with symmetric bidirectional models. In these models, the influence of comprehension on production is an exact mirror image of the influence of production on comprehension. Note that, as I have done in the chapter until now, I mean by production and comprehension the directions of optimization, respectively from meaning to form and from form to meaning. The relation of these directions to grammaticality and interpretation depends on the exact definition of optimality.

Strong bidirectionality Strong bidirectionality (Blutner, 2000), the type of bidirectionality that we have seen thus far, is the most restrictive and simple form. For convenience, I have repeated the definition in (97).

(97) A form-meaning pair $\langle f,m\rangle$ is grammatical under *strong bidirectional optimality* iff
  a. $\langle f,m\rangle \in Gen$
  b. and (*production optimality*)
     there is no $\langle f',m\rangle \in Gen$ such that $\langle f',m\rangle \succ \langle f,m\rangle$
  c. and (*comprehension optimality*)
     there is no $\langle f,m'\rangle \in Gen$ such that $\langle f,m'\rangle \succ \langle f,m\rangle$

In both directions, the optimal output is simply the most harmonic candidate. This means that production optimization and comprehension optimization do not influence each other. However, production does influence *interpretation*, because a meaning can be barred from being assigned to a form if that form is not production optimal. Likewise, comprehension influences grammaticality by preventing forms that do not have the right comprehension optimal meaning from being grammatical.

As Beaver and Lee (2004) point out, the reason that strong bidirectional OT is so restrictive is that the set of the optimal form-meaning pairs is the intersection of the optimal form-meaning pairs from production optimality and comprehension optimality. We have already seen that under a classic ranking regime the two perspectives excel

in opposite directions. Production has no problem modelling ambiguity, but struggles with optionality, and comprehension easily deals with optionality, but not with ambiguity. As a result, strong bidirectional OT copes well with neither. See Section 5.5 for more discussion.

Blutner's original proposal only allowed markedness constraints. The idea is that mapping between form and meaning is a kind of alignment of form and meaning markedness. The unmarked forms pair up with the unmarked meanings. In spirit, this renders faithfulness constraints superfluous; relations between form and meaning emerge from this alignment. However, in such a framework, one ends up with only *very* few optimal pairs. Typically, only one form-meaning pair – or to be more precise: one constraint profile – per candidate set is bidirectionally optimal.[33]

Weak bidirectionality There is an follow-up to strong bidirectionality that is designed to deal with so-called *partial blocking*. Strong bidirectionality pairs unmarked form with unmarked meaning. As a result, marked meanings are blocked from being expressed with unmarked forms and marked forms are blocked from expressing unmarked meanings. Weak bidirectionality tries to capture the (Gricean) intuition that a marked meaning may pair up with a marked form. Weak bidirectionality was proposed in Blutner (2000), along with strong bidirectionality. The definition of weak bidirectionality is given in (98).

(98) A form-meaning pair $\langle f,m\rangle$ is grammatical under *weak bidirectional optimality* iff
  a. $\langle f,m\rangle \in Gen$
  b. and (production optimality)
     there is no *comprehension optimal* $\langle f',m\rangle \in Gen$ such that $\langle f',m\rangle \succ \langle f,m\rangle$
  c. and (comprehension optimality)
     there is no *production optimal* $\langle f,m'\rangle \in Gen$ such that $\langle f,m'\rangle \succ \langle f,m\rangle$

[33] As an aside, it is tempting to present the definition of a strong bidirectionally optimal pair as simply the one most harmonic form-meaning pair. In general, this is a false simplification of the definition in (97), although it may be the case that it holds under certain conditions. The simplification fails for instance when the candidate set contains *unconnected* form-meaning pairs: pairs that cannot be linked to each other through form-meaning pairs that share form or meaning. No matter what the constraint profiles are like, if $Gen = \{\langle zie,\text{see}\rangle, \langle hoor,\text{hear}\rangle\}$, both candidates will be strong bidirectionally optimal. If one would pick only the most harmonic pair, one would not retrieve all strong bidirectional pairs. A more realistic example of incompleteness of the simplification with respect to the definition of strong bidirectionality is a grammar that includes faithfulness constraints. Again, we can end up with more than one bidirectional optimal. For instance, take $Gen = \{1,2\} \times \{a,b\}$, and constraints $1 \rightarrow a$, and $*a$, with $1 \rightarrow a \gg *a$. In this case, input 1 prefers output a, and input a is happy with either output 1 or 2, so $\langle 1,a\rangle$ is bidirectionally optimal. Also, 2 prefers b, and b prefers 2 (because of $1 \rightarrow a$). So $\langle 2,b\rangle$ is bidirectionally optimal, too.

I *suspect* that the simplification of strong bidirectional optimality to 'the most harmonic form-meaning pair is optimal' is complete when a) there are no faithfulness constraints involved, and b) the form-meaning pairings are exhaustive. Both conditions are met in Blutner's original proposal. With appropriate modifications, this little *nota bene* carries over to weak bidirectionality, too.

Note that this is like strong bidirectionality, with an added requirement on what counts as a competitor. Competitors have to be optimal in the opposite direction of optimization. Due to the stronger requirements on competitors, candidates become optimal that were otherwise outperformed by the excluded competitors. The set of weak bidirectionally optimal pairs is a superset of the set of strong bidirectional optimal pairs.

The definition of weak bidirectionality is recursive, but Jäger (2002) has shown that this is not a problem.[34] In a nutshell, the recursion bottoms out on the best constraint profile, of which there is only one. For a given form-meaning pair it is therefore always possible to calculate whether they are weak bidirectionally optimal. Procedurally, one can find weak bidirectionally optimal pairs amongst all form-meaning pairs by first finding the strong bidirectional optimal pairs, and then removing all pairs that share either form or meaning with a strong bidirectional optimal pair. Every strong bidirectionally optimal pair is weak bidirectionally optimal. The pairs that share either form or meaning with these strong pairs would lose out to the strong pairs in production (shared form) or comprehension (shared meaning). In this pruned candidate set, one then starts again to look for strong bidirectionally optimal pairs, etcetera. Each 'round' of optimization can yield additional optimal pairs.

It is not clear that we would gain anything by using weak bidirectional OT in the analysis of word order freezing. Note that weak bidirectional OT does not in itself allow one form to be paired with more than one meaning, or vice versa. Weak bidirectionality does therefore not suffice to introduce ambiguity and optionality in the grammar. Let me illustrate this for optionality. Take two pairs $\langle f_1, m\rangle$ and $\langle f_2, m\rangle$ that do not have the same constraint profile. These two will never both be weak bidirectionally optimal because either a) one of them is not comprehension optimal, or b) both are, in which case the most harmonic wins the production competition. With respect to ambiguity and optionality, weak bidirectionality is as restrictive as strong bidirectionality.

Weak bidirectionality does make form-meaning pairs grammatical that are considered ungrammatical under strong bidirectionality because they consisted of a marked form and a marked meaning. Since we have seen no indication in our word order data that this kind of analysis is called for, I do not see a need to abandon strong bidirectionality in favour of the more complicated weak bidirectionality.

[34]The introduction of variable ranking may break Jäger's result if one is not careful with the scope of the choice of full ranking in the definition. If one is allowed to pick a different ranking for every round/level of recursion, one effectively changes the harmonic ordering between two levels. This means that what is the best constraint profile may also change. If what is the best constraint profile flips to and fro between recursions, we can descend an endless chain of ever better profiles. As a result, the recursion does not stop, and the definition of weak bidirectionality is problematic. Of course, the problem is easily prevented by not allowing the full ranking to change between recursions.

Medium-strength bidirectionality Weak bidirectionality suffers from over-generation. The recursion will continue to add optimal pairs as long as not all candidates are blocked. Beaver and Lee (2003; 2004) show that this behaviour soon leads to undesirable results. To preserve the intuition behind weak bidirectionality, viz. that marked forms go with marked meanings, Beaver and Lee propose a compromise, dubbed *medium strength bidirectionality*. Procedurally, it amounts to weak bidirectionality with only one recursive step. It is used in the form of a special constraint *BLOCK in Beaver (2004) and Lee (2004). As with weak bidirectional OT, there are no reasons to adopt this approach in our current model. We have not seen any cases where we miss a form-meaning pair because it would involve a marked form and a marked meaning.

5.6.3 Asymmetric bidirectionality

Thus far the models have been symmetric in nature: Comprehension and production play equal parts in bidirectional optimization. However, asymmetric proposals can also been found in the literature. These models can be characterized as hybrids between strong and weak bidirectionality (Buchwald et al., 2003).

A comprehension filter on production In the beginning of the chapter, we saw how Lee (2001b) modelled recoverability by adding a separate comprehension requirement to the initial production model (Section 5.3). Another way of ensuring recoverability is by moving the comprehension requirement into production optimization. Only form-candidates that are interpreted as the input meaning are allowed to compete in production optimization. Thus comprehension acts as a filter on the production candidate set. The definition of grammaticality under this concept of asymmetric bidirectionality is given in (99).

(99) A form-meaning pair $\langle f,m\rangle$ is grammatical under *comprehension-filtered production optimality* iff
  a. $\langle f,m\rangle \in Gen$
  b. and (production optimality)
     there is no *comprehension optimal* $\langle f',m\rangle \in Gen$ such that $\langle f',m\rangle \succ \langle f,m\rangle$
  c. and (comprehension optimality)
     there is no $\langle f,m'\rangle \in Gen$ such that $\langle f,m'\rangle \succ \langle f,m\rangle$

We are in fact using the definition of comprehension from strong bidirectionality, and the definition of production from weak bidirectionality. This setup has been proposed in Wilson (2001) and used in Jäger (2004) and Buchwald et al. (2003, under the name of 'Recoverability OT').

The definition is appealing because it naturally captures a kind of fall-back behaviour. An otherwise suboptimal form may win the production competition because it is the best form that will be interpreted correctly. This can be exploited in a model of word order to capture freezing. To illustrate, consider a grammar COHERE ≫ 1ST1ST ≫ SU-LEFT. As the input to production we take the meaning call(frank, ella), in a context that sets up for Ella to be important. The output candidates are SVO and OVS. The two comprehension tableaux in (100) show that of these two, only SVO leads to the correct meaning. Therefore, only SVO participates in production in (100). The OVS candidate is included in the production tableau, but striked out to indicate that it is taken out of the competition.

(100) Context: ella
Production:

belt(frank, ella)	COHERE	1ST1ST	SU-LEFT
~~Ella belt Frank~~			*
✌ Frank belt Ella		*	

Comprehension:

Ella belt Frank	CO	1ST	S-L
☞ belt(ella, frank)			
belt(ella, frank)	*!	*	
belt(frank, ella)	*!	*	*
belt(frank, ella)			*!

Comprehension:

Frank belt Ella	CO	1ST	S-L
belt(ella, frank)		*	*!
belt(ella, frank)	*!		*
belt(frank, ella)	*!		
☞ belt(frank, ella)		*	

We see that the SVO form, which is otherwise suboptimal for an object-important meaning (1ST1ST ≫ SU-LEFT), is predicted to be grammatical when there is additional information about information structure (context), but none about grammatical function (case). Comprehension-filtered production OT would therefore predict that there are languages that allow frozen sentences with object-important interpretation, even though other canonical word order sentences cannot have this interpretation. Comprehension-filtered production is much like the Zeevat/Flack unidirectional models in this respect, although unlike Zeevat/Flack, comprehension-filtered production is able to use non-form information to prevent freezing. In Section 5.5.3, I explained that the stratified strong bidirectional model predicts that languages in which frozen sentences receive other interpretations than canonical sentences do not exist.

The fall-back behaviour just sketched is sometimes undesirable. For instance, with the constraints I used to model the ambiguity in questions (WH-LEFT ≫ {SU-LEFT, *SU/IND}), the strong bidirectional model predicted that object-questions with an in-

definite subject are ineffable (Section 5.5.2). Comprehension-filtered production would predict wh-in-situ to be grammatical in such cases. The example in (101) is incorrectly predicted to be grammatical.[35]

(101) *Een jongen belt wie
A boy calls who
'Who is a boy calling.'

Comprehension-filtered production makes this prediction because it is the only form that passes the comprehension filter. There is no way of formulating that WH-LEFT cannot be violated to satisfy the comprehension filter. Comprehension-filtered production shows this behaviour in general. To use the Rat/Rad-metaphor again, this architecture would predict that pronouncing /ra:d/ 'wheel/bicycle' as [ra:d] is grammatical to avoid confusion with /ra:t/. This is clearly undesirable.

As a way around this, one might model the comprehension filter as a violable constraint, instead of as an absolute requirement on candidates. A candidate may break the comprehension-filter, and still participate in competition, just as it in principle may violate any other constraint and still compete. Ranking the constraint – call it RECOVER – below WH-LEFT is saying that wh-fronting is not stopped by recoverability considerations. Likewise, ranking RECOVER below *VOICEDCODA expresses that we cannot pronounce wheel as [ra:d] to guarantee comprehension. RECOVER would have to be a very special constraint. To avoid problems of circularity, it has to be switched off in comprehension.

Jäger (2004) uses a modification of this asymmetric type of bidirectionality in his EvolOT-programme. EvolOT is a bidirectional constraint ranking algorithm based on the ranking algorithm associated with StOT. However, in order to avoid the problems with ineffability and frequency prediction that I discussed in the previous subsection, Jaeger makes the comprehension filter conditional: It is only applied if there is a candidate that passed the comprehension filter. This means that there always is an optimal form-meaning pair because the model reduces to unidirectional production optimization when there would be ineffability. Jäger's approach is equivalent to saying that there is a constraint RECOVER that always ranks highest. Note that this still leads to the undesirable fall-back behaviour explained above.

A production filter on comprehension The opposite of a comprehension filter on production would be a production filter on comprehension. In this setup, comprehension optimality refers to production optimality, but not vice versa. This is in fact Lexicon Optimization, as proposed in early work on OT (Prince and Smolensky, 1993/2004).

[35]The '*' refers to suitability as a neutral question. Echo questions, etcetera, are possible with wh-in-situ.

(102) A form-meaning pair $\langle f, m \rangle$ is grammatical under *production-filtered comprehension optimality* iff

a. $\langle f, m \rangle \in Gen$

b. and (production optimality)
there is no $\langle f', m \rangle \in Gen$ such that $\langle f', m \rangle \succ \langle f, m \rangle$

c. and (comprehension optimality)
there is no *production optimal* $\langle f, m' \rangle \in Gen$ such that $\langle f, m' \rangle \succ \langle f, m \rangle$

Zeevat (2000) also advocates such an approach, but proposes to use two completely different grammars for the directions. Production-filtered comprehension could be used to model word order freezing, although without variable ranking it suffers from much the same problems as Lee's (2001b) original proposal. Production-filtered comprehension also shows fall-back behaviour, but now in comprehension. An otherwise suboptimal meaning becomes optimal for a form when it is the best meaning that would be produced as this form. This would provide an explanation for the data leading to the ranking paradox in Section 5.5.3. The problem was that on the one hand canonical word order should be assigned a meaning that obeys COHERE, even when it means violating 1ST1ST. On the other hand, a non-canonical word order should be interpreted as satisfying 1ST1ST, even when it means violating COHERE. In production-filtered comprehension, this does not lead to a ranking paradox. COHERE $\gg$ 1ST1ST captures not only the canonical word order facts, but the non-canonical word order facts, too. Non-canonical word order is only production optimal for certain information structural configurations. Therefore, all candidates in comprehension in production-filtered comprehension will have this information structure. COHERE may be violated simply because there is no candidate that passes the production filter that satisfies the constraint.

I will leave investigation of further predictions of production-filtered comprehension for future work. However, I will point out a possible computational problem with this kind of asymmetric model. Kuhn (2003, ch6) has shown that the recognition problem for unidirectional production OT, in his LFG formalization, is undecidable. There is no reason to assume this rests on peculiarities of his formalization. The recognition problem for a grammar formalism is as follows: Given an arbitrary string and a grammar, decide whether the string is part of the language predicted by the grammar. Informally, Kuhn's proof proceeds as follows: For unidirectional production OT, a form is grammatical if that form is an optimal realization of some input. This seems unproblematic, because one can just start from any input and look at its optimal output. If it matches, we are done; if it doesn't, we have to try with another input. The problem is that because of faithfulness violations, we cannot be certain what the relation between the desired input meaning and the given output form is: Any amount of meaning can remain *unexpressed* in the output. Because there is no systematic way of exploring this infinite space of possible inputs, we will never stop looking. For instance, we cannot use the constraint profile to guide

our search because we do not know how harmonic the form-meaning pair that we are looking for will be. Hence, the recognition problem is undecidable for unidirectional OT. As production-filtered comprehension relies on unidirectional production to supply its candidate set for comprehension, it inherits this decidability issue. In production-filtered comprehension, even GEN is a problem.

Kuhn offers two remedies for the recognition problem. The first is to constrain the amount of meaning material that can remain unexpressed in form, by relating it to a representation of the context. As an example, Kuhn uses ellipsis. The fragment *and John* can be an elliptic utterance with any amount of meaning unexpressed in the first, elided, conjunct. However, the unexpressed meaning is restricted by the context and, if we assume the amount of material in the context that is suitable for elision is finite, the meaning search space should be finite, too. Note that this remedy applies vacuously to the domain of word order variation in this chapter, because there is only a finite amount of meaning material that remains unexpressed (grammatical function, information structure).

The second option Kuhn offers, which allows for a less trivial restriction of the search space, is to move to (strong) bidirectional OT. In strong bidirectional OT, the meaning that we are after has to be amongst the comprehension optimal candidates for the given form, which means that we have a way of systematically searching the space of meanings. We have seen in this chapter that a strong bidirectional model can be quite successful.

We have seen that combining production and comprehension in a bidirectional OT model can be done in many different ways. Each way of combining production and comprehension has its pros and cons. However, none of the combinations discussed here offered real advantages over the simplest combination – strong bidirectionality.

It is still an open question how to relate frequencies to the bidirectional model that I have been using. The choice for Anttila's stratified rankings over Stochastic OT is motivated by this. However, I do consider this to be a very important question. If we have a way to derive frequencies from the model directly, we could bridge the gap between the theoretical work, and empirical corpus investigations, without having to stipulate a connection.

5.7 Conclusion

In this chapter, I have discussed a wide range of aspects of word order freezing, and shown that freezing is a good argument for genuine bidirectionality in grammar. Also, I have argued that the straightforward implementation of bidirectionality in the form of Blutner's strong bidirectionality suffices and is even preferable over other formalizations in certain respects. If we give proper attention to the importance of interpretation preferences and context, the proposed stratified strong bidirectional model of word order is capable of

elegantly capturing freezing and exceptions to freezing. The analysis still contains many open areas that need to be filled in, and throughout the chapter we have seen many ways in which this research may be carried forward.

I have already announced at several places that the results of the theoretical investigations of this chapter will be used in a corpus investigation. A corpus investigation of freezing is important, because it may give a stronger empirical foundation to freezing. In the theoretical discussion in this chapter, we have identified several factors that influence the interpretation of word order: definiteness, animacy, focus/intonation and context. The next chapter will investigate whether the factors influence word order freezing in the specific way that bidirectional OT predicts.

More support for a bidirectional model may also come from other phenomena that can be fruitfully analyzed as word order freezing, or as similar in spirit to word order freezing. I would like to give one example of each. The first example is a constraint on *embedded object scrambling* in Dutch, which could be explained in terms of word order freezing. The second example is V2/V3 variation in German. Although this is not word order freezing as such, V2/V3 variation is subject to a constraint that is of similar nature. Further study will have to reveal whether these cases can indeed be modelled using strong bidirectional OT.

Embedded object scrambling Like the data in this chapter, the first example pertains to argument order variation. In this case, however, the domain is Mittelfeld scrambling in Dutch. A pronominal object that belongs to an embedded verb may scramble over an object belonging to the matrix verb. Consider the AcI construction in (103). Example (103a) gives the scrambled order, (103b) the canonical one. As an aside, the scrambled (103a) is preferred over the canonical (103b).

(103) a. Ik hoor **het** je al zeggen
I hear it you PART say
'I can hear you say it.'
b. Ik hoor je **het al zeggen.**

The direct object of the main verb is *je*, which is also the understood subject of the embedded verb; the direct object of the embedded verb is *het*. Interestingly, embedded object scrambling is restricted to inanimate embedded pronominal objects (Van der Beek, 2005). Van der Beek gives the following contrast:

(104) a. Ik heb ze Jo door zien slikken.
I have them Jo VPART see swallow
'I saw Jo swallow them.'
b. Ik heb ze Jo zien zoenen.
I have them Jo see kiss
'I saw them kiss Jo.' *Not:* 'I saw Jo kiss them.'

This contrast is understandable from a freezing perspective: Animacy is information about what is to serve as the subject of the embedded verb, and, in an AcI, therefore about what is the direct object of the main verb. If animacy does not distinguish between the two NPs in the Mittelfeld, word order is the only information available about which is the direct object, and which is the embedded object. Word order freezing is the result, as (104b) shows.

V3 in German German (and Dutch, too) allows certain adverbials to appear before the Vorfeld, which results in a V3 sentence. An example is given in (105). Note that the verbal part *gesagt* 'said' may be omitted (Meinunger, 2004).

(105) Ehrlich (gesagt), ich **bin** von dir total enttäuscht.
honestly said I am by you totally disappointed
'To be honest, I am completely disappointed with you.' (his 21/22)

These speech act adverbs may also appear in the Vorfeld, resulting in V2. However, in this case, only the more complete form is acceptable, as shown in (106).

(106) Ehrlich *(gesagt) **bin** ich von dir total enttäuscht.
honestly said am I by you totally disappointed
'To be honest, I am completely disappointed with you.' (his 24/33)

Meinunger (2004) convincingly argues that this is related to the fact that *ehrlich* 'honestly' on its own may also take a clause internal role, in the form of a manner adverb.

(107) EHRlich kann man sich in solchen Situationen GAR nicht verhalten.
Honestly can one REFL in such situations PART not behave
'It's just impossible to behave HONESTLY under those circumstances.' (his 37)

German also has adverbs that can only be understood as speech act adverbs. In contrast to *ehrlich*, these can appear on their own in the Vorfeld as speech act adverbs.

Although the constraint on variation that we see above cannot be expressed in terms of canonical word order, we can recognize in the data above the restriction that the intended reading has to be recoverable. In this respect the V2/V3 alternation is similar to word order freezing. When *ehrlich* appears in initial position of a V3 sentence (the topological *lead*), and/or when it is accompanied by *gesagt*, it is unambiguously recognizable as a speech act adverbial. However, when it appears on its own in the Vorfeld this is not the case – it might in principle be a manner adverb. The fact that the manner adverb reading ('disappointed in an honest manner' for 106) is odd does not play a role (cf. the fact that word order freezes in Japanese even though the resulting reading is unlikely).

A bidirectional model would be able to capture this, if it can be made plausible that there is a preference for adverbs in the Vorfeld to be interpreted as manner adverbs.

This could be because Vorfeld material is preferably interpreted as contributing to the semantics of the clause. A speech act adverb or speaker oriented adverb indicates the attitude of the speaker to the propositional content of the clause, but it does not contribute to this content. Just as SUBJECT-LEFT surfaces when grammatical function assignment is not indicated clearly enough (freezing), the preference on Vorfeld adverb interpretation becomes visible as soon as the Vorfeld material is not clearly marked as a speech act adverbial.

Let me conclude this chapter with a reflection on what could be the cause of the existence of freezing, in light of the model that I have used in this chapter. Some explanations for freezing assign a central role to *ambiguity*. Indeed, what relates many if not all examples of freezing is that as far as morphology is concerned, the sentence allows more than one grammatical function assignment. The exceptional enforcement of strict word order is then explained as a measure to avoid the looming ambiguity. Avoidance of ambiguity could be posited as a separate, general mechanism, but it is also what lies behind the contrast preservation or marking constraints proposed in the Flack/Zeevat accounts.

Given that language can be ambiguous in many ways and on many levels, and given that we have taken so much care to let our model predict ambiguity, citing ambiguity avoidance as the cause of word order freezing would not fit in the approach I have taken in this chapter. Instead, the stratified strong bidirectional OT model casts a different light on what causes freezing, bringing out a weaker version of Lee's recoverability requirement. In addition to the optimality requirement on a form of unidirectional OT Syntax, stratified bidirectional OT requires that the intended meaning is amongst one of the possible interpretations of the form. When this is not the case, the form is not grammatical. Therefore, what drives freezing is not the avoidance of ambiguity, but the avoidance of the situation in which the intended interpretation does not show up at all: guaranteed miscomprehension.

Chapter 6

A Corpus Investigation into Word Order Freezing

In the preceding chapter, I argued that speakers of Dutch take the chances of communicative success into account when they select a Vorfeld occupant. Estimating the chances of communicative success involves taking the standpoint of the hearer to see if the original meaning of an utterance can be retrieved. As support for the claim that word order is influenced by such communicative considerations, I extended a bidirectional Optimality theoretic account of word order freezing and showed that the bidirectional model was able to capture a range of intuition data better than alternative explanations. The defining property of a bidirectional model is that it combines speaker and hearer preferences to model grammaticality. In this chapter, I will gather further empirical evidence for the bidirectional word order model by means of a corpus study of word order freezing. In particular, I will investigate whether various predictions that the bidirectional model of word order makes about direct object fronting are observable in the spoken Dutch corpus.

The chapter starts with a discussion of the translation of the discrete predictions of the bidirectional model into hypotheses that can be tested on a corpus (Section 6.1). I will then continue with an investigation of the role of relative definiteness between the subject and the direct object in direct object fronting in Section 6.2. Similarly, Section 6.3 looks at the role of relative animacy. These two sections rely on statistical investigation of the spoken Dutch corpus. In Section 6.4, I will consider individual sentences from the corpus that may be expected to be counterexamples to word order freezing in Dutch. Section 6.5 concludes the chapter.

6.1 Preliminaries

Word order freezing refers to a lack of word order variation in otherwise free word order languages which occurs when there is not enough word order independent information to infer the correct grammatical function assignment from. Freezing has been observed in a range of languages that allow for word order variation. In Chapter 5, I argued that Dutch showed word order freezing as well, on the basis of examples like (1). Presented without indication of intonation or context, the sentence receives an SVO interpretation.

(1) De jongens zoeken de meisjes op.
the boys look the girls up
'The boys look up the girls.' (SVO)
Not, or strongly dispreferred 'The girls look up the boys.' (OVS)

In spite of the fact that it is possible to put a definite direct object in the Vorfeld, and to have the subject in the Mittelfeld in Dutch, sentence (1) is not easily interpreted as OVS. In Chapter 5, I defended the thesis that the lack of an OVS reading was due to the fact that there is nothing about (1) that tells a hearer that the sentence does not adhere to canonical word order. OVS is not available because there is nothing to indicate that the initial NP should be the direct object. For instance, there is no morphological information that indicates this; a situation that is traditionally connected to word order freezing.

The tendency of word order to freeze is easily overridden in Dutch. That is, there are many kinds of information that may bring a hearer to decide against canonical word order. The examples in (2) allow OVS readings, even though they lack distinguishing morphological information.

(2) a. Welke jongens zoeken de meisjes op?
which boys look the girls up.
Also: 'Which boys do the girls look up.' (OVS)
b. De \JONGens\ zoeken de meisjes op.
the boys look the girls up
Also: 'The girls look up the boys.' (OVS)
c. De telefoonnummers zoeken de meisjes op.
the telephone numbers look the girls up
Also: 'The girls look up the telephone numbers' (OVS)

In Chapter 5, I identified factors that preferred an OVS reading over an SVO reading for each of the cases above and explain why the sentences are not frozen to SVO. In (2a), it is the difference between definiteness of the first NP and the second that triggers OVS: Wh-constituents can be regarded as indefinite and subjects are preferably not indefinite (Section 5.5.2). In (2b), the Vorfeld constituent is focussed. By reading the

sentence as OVS, one avoids focussing the subject (Section 5.5.4). In (2c), SVO would make the subject inanimate and the direct object animate, which is a marked situation (Section 5.5.5). In each case, hearer preferences in the relation between grammatical function and constituent properties explain the availability of a reading that deviates from canonical argument order. The bidirectional model of word order of Chapter 5 predicts that when there is no information against SVO available, word order is frozen to canonical argument order.

The goal of this chapter is to see whether we can supplement the intuition data that supports the bidirectional model of word order with corpus data. This would give the claim that Dutch shows freezing, and that a model of word order should therefore be bidirectional, much needed empirical support. I will look for this support by investigating whether we can observe word order freezing in the direct object fronting data. In principle, the bidirectional model of word order could tell us when direct object fronting is allowed, and when it is not allowed because of freezing. A corpus investigation of the effect of word order freezing on direct object fronting would then involve locating these circumstances. There should be no instances of direct object fronting in these circumstances. In any other case, when there is *any* information present that non-canonical argument order is intended, direct object fronting should go through undisturbed.

I indicated in Section 5.1.3 that this direct interpretation of the bidirectional model leads to a highly unlikely and even impractical hypothesis to evaluate on a corpus. The hypothesis is unlikely because, as we have seen in this dissertation so far, there are few things about the variation in Vorfeld occupation that involve black-and-white contrasts. Rather, factors like definiteness cause gradual differences, slightly raising or lowering the likelihood of a certain word order. The data contains many tendencies, but hardly any clear distinctions. Moreover, the prediction is impractical because we do not know which information exactly needs to be absent to trigger freezing, nor do we know how to measure some of the bits of information that we have established as being important. It is relatively straightforward to measure, say, definiteness, animacy, case and agreement. But establishing the relevant features of prosody or, especially, context is hard even on an individual example. To do this for a larger amount of data is not realistic. One might even argue that in real data there is almost always some disambiguating information present in the context, which would mean that freezing is basically never observed.

Therefore, I will investigate word order freezing on the basis of a quantitative interpretation of the bidirectional model of word order. The model predicts word order freezing when there is no information besides word order to guide grammatical function assignment by the hearer. The hypothesis that I will investigate in a corpus is that the probability with which we observe a deviation from canonical subject-object order is dependent on the *amount* of information available to guide grammatical function assignment. For instance, if the subject is animate and the object is not, the hearer can correctly identify

the subject by following the preference for animate subjects. In this case the speaker has a better chance of getting his message across with a non-canonical word order than when the subject is inanimate or the direct object is animate. We thus predict that non-canonical word order is more frequent than when the subject is inanimate, or the object is animate. This approach has the advantage that one can try to estimate the influence of one hearer preference independently from any other hearer preference. For instance, we will not be able to investigate the relation between intonation and the frequency of direct object fronting that is suggested by example (2b). The existing annotation in the corpus does not allow us to do so in a reasonable time frame. However, this does not stop us from investigating the influence of animacy and definiteness. Each hearer preference individually contributes to the amount information that helps the hearer.

In the next two sections I will investigate three freezing hypotheses. The first tries to explain the fact that personal pronoun subjects front infrequently from the fact that they are highly recognizable subjects. I will contrast this hypothesis with the explanation that we gave the behaviour of subject pronouns in Chapter 4. In this chapter, I concluded that it was the incompatibility between the Vorfeld as a place for important information, and personal pronouns in general as unimportant information that caused this behaviour. The second and third freezing hypotheses are derived more directly from the bidirectional model. The second freezing hypothesis says that the frequency of direct-object–before–subject depends on whether the preference for highly definite subjects correctly guides grammatical function assignment. The third freezing hypothesis says the same for animacy.

The corpus investigations of Chapter 4 looked at constituents individually. We saw how properties of one constituent related to the chance that that constituent was place in the Vorfeld. This approach will not do if we want to investigate freezing. If we want to know whether a hearer can rely on animacy to correctly recognize the subject, we have to know the animacy of the subject NP as well as the animacy of the object NP. Only when the subject is animate and the object is not will the preference for animate subjects lead the hearer to the actual subject. If the subject NP and object NP are of equal animacy, the preference for an animate subject favour both equally. If the object NP is animate and the subject NP is inanimate, the preference for an animate subject will even make the hearer draw the wrong conclusion. It is *relative* animacy that we need to determine whether animacy is actually of any help to the hearer. In this chapter, I will therefore look at sentences, represented as pairs of subjects and direct objects. The findings of Chapter 4 are still relevant here, however, because we are interested in knowing whether the relative properties tell us anything about the fronting behaviour of subjects and direct objects over and above the properties of the constituent alone.

Table 6.1: Definiteness of subject vs direct object

	Object Def.			
Subject Def.	ind. full NP	def. full NP	pronoun	Total
indefinite full NP	162	88	114	364
definite full NP	664	477	373	1 514
pronoun	5 421	2 875	5 972	14 268
Total	6 247	3 440	6 459	16 146

Note: The lighter highlighted cells are cases of definiteness superiority, the darker highlighted cells are cases of definiteness inferiority.

6.2 Relative definiteness

From the dataset used for the grammatical function, definiteness, and grammatical complexity investigations of Chapter 4, we select all sentences that contain a nominal subject and a nominal direct object and determine definiteness, grammatical complexity and position for these two arguments. In total, 16 146 subject-object pairs were extracted. Not all direct objects in the dataset of Section 4.3 were included, because sentences for which the subject properties could not be reliably established were excluded. In Table 6.1, we find the distribution of subject and object definiteness.

Table 6.1 shows an interesting but complex picture. Let us begin by looking at the overall distribution of definiteness in subjects and objects. Of the objects, 40% are pronoun, 21% are definite full NPs and 39% are indefinite full NPs. This is very similar to the distribution we saw in Section 4.3. This is as expected because it is basically the same dataset. The subjects differ in definiteness from what we saw in Chapter 4. Of the subjects in transitive clauses (that is, those in Table 6.1), 88% are pronoun, 9% are definite full NP and 2% are indefinite full NP. For comparison, the overall subjects are 83% pronoun, 13% definite full NP, and 4% are indefinite full NP (counts in Table 4.10, s4.3, p109). Apparently, transitive subjects, which form about a quarter of all subjects, are more often pronouns than intransitive ones.[1]

The majority of the subjects in Table 6.1 are higher on the definiteness scale than the direct objects in the same sentence are. I will refer to this situation as *definiteness superiority*. In the table, definiteness superiority is indicated by light gray highlighting. Definiteness superiority holds in 55% of the sentences. Approximately 41% of the sentences have a subject and object that are equally high on the definiteness scale. Only

[1]Comparison of the estimated odds of being indefinite, definite or pronominal in the transitive and intransitive subject data shows this, too. The odds of being an indefinite or definite full NP are lower in the transitive data, the odds of being a pronoun are higher (OR ind.= .42, OR def.=0.64 , OR pron.=1.78 , all ORs $\frac{\text{transitive}}{\text{intransitive}}$)

4% of the sentences are cases of *definiteness inferiority*, in which the object is higher on the definiteness scale than the subject. Definiteness inferiority is highlighted with a darker gray background in the table. Definiteness superiority and inferiority will be relevant in the investigation of the second freezing hypothesis in Section 6.2.2. In cases of definiteness superiority, a hearer can correctly identify the subject by selecting the more definite NP. Definiteness inferiority means that the hearer would incorrectly select the intended object as the subject. In the bidirectional model, we predict that this has an effect on word order freedom.

Comparison of the observed counts in Table 6.1 with the counts expected under non-association of subject and object definiteness suggests that subject and object definiteness are associated: Subjects and objects have a tendency to be both pronouns or both full NPs. Two combinations of subject and object definiteness stand out in this pattern. First, the combination of definite full NP subject and definite full NP object occurs more often than expected on the basis of non-association (observed 477, expected $\frac{3440 \times 1514}{16146} = 323$). Secondly, the combination of definite full NP subject and pronoun object is observed less often than expected under non-association (observed 373, expected $\frac{6459 \times 1514}{16146} = 606$).

The bidirectional model of word order presented in Chapter 5 predicts that whether the direct object is allowed to precede the subject depends on the availability of information about grammatical function assignment. In order to study the interaction between subject and object definiteness and word order, we will divide the data of Table 6.1 according to whether the subject precedes the object or not. If we take Vorfeld occupation into account in this division, there are four logically possible configurations, respectively:

SVO: The subject occupies the Vorfeld, the direct object is in the postverbal domain. The subject precedes the object.

XVSO: The subject and direct object are both in the postverbal domain. The subject precedes the object.

OVS: The direct object occupies the Vorfeld, the subject is in the postverbal domain. The object precedes the subject.

XVOS: The direct object and subject are both in the postverbal domain. The object precedes the subject.

We expect that the XVOS construction is rare in the corpus. In the Dutch Mittelfeld, generally only indirect objects, and reflexive objects of certain verbs are allowed to precede the subject (see Section 2.6.1). Indeed, in the dataset, we find only three examples of XVOS. Two of those might be considered to involve indirect objects rather than direct objects, because they contain an object experiencer. One of the two cases of an object experiencer that precedes the subject in the Mittelfeld is given in (3). A shortened version of the third case of XVOS is given in (4). The subjects are boldfaced. The number of

missing words in (4) is given within brackets. The translation is of the complete sentence, with the left out material between square brackets.

(3) op den duur na drie uur overvalt je **een zekere verveling toch**
after a while after three hours overcomes you.SG a certain boredom PART
'After three hours, one gets bored.' (Vl-a 400401:59)

(4) in 't hele verhaal speelt een grote rol **de brief die [+7w] waarin [+14w].**
in the whole story plays a big role the letter that in which
'The letter that [Willem-Alexander supposedly wrote four years ago], in which [he promises the NOC-NSF not to become a member of the International Olympic Committee], plays an important role in the events.' (Nl-f 7150:65)

The example in (4) contains an extremely heavy subject (25 words in total, two dependent relative clauses), which appears to the right of the object *een grote rol* in sentence final position. This order could be explained by the tendency of heavy constituents to appear at the right periphery (Section 4.4). I will ignore these 3 cases of XVOS in the data, and focus on SVO, OVS and XVSO. This has the advantage that we can equate object-precedes-subject with object topicalization. This will simplify the investigation of word order freezing as well as the comparison of the results in this chapter with the results of Chapter 4.

In Table 6.2, p248, we find the definiteness distributions of subject and object, across the three configurations SVO, XVSO and OVS. Table 6.2 must be understood as the data of Table 6.1 cut in three slices according to configuration. The percentages in italics behind each count in Table 6.2 refer to the proportion of sentences that are of the configuration for which the counts are reported. Summing the percentages of corresponding cells between slices gives 100%. For instance, sentences with indefinite subjects and indefinite objects are 83% SVO (top left cell in the SVO-slice), 15% XVSO (top left cell in the XVSO-slice) and 1% OVS (top left cell in the OVS-slice).

The findings of Section 4.3 with respect to definiteness can be recognized in Table 6.2 by looking at the totals. The proportion of direct objects that appears in the Vorfeld rises when the objects are higher on the definiteness scale. The bottom row in the table shows that 3% of indefinite full NP direct objects are OVS, 9% of definite full NP direct objects are OVS, and 42% of pronoun direct objects. On average 20% of the direct objects appear in an OVS sentence, that is, 20% of the direct objects occupies the Vorfeld. For subjects, we see the initial rise in Vorfeld occupation when we move from indefinite full NPs (75%, right-hand side SVO-slice) to definite full NPs (79%). Pronoun subjects only appear in the Vorfeld in 65% of the sentences. I will discuss the behaviour of pronoun subjects in more detail in Section 6.2.1. For now, we can observe that this fall is more pronounced than in the overall data in Section 4.3 – overall 74% of definite full NP subjects occupied the Vorfeld and 70% of pronoun subjects. Inspection of the transitive

Table 6.2: Subject and object definiteness per configuration

		Object Definiteness							
Conf.	Subject Def.	indefinite	%	definite	%	pronoun	%	Total	%
svo	ind. full NP	135	*83*	63	*72*	74	*65*	272	*75*
	def. full NP	537	*81*	399	*84*	258	*69*	1 194	*79*
	pronoun	3 798	*70*	1 805	*63*	3 398	*57*	9 001	*63*
	Total	4 470	*72*	2 267	*66*	3 730	*58*	10 467	*65*
xvso	ind. full NP	25	*15*	24	*27*	1	*1*	50	*14*
	def. full NP	113	*17*	69	*14*	2	*1*	184	*12*
	pronoun	1 452	*27*	769	*27*	33	*1*	2 254	*16*
	Total	1 590	*25*	862	*25*	36	*1*	2 488	*15*
ovs	ind. full NP	2	*1*	1	*1*	37	*32*	40	*11*
	def. full NP	13	*2*	9	*2*	113	*30*	135	*9*
	pronoun	171	*3*	300	*10*	2 541	*43*	3 012	*21*
	Total	186	*3*	310	*9*	2 691	*42*	3 187	*20*

Note: The percentages in italics refer to proportion across sentence type.

pronoun subject data of Table 6.2 shows that the strong drop in Vorfeld occupation can be related to an extremely high proportion of personal pronouns. Of the transitive pronoun subjects, 93% are personal pronoun. From Chapter 4 we know that personal pronouns in any grammatical function are dispreferred as Vorfeld occupants.

A striking trend in Table 6.2 is the almost complementary nature of the XVSO and OVS. When the object is a full NP, XVSO accounts for the majority of non-subject-initial cases. When the object is an indefinite full NP, $100\% - 72\% = 28\%$ of the sentences are not subject-initial, of which 25% are XVSO. When the object is a definite full NP, 34% of the cases are not SVO, of which 25% are XVSO. Even though there is a rise of OVS cases, XVSO is still the majority. However, when the object is pronominal, there are virtually no cases of XVSO (1%), even though the percentage of non-SVO sentences rises considerably to 42%. Briefly put, when the subject is not in the Vorfeld, it depends on the pronominality of the object whether it or another constituent occupies the Vorfeld. We can also turn this around: When the object is highest on the definiteness-scale, it is typically such a good Vorfeld candidate that only subjects can prevent it from occupying the Vorfeld.

After these general observations about the relation between subject definiteness and object definiteness (relative definiteness), and the relation between relative definiteness and the position of subject and object in the sentence, it is time to investigate two hypotheses about word order freezing in the relative definiteness data.

6.2.1 Pronominal subjects and object placement

An interesting observation can be made about the data in Table 6.2: Direct object fronting becomes more frequent when the *subject* is a pronoun. The proportion of objects that appears in the Vorfeld is low when the subject is an indefinite or definite full NP, respectively 11% and 9% (right-hand side margins OVS-slice). However, when the subject is pronominal, 21% of the direct objects appears in the Vorfeld. This is not just due to the association between pronominal subjects and pronominal objects. As well as in the average case, the trend is observed when the direct object is pronominal (right column OVS-slice, from 32%/30% to 43%) and when it is a definite full NP (middle column OVS-slice, from 1%/2% to 10%).

Now let us consider two hypotheses about the relation between definiteness of the subject and the position of the direct objects. The first hypothesis is based on our findings of Chapter 2 and 4, and concerns the behaviour of personal pronouns. It was found that personal pronouns in general did not like to appear in the Vorfeld. This was observed as a relatively low proportion of Vorfeld occupation by personal pronouns across the grammatical functions. This behaviour of personal pronouns was expected under the hypothesis that the Vorfeld is a position for important (new, contrastive, unexpected) material, cf. Gundel's (1988) *first-things-first* principle. If personal pronoun subjects shy away from the Vorfeld, that is, they occupy the Vorfeld relatively infrequently, then the proportion of other material in the Vorfeld must go up by necessity. There is exactly one Vorfeld constituent, and if it is not the subject, it must be something else. This relates to the data in Table 6.2 in the following way. As noted before, the overwhelming majority of the pronoun subjects in the table are personal pronouns (93%). Thus, we should see a strong decrease in Vorfeld occupation by pronoun subjects compared to full NPs (74%/79% to 63%). This gap has to be filled by constituents of other functions. On the assumption that this is more or less equally spread out over the other constituents, the proportion of direct objects that occupies the Vorfeld goes up. Let use call this explanation the *unimportance hypothesis*, since the explanation of the pattern in the direct object data is based on the behaviour of pronoun subjects as material that does not meet the (soft) requirement that Vorfeld material be important.

On the basis of the insights with respect to word order freezing gained in Chapter 5, we can also formulate a second possible explanation for the effect that subject pronominality has on direct object fronting. Recall from the previous chapter that word order freezing

was inversely dependent on the recognizability of the arguments. If the subject is recognizable as a subject on independent grounds, word order can be used for other purposes than marking subjecthood. Personal pronouns are highly recognizable as subjects: They are high on the definiteness scale (Section 5.5.2), they may show case, and they can be expected be part of the information structural background (Section 5.5.4) since they are not likely to realize new or contrastive material. When the subject NP does not have all these good subject properties, putting an object before it – that is, in the Vorfeld – may lead to misinterpretation. Thus, when the subject is not a personal pronoun, more of the sentences are frozen. However, when the subject is a personal pronoun, a speaker is free(r) to give in to the desire to put a direct object in the Vorfeld. Consequently, the proportion of direct objects that appears in the Vorfeld is higher when the subject is a personal pronoun. Briefly formulated, this second possible explanation of the behaviour of subject pronouns is as in (5).

(5) *First freezing hypothesis* Pronominal subjects are preceded by direct objects more often than other subjects because they are highly recognizable as subjects independent of word order.

Both hypotheses start from the assumption that the effect of subject pronominality on direct object fronting is due to personal pronoun subjects, which make up the majority of pronominal subjects in the transitive data. What distinguishes the unimportance hypothesis from the first freezing hypothesis is that the unimportance hypothesis can be formulated in terms of speaker preferences alone. The Vorfeld is best occupied by important material and if a speaker were to place a personal pronoun in the Vorfeld, he would go against this preference. The first freezing hypothesis crucially refers to hearer preferences: A speaker does not need to rely on canonical word order as much to get the intended grammatical function assignment across, because the subject is such that it is very likely to recognized as the subject by a hearer.

Now let us look at the data. The freezing hypothesis correctly predicts the rise in OVS that is seen in Table 6.2. However, the freezing hypothesis does not predict that other constituents also front more often when the subject is a personal pronoun. After all, word order is still a good cue for correct grammatical function assignment if we put a non-object constituent in the Vorfeld, but maintain subject-before-object order. The freezing hypothesis only predicts that material that can be mistaken for a subject should not precede a poor subject. Yet, in Table 6.2, we can see that the effect of subject pronominality is also visible in the XVSO slice. When the subject is an indefinite or definite full NP, 14% respectively 12% of the sentences are XVSO. When the subject is a pronoun, 16% of the sentences are XVSO. The effect is less pronounced than in the OVS data. Recall from the general discussion of the relative definiteness data that there is a positive correlation between subject and object pronominality, and that pronoun objects

hardly occur in XVSO. As a result, the proportion of XVSO when the subject is a pronoun is pushed down. This may explain the difference in effect of subject pronominality on OVS and XVSO.

In addition, we can also find the effect of subject pronominality in subject data that is not part of the transitive data in Table 6.1. In this data (not tabulated), $\frac{4512}{6205} = 73\%$ of definite full NPs occur in the Vorfeld. For *personal* pronouns, this proportion lies lower at $\frac{16596}{25357} = 66\%$. Again, the effect is less pronounced than in the object data, but clearly visible.

In neither the transitive XVSO data nor the 'intransitive' subject data does object-before-subject word order result when the subject moves out of the Vorfeld. Therefore, the freezing hypothesis says nothing about the behaviour of personal pronoun subjects in these cases. In contrast, the unimportance hypothesis predicts that personal pronoun subjects in any context have a preference for not appearing in the Vorfeld. The fact that objects and any other material front more often when the subject is a personal pronoun is thus explained by the unimportance hypothesis.

I conclude that the unimportance hypothesis is to be preferred over the freezing hypothesis as an explanation of the effect of personal pronoun subjects on object fronting. The unimportance hypothesis explains that we see similar effects of subject pronominality on other constituents, whereas the explanation offered by the freezing hypothesis is limited to object fronting alone.

6.2.2 Relative definiteness and object placement

In the bidirectional OT model of word order presented in the previous chapter, I proposed that the interpretation of wh-questions motivates the adoption of constraints that link subjecthood to high definiteness (Section 5.5.2). These constraints also apply to declarative sentences: They predict that a subject is recognizable when it is higher on the definiteness scale than the object is (definiteness superiority). Recognizability of subject and object can be linked to word order variation. In particular, if the subject is recognizable because it is higher on the definiteness scale than the object, the object is free to be placed in front of the subject. Under the quantitative interpretation of the bidirectional model that I explained in Section 6.1, we predict that non-canonical word order is correlated with the amount of information that guides the hearer towards the correct grammatical function assignment. In the case of relative definiteness, we thus predict that direct object fronting is more frequent when the subject is higher on the definiteness scale than the direct object is.

A quantitative interpretation also makes sense in the opposite situation. When the subject is below the direct object on the definiteness scale (definiteness inferiority), selecting the NP that is highest on the definiteness scale as the subject will give an incorrect result. In a sentence with definiteness inferiority, relative definiteness is misleading information

to the hearer. I speculate that in this case, word order is more restricted than when the subject is not inferior. To be precise, we predict that direct object fronting is less frequent when the subject is lower on the definiteness scale than the direct object is.

This means that the bidirectional model of word order freezing, under the gradient interpretation that I explained in the introduction to this and the previous chapter, connects relative definiteness to direct object fronting at three levels. The *second freezing hypothesis* is summarized as (6).

(6) *Second freezing hypothesis* Direct object fronting is influenced by relative definiteness as indicated in the following schema.

Subject > Object	Subject = Object	Subject < Object
Superiority		Inferiority
Object fronting more frequent	No effect	Object fronting less frequent

In the discussion of Table 6.1, p245, I noted that definiteness superiority is observed in 55% of the subject-object pairs (light gray in the table), whereas definiteness inferiority is observed in only 4% (darker gray). As a first stab at investigating the second freezing hypothesis, we can study the relative frequency (as an estimate of probability) of direct object fronting in those 55%, and in the 4%, and compare this frequency to the relative frequency of direct object fronting in the remaining 41% where the subject and direct object are at the same definiteness level. We can calculate the numbers needed from Table 6.2.

The relative frequency of object fronting in each case is the average proportion of OVS in the areas corresponding to each case in Table 6.2. The relative frequency of OVS in the neutral case (no inferiority, no superiority) is the average proportion of OVS on the top-left–bottom-right diagonal: $\frac{2552}{6611} = 39\%$. The relative frequency of OVS in the case of definiteness superiority is the proportion of OVS in the three bottom-left cells: $\frac{484}{8960} = 5\%$. The relative frequency of OVS in the case of definiteness inferiority is the proportion of OVS in the three top-right cells: $\frac{151}{575} = 26\%$.

If we compare the relative frequencies of OVS in the three cases, we would have to conclude that the second freezing hypothesis is not even near the truth. The relative frequency of OVS in the definiteness inferiority data is lower than in the neutral data (26% vs 39%). This is as is predicted by the second freezing hypothesis. However, the relative frequency of OVS in the definiteness superiority group is also lower than in the neutral group. In fact, it is a factor of eight lower (5% vs 40%). The relative frequency of OVS is even lower in the definiteness superiority data than in the inferiority data.

This way of investigating the second freezing hypothesis ignores the fact that there are other factors that shape the data in Table 6.2. For instance, we know from Chapter 4 that indefinite full NP direct objects do not front frequently. This trend is also clear in Table 6.2: Only 3% of the indefinite full NP objects appear in OVS (first column, bottom

margin OVS-slice). If we combine that with the fact that by far most of the subjects are definite full NPs or pronouns and, thus, higher than indefinite full NPs, it becomes clear that the very low 5% relative frequency of OVS in definiteness superiority sentences is a reflection of the behaviour of indefinite full NP direct objects. Likewise, the relatively high 40% relative frequency of OVS in the neutral case results from the fact that the definiteness combination with the highest proportion of OVS falls in the neutral case. When both subject and object are pronouns, the relative frequency of object fronting is 43%. And in turn, the relative frequency of object fronting when both the subject and object are pronouns is probably related to the fact that in this group the subjects tend to be personal pronouns (93% personal pronoun), whereas the objects tend to be demonstrative pronouns (60% demonstrative pronoun). In Chapter 4, we learnt that personal pronouns do not like to appear in the Vorfeld (cf. the unimportance hypothesis), but that demonstrative pronouns front very frequently. All in all, before we can estimate the effect of relative definiteness on direct object fronting, we have to control for the average behaviour of subjects and objects of different definiteness levels and/or pronominal forms.[2]

More generally, we should investigate whether relative definiteness per the second freezing hypothesis adds anything to what we learnt about subject- and object-fronting in Chapter 4. In that chapter, we saw that grammatical function, definiteness or NP form, and grammatical complexity each influenced Vorfeld occupation in their own, possibly complex, way. Thus, the findings of that chapter tell us what makes a constituent attractive as a Vorfeld occupant to the speaker. In this chapter, we want to know whether a speaker, in addition to being sensitive to these constituent properties, is also sensitive to the effect that a certain word order may have on the hearer. Our investigation of the second freezing hypothesis should take as much as possible of everything we know about subject- and object-fronting into account.

A good way to see whether relative definiteness explains more about object fronting than the results of Chapter 4 is to fit a logistic regression model of the transitive data. The model predicts for each sentence whether it is OVS or not, which is the same as predicting whether the object precedes the subject or not. The model incorporates all knowledge from Chapter 4 by building on the logistic regression model of subject, direct object, and indirect object fronting in Section 4.5, Model 1. Model 1 predicted the chance that an argument constituent appears in the Vorfeld on the basis of grammatical function,

[2]The discussion in this paragraph is reminiscent of the discussion in Chapter 5 of the predictions of the bidirectional OT grammar in (78), p213. Under this grammar, it was incorrectly predicted that a sentence with an initial indefinite full NP and a postverbal definite full NP (also) receives an OVS interpretation. This interpretation was predicted by the preference for definite subjects. In my discussion of the example, I pointed out that the model did not take into account that indefinite full NP direct objects were not very likely to appear in the Vorfeld to begin with and that a better model should contain constraints that better capture the speaker preferences with respect to Vorfeld occupation that were found in Chapter 4. Looking at raw percentages of object fronting in definiteness superiority or inferiority is making the same mistake.

NP form and grammatical complexity, and parameters that corrected for the influence of NP form and grammatical complexity on subjects. Recall that the six-level variable NP form contains information about definiteness, as well as pronominal form and that NP form also distinguishes bare nouns and proper names from other indefinite and definite full NPs (Section 4.3.1). The distinction between the pronominal forms personal and demonstrative is relevant because the two differ in whether they realize important material, and therefore show very different Vorfeld behaviour. The distinction between bare nouns and indefinite determiner NPs is relevant because we concluded in Section 4.3.1 that many of the bare nouns are arguably definite full NPs. Grammatical complexity is measured as the (natural logarithm of) the number of words in a constituent (Section 4.4). Because of an association between complex constituents and the right periphery, less complex material has a better chance of appearing in the Vorfeld.

Model 2 predicts for a sentence (a subject-object pair) whether the sentence is OVS on the basis of properties of the direct object, properties of the subject, and relative definiteness. The fact that we have separated the properties of the subject and direct object means that we have no need for the factor grammatical function, or the factors that adjusted the influence of NP form and grammatical complexity on subjects. The inclusion of subject properties to predict OVS is motivated by the observation that the chance that the subject itself occupies the Vorfeld may influence the chances for the direct object, as became clear in the discussion of the unimportance hypothesis of the previous subsection. Note however that when the direct object is postverbal (not-OVS), the subject may be in the Vorfeld (SVO) or in the postverbal domain (XVSO). A low chance of subject fronting does therefore not translate directly to a high chance of direct-object fronting.

Relative definiteness is a three-valued variable: superior, neutral, inferior. I will consider neutral as the base level, so that the second freezing hypothesis predicts that the level superior is associated with an increase in OVS, and the level inferior is associated with a decrease of OVS. The levels are defined on the three-valued definiteness scale.

The result of fitting Model 2 to 16 146 SVO, XVSO and OVS sentences is given in Table 6.3. The model is a good predictor of OVS (c-index $= 0.927$) and shows no signs of overfitting. The influence of the factor Object NP Form is as we have seen in Chapter 4: Definite full NPs (definite determiner and proper name) are more likely to appear in the Vorfeld than indefinite full NPs (indefinite determiner and bare noun) are. Demonstrative pronouns front very often, but personal pronouns show a strong tendency of avoiding the Vorfeld. The influence of Complexity on object-fronting is also as expected: Longer direct objects have a lower chance of occupying the Vorfeld.

The combined information about Subject NP Form significantly improves the model of direct object fronting ($G^2 = 236.1$, df $= 5$, $p < .001$). Table 6.3 shows that the effect of Subject NP Form on direct object fronting is limited to Subject NP Form=demonstrative pronoun. When the subject is a demonstrative pronoun, it makes such a good Vorfeld

Table 6.3: Model 2. Predicting OVS with relative definiteness.

Parameter		Estimate	OR (lo–hi)		p
α		-4.729			
Subject Complexity		0.083	0.75	1.57	.653
Object Complexity		**-0.721**	0.40	0.58	$<$ .001
Subject NP Form	indefinite determiner *(base level)*				
	bare nominal	-0.722	0.17	1.37	.171
	definite	-0.220	0.43	1.49	.480
	proper name	-0.450	0.32	1.24	.184
	demonstrative pronoun	**-2.498**	0.03	0.23	$<$ .001
	personal pronoun	0.295	0.56	3.19	.503
Object NP Form	indefinite determiner *(base level)*				
	bare nominal	**1.063**	2.12	3.94	$<$ .001
	definite	**1.742**	4.41	7.38	$<$ .001
	proper name	**1.949**	4.83	10.19	$<$ .001
	demonstrative pronoun	**5.459**	101.96	541.06	$<$ .001
	personal pronoun	**-2.228**	0.02	0.44	.002
Relative Def.	superior	**1.214**	1.55	7.31	.002
	neutral *(base level)*				
	inferior	-0.437	0.35	1.200	.163

Note: The parameter estimates in boldface indicate significant contribution of the parameter according to Wald's test. The Wald's test p-values are reported in the last column. In the OR column, the 95% confidence intervals are given for the estimated odds ratios. See Section 3.5 for more explanation.

occupant that the chances of OVS are a least lowered by a factor of 0.23 (upper OR confidence limit). There is no significant effect of Subject Complexity on object fronting ($G^2 = 0.2$, df $= 1$, $p = 0.65$, see also the results of the Wald's test in the table).

The negative influence that demonstrative pronoun subjects have on object-fronting was not captured by Model 1, since Model 1 only used properties of the constituents themselves. Although a model like Model 1 may give a good approximation (many of the parameter estimates are similar), ultimately a better understanding of Vorfeld occupation has to take into account that there can be several contenders for the Vorfeld position in a sentence. It can not be excluded that a constituent that would make a very good Vorfeld occupant is not placed into the Vorfeld simply because there is a better candidate.

Now let us turn to the predictions made by the second freezing hypothesis. This hypothesis said that when a subject is above the object on the definiteness scale, object fronting should be more frequent, and when the subject is below the object, fronting

should be less frequent. This means that in Model 2, the factor Relative Definiteness should contribute in a specific way: The value superior significantly increases the chance of OVS, the value inferior decreases it. The whole factor Relative Definiteness increases model fit significantly ($G^2 = 11.1$, df $= 2$, $p = .004$). Looking at the parameter estimates of Model 2 highlighted in Table 6.3, we can see that the parameter estimates are in the direction that were predicted by the freezing hypothesis: Relative Definiteness=superior is positive, Relative Definiteness=inferior is negative. However, Wald's testing shows that only the estimate for superiority is significant. When the subject is above the object on the definiteness scale, the chance that the object appears in the Vorfeld increases by at least a factor 1.5 (OR between 1.55 and 7.31). The opposite trend cannot be established: Even though the estimate is negative, we have no evidence that it is significantly different from zero. The odds ratio confidence interval reflects this as it contains 1. This means that it cannot be excluded that the odds of OVS between inferior and neutral differ by a factor of 1, that is, that they are really the same.

What does this finding mean for the second freezing hypothesis? I would like to argue that the second freezing hypothesis can be considered confirmed, although perhaps not to the highest degree. The moderately sized positive effect of definiteness superiority provides support for the freezing hypothesis. Although no significant negative effect of definiteness inferiority was found, no significant positive effect was found either. So, even though this part of the prediction is not borne out, there is no evidence that the hypothesis is completely on the wrong track, either. Moreover, we do not have another hypothesis that would straightforwardly predict the positive effect of superiority. The second freezing hypothesis can be accepted. The goal in this chapter to find additional empirical support for word order freezing and the bidirectional model of word order is achieved. We have found corpus evidence that word order freezing in Dutch is a real phenomenon, that is not only observed in specially crafted isolated sentences without intonation, but also in a corpus of naturally occurring spoken language.

6.3 Relative animacy and object placement

In Chapter 5, I argued for a bidirectional OT model that used constraints on definiteness and grammatical function to capture the effect of subject recognizability on word order freezing. In the previous section, we have seen that a quantitative interpretation of the bidirectional model's predictions with regards to relative definiteness and object-fronting is confirmed in the corpus of spoken Dutch CGN. In Chapter 5, I also argued that we should adopt constraints on animacy and grammatical function that are like the constraints on definiteness and grammatical function. In a bidirectional model, constraints that prefer subjects to be animate (Aissen, 1999) capture the observation that animacy is used by a Dutch hearer in grammatical function assignment. When two NPs of unequal animacy

are considered for the function of subject, a hearer will prefer to assign subject function to the NP that is highest on the animacy scale. This means that a hearer can use animacy to correctly identify the subject when the subject is above the object on the animacy scale. When the subject is below the object on the animacy scale, using animacy to select the subject gives the wrong result. If it is true that the chance a speaker deviates from canonical argument order is dependent on the amount of information the hearer has in grammatical function assignment, then we should expect to see an effect of relative animacy on object fronting.

I will formulate the expectation about the effect of *relative animacy* as the third freezing hypothesis, modelled after the second freezing hypothesis.

(7) *Third freezing hypothesis* Direct object fronting is influenced by relative animacy as indicated in the following schema.

Subject > Object	Subject = Object	Subject < Object
Superiority		Inferiority
Object fronting more frequent	No effect	Object fronting less frequent

Unfortunately, we shall see that investigating the third freezing hypothesis is not as straightforward as investigating the previous two was.

In order to investigate the freezing hypothesis, I manually annotated a random sample of 2 345 sentences from the dataset used in the previous section for animacy. The animacy information was added to the existing CGN syntactic annotation, so that we have access to combined animacy, definiteness and word order information. For reasons of time, reliability of encoding, and size of the dataset, an extremely simple annotation scheme was used, consisting of just two categories. Human and animal referents, as well as organizations, countries, collectives and companies, were classified as *animate*; concrete objects such machines and body parts, and 'material masses' like air, wind, water and land were classified as *inanimate*, as well as abstract objects, like propositions, utterances, times, places and amounts of money. The scheme can be considered a dichotomization of the scheme described in Zaenen et al. (2004). Given the large proportion of personal and demonstrative pronouns in the corpus, it is important to note that anaphors were assumed to be resolved, and were classified according to their interpretation.

The annotation revealed a distribution of animacy across subject and object as given in Table 6.4, p258. Direct objects are overwhelmingly inanimate ($\frac{2102}{2345} = 90\%$), whereas the great majority of subjects is animate ($\frac{2262}{2345} = 96\%$). As a result, we can observe that the subject is higher in animacy than the object about 86% of the sentences, of equal animacy in about 13% of the cases, and lower in animacy in less than 1%. Animacy superiority and inferiority are highlighted in the table. There is no evidence for a correlation between subject animacy and object animacy (OR: 1.08, $p = 1$, 2-t Fisher's). These results are in line with results for Norwegian (Øvrelid, 2004) and Swedish and English (Dahl and Fraurud, 1996; Zeevat and Jäger, 2002).

Table 6.4: Animacy of subject vs direct object

	Object Animacy		
Subject Animacy	inanimate	animate	Total
inanimate	75	8	83
animate	2027	235	2262
Total	2102	243	2345

Note: The lighter highlighted cell contains cases of animacy superiority, the darker highlighted cell cases of animacy inferiority.

This particular distribution of animacy is probably highly genre specific. In the discussion of the effect of definiteness on Vorfeld occupation in Section 4.3, we observed that most subjects are personal pronouns, and that a good portion of these pronouns are 1st and 2nd person. Even though we theoretically (Aissen) and empirically (the corpus studies cited above) expect subjects to be animate and objects to be inanimate, the degree of subject animacy is extreme. This skew in the animacy distribution spells trouble for our investigation. The fact that 86% of the sentences show animacy superiority does not leave us with much material to compare with.

Inspection of the 8 cases of animacy inferiority reveals that they are all SVO. The examples typically have a demonstrative subject in the Vorfeld (8a), but (8b) is also observed.

(8) a. dat zal jullie niet verbazen
this will you.PL not surprise
'This will not come as a surprise.' (Nl-h 9128:40)

b. een grotere school trok me wel wat aan
a bigger school attracted me somewhat VPART
'I kind of liked the idea of a bigger school.' (Nl-b 86:26)

One could argue that these examples do not involve direct objects, but indirect objects, since they feature an object experiencer and a psych-verb. Sentence (9) illustrates that animacy inferiority is possible in a Dutch sentence without a psych-verb.

(9) dat mag ons niet stoppen
this may us not stop
'This can't stop us.' (Vl-f 600373:72)

The fact that only two examples like (9) are found in a sample of 2 345 does suggest that it is highly unusual.

Table 6.5: Subject and object animacy per configuration.

		Object Animacy					
Conf.	Subject Anim.	inanimate	%	animate	%	Total	%
svo	inanimate	64	*85*	8	*100*	72	*87*
	animate	1273	*63*	181	*77*	1454	*64*
	Total	1337	*64*	189	*78*	1526	*65*
xvso	inanimate	9	*12*	0	*0*	9	*10*
	animate	320	*16*	31	*13*	351	*16*
	Total	329	*16*	31	*13*	360	*15*
ovs	inanimate	2	*2*	0	*0*	2	*2*
	animate	434	*21*	23	*9*	457	*20*
	Total	436	*20*	23	*9*	459	*20*

Note: The percentages in italics refer to proportion across sentence type.

The lack of OVS for object experiencer verbs was also observed for Norwegian (Øvrelid, 2004). Note that this would follow directly from a model of word order that takes the hearer perspective in consideration like the bidirectional model of word order does. A sentence that is uttered as O[anim]VS[inan] is likely to be misunderstood as SVO. Both the preference for canonical argument order, and the preferences for animate subjects and inanimate objects point towards an SVO interpretation. This analysis does rest on the assumption that linking subjecthood to animacy is the correct way of characterizing the preferences of a hearer. If a hearer does not care about animate subjects in general, but is just interested in fulfilling the selection criteria of the verb, there should not be a problem with O[anim]VS[inan] when the main verb is a psych-verb. The verb will demand an object that is capable of experiencing, so that the inanimate second NP makes a poor object.

As mentioned, most of the examples of animacy inferiority involve a demonstrative subject and a personal pronominal object, which would mean that we can predict data like (8) and (9) from the properties of subjects and objects alone. Demonstrative subjects hardly appear outside of the Vorfeld and personal pronoun objects hardly appear in the Vorfeld. We can almost be certain that a sentence that features both is SVO. This way, we also explain why there are no XVSO cases in the animacy inferiority data. Although it is not certain that the lack of XVSO is not purely a matter of chance.

The distribution of animacy across configuration is given in Table 6.5. We have already seen that the 8 cases where the subject is inanimate and the object animate are all SVO. The data in the table is modeled well by a log-linear model that only allows interaction

between subject animacy and configuration and object animacy and configuration ($G^2 = 1.97, df = 3, p = 0.57$). This means that on the basis of this dataset we have no reason to assume that there is interaction between object and subject animacy in any sentence type. The resulting log-linear model also indicates that animate subjects are involved more in OVS than the other sentence types and that animate objects occur slightly less in OVS, and more in SVO and XVSO. Put together, it means that the combination of an animate subject and inanimate object has an increased chance of being observed in a sentence where the object precedes the subject (OVS). In Table 6.5, this effect is seen in the high proportion of OVS (21%, bottom left cell, OVS-slice), compared to the much lower proportions in the other three cells (respectively, 2%, 9%, and 0%). This state of affairs would be predicted by the third freezing hypothesis: When the subject is superior to the object in animacy, the hearer can identify the subject on the basis of relative animacy and the speaker is free to move it the object in front of the subject.

However, like before, definiteness may provide an explanation for the data in Table 6.5. Inspection of the data shows that animate subjects are overwhelmingly personal pronouns, whereas inanimate subjects are more often demonstrative pronouns or definite full NPs. Animate objects are typically indefinite full NPs or personal pronouns. Inanimate objects are indefinite full NPs or demonstrative pronouns. As a result of these correlations between definiteness and animacy and between pronominal form and animacy, the effects of definiteness and pronominal form on subject and object placement may yield the observed animacy data as a side effect. Demonstrative pronouns (many inanimate) front often in either category, which leads to more inanimate objects or inanimate subjects in the Vorfeld. Personal pronouns (many animate) do not front often, which makes many animate objects and animate subjects postverbal.

In the investigation of relative definiteness, I addressed the problem of correlations in the factors by fitting a logistic regression model. I will use the same technique here but with some reservations. The combination of highly correlated factors and the complete lack of positives (OVS) for certain values of the factors (no OVS when the subject is inanimate, and the object animate) means that logistic regression analysis is problematic. I will circumnavigate this problem in two ways: I will reduce the factors as much as is reasonable on the basis of what we know now, and I will use a special fitting procedure that is especially suited for sparse data sets like ours (Heinze and Ploner, 2003).[3] Compared to Model 2, the model I will fit here will make fewer distinctions in its factors, and the estimates of the effect sizes will be large and imprecise. Even though the conclusions about which factors contribute significantly and in what direction can be trusted, I do consider the results I will present below as preliminary.

[3]The model will be fitted with *Firth's penalized maximum likelihood estimation*, provided by the `logistf`-package in *R*. See Heinze and Ploner (2003) for references, notes on the implementation and discussion of the method.

Table 6.6: Model 3. Predicting OVS with relative animacy.

Parameter		Estimate	OR (lo–hi)		p
α		-7.160			
Object Complexity		**-0.465**	0.40	0.93	.022
Subject NP Form	not demonstrative pronoun *(base level)*				
	demonstrative pronoun	**-4.326**	0.00	0.09	$<$ .001
Object NP Form	indefinite determiner *(base level)*				
	bare nominal	**1.087**	1.12	7.95	.029
	definite full NP	**2.213**	4.55	20.66	$<$ .001
	demonstrative pronoun	**6.738**	177.36	$>$1000	$<$ .001
	personal pronoun	-1.019	0.00	9.18	$<$ .568
Relative Defin.	superior	**2.029**	2.00	67.93	$<$ .001
	not superior *(base level)*				
Subject Animacy	inanimate *(base level)*				
	animate	-5.009	0.00	1.40	.061
Object Animacy	inanimate *(base level)*				
	animate	**6.508**	2.76	$>$1000	.028
Relative Animacy	superior	**6.333**	2.28	$>$1000	.031
	not superior *(base level)*				

Note: The parameter estimates in boldface indicate significant contribution. 95% confidence intervals and significance are based on *profile penalized log likelihood*. See also footnote 3, p260.

Model 3 predicts OVS on the basis of properties of the subject and object, relative definiteness, and relative animacy. The only information about the subject that is part of the model is whether it is a demonstrative or not. Model 2 showed that the other distinctions in Subject NP Form did not contribute significantly, nor did Subject Complexity. Object NP Form will make five distinctions in stead of six: Proper names and definite determiner NPs are collapsed into one level (definite full NPs) because they have an equal effect on fronting (see also Section 4.5). Relative definiteness is reduced to definiteness superiority alone, since definiteness inferiority did not contribute significantly in Model 2. In addition to these reduced factors from Model 2, Model 3 contains information about animacy of subject and object, and relative animacy. The latter will only show the distinction animacy superiority or not. This has a separate motivation: If we know the effect of subject animacy and object animacy on OVS, and we know the effect of animacy superiority on OVS, the effect of animacy inferiority on OVS is not free to vary. On a scale that makes more than two distinctions (like the definiteness scale did), it would have been possible to investigate superiority and inferiority independently.

The results of fitting Model 3 on the 2 345 SVO, XVSO and OVS sentences that have received additional annotation for animacy are given in Table 6.6, p261. The model is a good predictor of OVS (c-index = 0.932). The model has not been inspected for signs of overfitting. Most of the parameters and distinctions that were found to be significant in Model 2 are significant in Model 3, too. The effects are in the same direction. A notable finding is that the effect of relative definiteness is confirmed in spite of the smaller dataset and the additional factors. Compared to Model 2, the estimated effect sizes are large, which is an effect of removing non-significant distinctions and of the data sparseness. Some of the confidence intervals for the ORs are very large, which again can be blamed on the small dataset. I will ignore the estimates and the related intervals, and concentrate on the significance and direction of the contributions alone.

According to Model 3, animate objects front more frequently than inanimate objects. Animate subjects discourage OVS, which may be because animate subjects want to occupy the Vorfeld themselves. We may speculate that animacy of a constituent promotes Vorfeld occupancy. Investigating this highly speculative hypothesis is a topic for future research. If the third freezing hypothesis is correct, animacy superiority should lead to an increase of OVS. According to Model 3 this is indeed the case: Relative Animacy=superior has a significantly positive parameter estimate. Thus, we can consider the third freezing hypothesis to be preliminarily confirmed. The finding has to be treated as a good indication rather than solid evidence. Furthermore, Model 3 suggests that the effects of definiteness superiority and animacy superiority can be observed at the same time in the data. Further testing of the freezing hypotheses on a larger corpus annotated for animacy is needed in order to draw firmer conclusions. To the extent that we accept the results of Model 3, we have found a second piece of empirical support for the bidirectional model of word order, and the reality of word order freezing in spoken Dutch.

Future work on animacy and object fronting should also benefit from comparison with other languages. Øvrelid (2004) presents corpus evidence for the influence of animacy on Norwegian object-fronting. She proposes a Stochastic OT comprehension model that captures the predictions of the third freezing hypothesis. The problems associated with using Stochastic OT in a bidirectional setting were discussed in Section 5.6. Another interesting finding in relation to object-fronting and animacy is presented in Snider and Zaenen (2006). Snider and Zaenen find that inanimates in spoken English have a greater chance of appearing in topicalized position. This result looks very similar to the data presented in Table 6.5: Inanimate objects fronted often. For the Dutch data, I offered an explanation in terms of animacy superiority and word order freezing. However, for English, such an explanation is unlikely: English word order is hardly ever ambiguous with respect to what is subject and what is object, so word order freezing is never needed. Snider and Zaenen (2006) do not offer a explanation for the effect they observe, but future work on animacy and fronting in Dutch and English data should try to take this similarity between Dutch and English topicalization into account.

6.4 Negative evidence for word order freezing?

We have thus far seen three hypotheses that were inspired by the bidirectional model of word order freezing. The three freezing hypotheses were:

First Personal pronoun subjects may be preceded by direct objects because they are highly recognizable as subjects.

Second Definiteness superiority positively influences object fronting, definiteness inferiority negatively influences it.

Third Animacy superiority positively influences object fronting, animacy inferiority negatively influences it.

The first freezing hypothesis was rejected in favour of an alternative that explained more of the data (the unimportance hypothesis). The second freezing hypothesis was found partly confirmed using a logistic regression model, which lead us to accept it – there is a positive effect of relative definiteness on object fronting. The third hypothesis found preliminary support in the data – there is suggestive evidence for the existence of a positive effect of animacy on object fronting. To summarize, we have found support for the claim that Dutch shows word order freezing and we have also seen that some patterns in the data that look like they could be result of word order freezing are better explained in other ways. In this section, I wish to take a brief tour of a part of the data to see if there are any blatant violations of word order freezing. Whereas the former sections where concerned with positive evidence for freezing, this section will be about negative evidence.[4]

Clear counterexamples to the particular account of word order freezing that I have elaborated upon in the previous chapter would be instances of OVS in which there is no formal information as to which NP is the subject, nor information from definiteness or animacy superiority. In these cases, the direct object would precede the subject, even though there is no linguistic information as to what the subject is.

There are 25 instances of OVS that meet the requirement of not having a pronominal subject and object, which excludes case as formal information.[5] Of these, 8 identify their subject by means of agreement. In 17/25 cases, one can identify the subject on the basis of relative animacy. There are no cases in which the object is higher in animacy than the subject, which would have meant that relative animacy would have led the hearer to select the wrong NP as the subject. The 8/25 cases in which animacy supe-

[4]Negative evidence is used here as meaning 'evidence from an event not occurring'. The language acquisition related meaning of 'corrective response' is not the intended one.

[5]It is fully conceivable that not all of the other OVS instances have distinguishing case, however I will for reasons of time not look into all of these, and focus on the 25 selected sentences.

riority does not identify the subject are thus sentences in which the subject and object are of equal animacy.

In 11/25 cases, the subject can be recognized on the basis of relative definiteness. In 2/25 cases, the object is higher than the subject in definiteness. One example is given in (10). Relative animacy and the order of the postverbal constituents *een huisarts* 'a general practitioner' and indirect object *u* 'you.FORM' provide clues as to which is the subject. The direct object and subject are in boldface.[6]

(10) **hetzelfde verhaal** kan **een huisarts** u vertellen.
the same story can a general practitioner you tell
'A GP will tell you the same.' (Nl-b 137:86)

In the other case of definiteness inferiority, animacy superiority and agreement guarantee recoverability of the subject. The example in (11) is a example where only animacy information, and no definiteness information or formal information like word order in the Mittelfeld or agreement, allows one to identify the subject. The example has been shortened, but the translation is of the entire sentence.

(11) **deze maatregel** wil **staatssecretaris Ella Kalsbeek** nemen om [+9w]
this measure wants secretary of state Ella Kalsbeek take to
'Secretary of state Ella Kalsbeek wants to take this measure in order to [reduce the enormous influx of young refugees].' (Nl-k 1800:2)

In the set of 25 OVS sentences, there is only one case where, possibly, one would need to appeal to more complex factors than agreement, relative definiteness, or relative animacy to recognize the subject. It is given in (12).

[6]Henriëtte de Swart (pc) points out that the order of the subject *een huisarts* and the indirect object *u* can be reversed without losing the reading in which the GP is subject:

(i) hetzelfde verhaal kan u een huisarts vertellen
the same story can you.FORM a general practitioner tell
'A GP will tell you etc.' (DO V IO SV) *Or*: 'You will tell a GP etc.' (DO VS IO V)

I have concentrated on the argument order of subject and direct object, so I do not have a full explanation of why the order of subject and indirect object in the Mittelfeld of (i) is not frozen. Also, I would like to point out that the example in the body of the text only has the indicated reading (DO VS IO V). As to the indirect object-before-subject reading of (i), I think this can ultimately be explained in a bidirectional model of word order. In production, definiteness favours the order pronoun before indefinite full NP. Section 4.3 showed that both subjects and indirect objects are unlikely to be indefinite NPs and both are likely to be personal pronouns. Also, we can expect that indirect objects are as or even more animate than subjects. Relative definiteness and relative animacy need thus not be helpful in comprehension. The indirect-object–before–subject reading may come about by a preference in Dutch for full NP indirect objects to be realized as PP-arguments, and not NPs (Van der Beek, 2005). It might be that this preference is even stronger for indefinite NPs. Whether this preference can be formulated in such a way that it supports recoverability in comprehension optimization remains to be investigated.

(12) **het filerecord** heeft **Oostenrijk** met dertig kilometer auto's [+6w]
the congestion record has Austria with thirty kilometer cars
'Austria is the keeper of the congestion record with thirty kilometers of cars [north of the Tauerntunnel].' (Nl-k 6009:3)

Here, in order to recognize that Austria is the subject, if one is not willing to assign Austria to the group of animates or some kind of intermediate group, one will have to appeal to some kind of knowledge of the world, and say that a country holding a record is a much more likely state of affairs that a record holding a country.

In this section, I have very briefly looked at a group of sentences that one might expect to contain some counterexamples to freezing, if these exist. There was one sentence in the 25 sentence OVS subset that we would have predicted to be frozen using only the simple information sources that we have focused on so far: formal information like agreement, relative definiteness, and relative animacy. This example can readily be explained using knowledge of the world, although, of course, allowing this without further specification considerably weakens our account. In all other examples the subject could be found without needing to use knowledge of the world. This finding is encouraging, because it suggests that the identification of freezing-preventing factors in Chapter 5 is on the right track. Of course, without knowing whether the situation is different in the SVO data, we cannot claim that the results in this section constitute evidence for word order freezing.

6.5 Conclusions

In this chapter, I have looked for empirical support for the claim that word order freedom in Dutch is dependent on the availability of word order independent information about the correct grammatical function assignment. The bidirectional model of word order in Chapter 5 is built on this claim, and in the course of developing this model we were able to identify specific sources of information that are used by a hearer to assign grammatical function. The search for empirical support was guided by a quantitative interpretation of the bidirectional model of word order. I have investigated the hypothesis that the proportion of object initial sentences is related to the amount of information available to guide the hearer in grammatical function assignment.

I started with investigating whether the increase in OVS that is observed when the subject is pronominal is related to the fact that (personal) pronoun subjects are highly recognizable as subjects. This first freezing hypothesis was rejected in favour of an analysis that relied on the low tendency of personal pronoun subjects to appear in the Vorfeld. This behaviour of personal pronoun subjects is fully expected in light of the

conclusions of Chapter 4: Personal pronouns in general make poor Vorfeld occupants because they do not realize important information.

The second and third freezing hypotheses were based on one template. Subjects tend to be highly definite, and objects tend to be indefinite. Likewise, subjects tend to be animate, and objects tend to be inanimate. This default association between subjecthood and definiteness or animacy can be considered to be information that helps the hearer in grammatical function assignment, but only when the subject of the sentence is in fact more definite or animate than the object. We refer to these cases as definiteness- or animacy-superiority. In this chapter I have shown that object fronting increases in the presence of definiteness superiority. We have also seen preliminary evidence that object fronting increases in sentences that show animacy superiority. Speakers of Dutch rely less on canonical argument order when the subject and object are identifiable on the basis of their relative definiteness or relative animacy. Negative effects on object fronting that we might expect when the default association between subject and definiteness leads to the wrong grammatical function assignment, that is, when the subject is lower in definiteness than the direct object, were not observed. We can conclude that Dutch shows signs of word order freezing, albeit as a trend rather than an absolute effect.

Chapters 5 and 6 combined provide evidence that communicative success is a factor in the selection of a Vorfeld occupant. In addition to the properties that make a constituent a good or a poor Vorfeld constituent in general, discussed in Chapters 2 and 4, the chance that a direct object ends up in the Vorfeld depends on whether the resulting word order will be interpreted correctly.

Future work could proceed in two directions. The first is to try to derive the quantitative predictions directly from the bidirectional model, so that we do not have to rely on an interpretation step to test the model against a corpus. Moreover, the results in this chapter have shown that word order freezing in Dutch is a gradient affair: When there is less information about grammatical function assignment, non-canonical word order is less common. Ideally the bidirectional model should predict this gradient nature.

The second direction for future work tries to gather more corpus evidence, and to make the existing evidence more solid. The results with respect to relative animacy need to be investigated on a larger corpus. It may also be a good idea to compare relative animacy as a source of grammatical function assignment information to selectional restrictions. The latter may of course vary from verb to verb, and therefore predict freezing in different situations. In addition, I proposed in Chapter 5 that a hearer uses information structure, or intonation as a proxy for it, to base grammatical function assignment on. It would be interesting to check this hypothesis against a corpus just as we have done with definiteness and animacy in this chapter. However, this would require a corpus with specific annotation. As far as I am aware, there is no large scale corpus with such annotation available for Dutch.

Chapter 7

Conclusions

There are many ways to start a sentence in Dutch. We may decide to mention the subject of a sentence first (1a), the direct object (1b) or the indirect object (1c); to name only three of the many possibilities.

(1) a. **Ik** zal haar de hardste klapzoen ooit geven.
I will her the loudest kiss ever give

b. **Haar** zal ik de hardste klapzoen ooit geven.
her I the loudest kiss ever

c. **De hardste klapzoen ooit** zal ik haar geven.
the loudest kiss ever I her
'I will give her the loudest kiss ever.'

In the previous six chapters, I have investigated the variation exemplified in (1), known as *Vorfeld occupation*. Vorfeld occupation of subject, direct object, and indirect object in spoken Dutch was investigated from a theoretical and empirical perspective. In the first half of the thesis, I explored constituent properties that promote or discourage Vorfeld occupation. The investigations were carried out on the Spoken Dutch Corpus (CGN, 2004). Statistical analysis showed that each of the investigated constituent properties contributed independently. In the second half of the thesis, I considered the relation between word order variation and the distinguishability of subject and object. I extended an Optimality-theoretic model that is able to capture this relation. I showed that the predictions of this model can be found as trends in spoken Dutch using statistical analysis of the spoken Dutch corpus.

In this final chapter, I will summarize the main findings of the investigation of Vorfeld occupation, and discuss directions for future research.

7.1 Summary of main findings

The mere observation that Vorfeld occupation may vary does not tell us much about Vorfeld occupation itself. For instance, it does not tell us whether each of the variants in (1) can be used in the same situations, whether they are equally common, and what drives the choice between variants. About 70% of all Dutch main clauses begin with their subject. Although the options to start with a direct object or indirect object are utilized a lot less in comparison, these options are not marginal or uncommon. In Chapters 2 and 4, I showed that the choice for a Vorfeld occupant can be understood in terms of a combination of general word order trends and the special property of the Vorfeld as a position for highlighted or foregrounded material.

Motivated by word order trends in the Dutch and German *Mittelfeld*, between the finite verb in second position and the non-finite verbs towards the end of the sentence, and by previous research on word order in Dutch, German, and English, I investigated the influence of three factors on Vorfeld occupation: grammatical function, definiteness and grammatical complexity. For each factor, I found that they significantly influence the choice of Vorfeld occupant. Below I will discuss the influence of each of these factors on Vorfeld occupation, and their relation to word order in the Dutch Mittelfeld.

1. Grammatical function Canonical word order in the Dutch Mittelfeld is subject before object, and indirect object before direct object. As for Vorfeld occupation, we find that subjects have the strongest tendency to appear in the Vorfeld. The difference between indirect object and direct object is small, but indirect objects appear to topicalize slightly more often than direct object. More indirect object data is needed to make this claim more solid.

The influence of grammatical function on Vorfeld occupation can therefore be summarized as in (2). The symbol '$\prec$' should be read as 'has a stronger tendency to occur sentence initially'.

(2) *Positive relation between grammatical function and Vorfeld occupation*:
subject $\prec$ indirect object $\preceq$ direct object.

The grammatical function hierarchy in (2) also describes canonical word order in the Dutch Mittelfeld. The behaviour of arguments in the Vorfeld therefore reflects facts about word order in the Mittelfeld.

2. Definiteness The relation between definiteness and Vorfeld occupation is rather complex. We can draw up a scale of definiteness: pronouns, definite full NPs, indefinite full NPs. Word order in the Dutch and German Mittelfeld is sensitive to this threefold distinction: Elements at the left of the scale have a stronger tendency to be realized early in the Mittelfeld.

In line with word order in the Mittelfeld, definite full NPs are more likely to appear in the Vorfeld than indefinite full NPs are. However, with respect to the behaviour of pronouns, we see differentiation between the Vorfeld and the Mittelfeld. As is well known, reduced personal pronoun objects are not allowed to appear in the Vorfeld. The corpus results showed that reduced personal pronouns have a strongly reduced chance of appearing in the Vorfeld, not only in the object data, but also in the subject data. In fact, personal pronouns in general front less often than definite full NPs do. Demonstrative pronouns, on the other hand front very frequently, at levels far above definite full NPs.

We may summarize the two conflicting trends found in the relation between definiteness and Vorfeld occupation as (3) and (4).

(3) *Positive relation between definiteness level and Vorfeld occupation:*
pronoun $\prec$ definite full NP $\prec$ indefinite full NP

(4) *Negative relation between pronominal form and Vorfeld occupation:*
reduced personal pronoun $\succ$ full personal pronoun $\succ$ demonstrative pronoun

The discouraging effect on Vorfeld occupation of being a personal pronoun is stronger than the positive effect of appearing further to the right on the definiteness scale.

The trends summarized in (3) and (4) are observable in the subject and object data alike, although the contrasts between different definiteness levels and pronominal forms are stronger in the object data. The pronominal form scale of (4) can be interpreted in terms of a universal word order principle proposed by Gundel (1988). The *first-things-first* principle states that important information should be uttered first, where important information is information that is new, unpredictable or contrasted. Personal pronouns, and especially reduced personal pronouns, realize material that is highly predictable. Consequently, personal pronouns in general do not realize important material. The fact that in all functions personal pronouns front less frequently than definite NPs means that in Dutch Vorfeld occupation, the first-things-first principle is more important than the definiteness scale.

3. Grammatical complexity The claim that there is a tendency to order simpler material before more complex material is at least as old as Behaghel (1909). For Dutch, Haeseryn et al. (1997) formulate a *complexity principle* to the same extent. Indeed, in the spoken Dutch corpus, we find that material in the Vorfeld is less complex than postverbal material.

However, the conclusion that complexity directly influences Vorfeld occupation was not warranted. Upon closer inspection, complexity affects placement at the right periphery. Constituents at the right periphery are more complex than constituents before it, which includes constituents in the Vorfeld. In the rest of the sentence, there are no word order effects of complexity. Even though there is no direct influence of complexity on Vorfeld occupation, we can again conclude that the Mittelfeld and Vorfeld show parallel behaviour.

Early realization in the Mittelfeld, and Vorfeld occupation are both associated with simple (or: less complex) constituents, since both positions are not at the right periphery.

The three factors grammatical function, definiteness, and grammatical complexity were each found to influence the decision of what to put into the Vorfeld. Generally put, subjects, demonstrative pronouns and grammatically simple material like to appear in the Vorfeld. Direct objects, indefinite full NPs and reduced personal pronouns, and grammatically complex constituents have a tendency to avoid the Vorfeld. Vorfeld occupation is influenced by the combination of global word order determinants (grammatical function, definiteness scale), a factor local to the Vorfeld (important material), and a factor local to the right periphery (complex material).

The investigation of Vorfeld occupation thus far has implicitly taken a speaker-oriented perspective. We may imagine that one of the tasks of the speaker is to distribute constituents over the sentence. One of the choices the speaker has to make in this task is the selection of a Vorfeld occupant. We have seen that this choice is influenced by constituent properties such as grammatical function, definiteness, and grammatical complexity. In fact, one of the statistical tools used explicitly takes this speaker perspective: In the *logistic regression* model in Section 4.5, the probability that a constituent appears in the Vorfeld is predicted from given information about the properties of this constituent; definiteness, length and grammatical function.

However, there are other considerations that come into play, too – considerations that cannot be straightforwardly phrased as constituent properties. This becomes clear when we look at the impact of word order on a hearer. One of the tasks of the hearer is to assign grammatical function to each of the constituents, given information about their position in the sentence, definiteness and grammatical complexity. Because grammatical function assignment is not given for the hearer, and position in the Vorfeld is not limited to one grammatical function, there is the risk that the hearer assigns a grammatical function to the Vorfeld constituent that was not intended by the speaker.

For free word order languages as diverse as German, Russian, Hindi and Japanese, it has been observed that word order variation is restricted or even banned when grammatical function assignment cannot be determined by other factors besides word order. In that case, word order is limited to canonical word order. This situation is referred to as *word order freezing*. From a hearer's perspective, word order freezing means that word order becomes a reliable source of information of grammatical function assignment. From a speaker's perspective, word order freezing is a limiting factor in word order variation.

Word order freezing can be observed in relation to Vorfeld occupation in Dutch as a tendency rather than an absolute effect. An example is (5).

(5) Jan heeft Piet aan de kant gezet.
Jan has Piet aside put.
'Jan has dumped Piet.' *Not, or much less likely:* 'Piet has dumped Jan.'

The object-initial interpretation of (5) can easily be obtained by intonation or context. The subject-initial interpretation needs no such help, however. The availability of a non-subject initial word order is contingent on the recognizability of the arguments. As a result, we can expect that Vorfeld occupation by, say, a direct object is not only influenced by the factors of grammatical function, definiteness and complexity presented above, but also by how well the direct object can be recognized as a direct object.

Building on previous work by primarily Lee (2001b), I argued in Chapter 5 that word order freezing is best modeled by combining the speaker and hearer perspectives in a model of word order variation. In a *bidirectional Optimality-theoretic* model, the speaker and hearer tasks can be modeled formally. In such a bidirectional model, the speaker's choice for a word order variant is influenced by the hearer's ability to correctly recognize the arguments. The hearer may use information like morphology, context, intonation, and the association between definiteness and subjecthood and between animacy and subjecthood for this task. Even if an object-initial word order is preferred from a speaker's perspective, it may be disallowed because from the hearer's perspective, the resulting sentence receives a subject-initial interpretation.

In Chapter 6, I gave the discrete predictions of the bidirectional Optimality-theoretic model a quantitative interpretation and evaluated these predictions on a subset of the spoken Dutch corpus. Word order is not restricted when the subject and object can be discerned from each other on independent grounds. Two factors that help in distinguishing the subject from the object are relative definiteness and animacy. Subjects tend to be highly definite and animate, whereas direct objects tend to be indefinite and inanimate. We therefore predict that when the subject is highly definite or animate, and the object is not, non-canonical argument order is more frequent (positive effect). Investigation of transitive sentences in the spoken Dutch corpus confirmed the positive effect of relative definiteness. Analysis of a manually annotated subset of the data also strongly suggested that the positive effect of relative animacy can be observed. One may also predict that there is a decrease in non-canonical argument order when the subject is lower on the definiteness scale than the object (negative effect). However, the negative effects of relative definiteness and animacy were not found confirmed in the corpus.

The positive effects of relative definiteness and animacy, that were predicted to exist on the basis of the bidirectional model of word order, were found on top of the factors in Vorfeld occupation established in Chapter 4. In spoken Dutch, a speaker uses canonical word order more often when there is a risk that the listener will understand the utterance incorrectly. I conclude that word order freezing is a real, but not absolute, phenomenon

in spoken Dutch discourse, and not just in abstract sentences in isolation and without intonation such as the sentence presented in (5).

The central question of this thesis is what determines the choice between word order variants that differ in Vorfeld occupant. We have seen that the overall picture is painted by factors of very diverse nature and scope. The choice for a Vorfeld occupant is influenced by global word order considerations (grammatical function and definiteness), by local factors that relate certain material to certain positions (important material in the Vorfeld, complex material at the right-periphery), and by the requirement that the intended meaning has a chance of being communicated (word order freezing). The combination of theoretical modeling, large scale corpus study and statistical modeling has allowed us to identify and investigate these factors and their interaction in the choice of a Vorfeld occupant.

7.2 Directions for future work

I would like to use the rest of the chapter to indicate some of the many directions that the work of this dissertation can be extended in. I will discuss these directions in two steps corresponding to the two halves of this dissertation.

The functional characterization of Vorfeld occupation I used in this thesis involved the intuitive notion of importance. The effects of these functional properties were most clearly manifested in the diverse behaviour of different pronominal forms. In support of the idea of informational importance, I have cited typological work like that of Gundel (1988). However, a more formal and precise definition of the information structural properties of the Vorfeld has not been given. A lot more work is needed before these properties can be defined in a formal way. I hope that the corpus results of Chapter 4 can contribute in this effort. However, I suspect that the type of corpus investigation presented in this dissertation will ultimately be unable to address this question because of the reliance on surface form correlates of semantic/pragmatic distinctions. This does not mean that surface properties will become redundant. Bresnan et al. (2007) show that discourse-semantic properties like givenness as well as highly correlated surface properties like pronominality and definiteness independently influence dative alternation in English. Even though we can treat the discourse-semantic properties and the surface properties as correspondents, we cannot reduce one to the other.

The corpus investigation in the first half of the dissertation can be extended in several ways. To begin with, indirect objects are so rare that the syntactically annotated part of the spoken Dutch corpus is not large enough to draw very solid conclusions about the Vorfeld behaviour of indirect objects. More indirect object data should be analyzed to test whether the conclusions about indirect object fronting I drew in this dissertation truly hold.

Furthermore, future research into other factors that could influence word order and thereby Vorfeld occupation may include information about the verb. Informal inspection of the data reveals that many of the Vorfeld direct objects occur in sentences whose main verbs were verbs of saying or knowing. It may be the case that verbs of saying and knowing promote Vorfeld occupation by their direct objects. The syntactic annotation of the corpus used in this dissertation does not contain information about what the main verb in a sentence is. However, there is a good chance that we may infer this automatically from the existing annotation with an accuracy that is good enough to allow statistical investigation of the role of the main verb. Similarly, Jansen (1981) noted that the combination of a deontic auxiliary and a second person subject frequently occurred in OVS order. In this case, it may be that the particular combination of auxiliary and subject promotes direct object fronting.

One possible confound in the corpus study of Vorfeld occupation was the existential construction (EC), which has probably influenced the definiteness effects in the subject data. On the basis of corpus studies of the existential construction in English and Dutch, Beaver, Francez, and Levinson (2006) show that one can draw scales of determiners of subjects, according to how strongly they are associated with realization in an EC. They argue that these can be related to universal subject markedness scales. The resulting scales have a lot in common with the definiteness scales I have assumed in this dissertation. It would be interesting to see whether future research can shed more light on the relation between subjects appearing in the Vorfeld (or early realization in general) and those in an EC (see Birner and Ward, 1998 for related work on English).

Finally, a methodological follow-up on the corpus study should address possible concerns about the validity of the results in this thesis given that there could be inter-speaker variation. Snider and Zaenen (2006) and Bresnan et al. (2007) use statistical techniques to show that their findings about word order in English are robust against influences of speaker variation. On the one hand, this is a reassurance for the results for this thesis – the *type* of findings presented here can be reliably found in a multi-speaker corpus – on the other hand it suggests that the results in this thesis can and should be made stronger by using one of these statistical techniques.

In the second part of the thesis, I have argued in favour of a bidirectional approach to grammar. Bidirectional Optimality-theoretic grammars are different in nature from their (standard) unidirectional counterparts and rather restrictive in comparison to them. The choice for bidirectional grammar should therefore be as well motivated as possible. The predictions regarding relative definiteness and animacy derived from the bidirectional model of word order freezing have been partly confirmed, but especially the influence of animacy on word order and word order freezing needs to be investigated in a larger corpus. We should also try to find evidence for bidirectionality outside of word order

freezing and Vorfeld occupation. In Section 5.7, I speculated that the V2/V3-variation in German and embedded object scrambling in Dutch are sensitive to hearer considerations just like Vorfeld occupation is. If these phenomena, and others like them, can be captured in the bidirectional model, this would further justify the increased complexity of bidirectional grammar.

A consequence of the bidirectional grammar architecture that I have been using in this dissertation is that all variation in language must have a double explanation: from the speaker's perspective and from the hearer's perspective. For the word order variation treated in this dissertation, finding double explanations has proven to be possible. It remains to be seen whether all instances of ambiguity and optionality can be analyzed in this way.

The corpus investigation into word order freezing of Chapter 6 relied on a quantitative interpretation of the non-quantitative, discrete bidirectional model of word order. It would be a great improvement if we were able to derive these predictions about corpus trends more directly from the bidirectional model. Section 5.6 discusses some of the hurdles that have to be taken before we are able to do this. As a first step, one could consider the stochastic bidirectional OT work presented in Jäger (2004). A model of word order that is capable of making gradient or quantitative predictions may also integrate the trends found in the first half of the thesis better.

In this dissertation, I have teased out some of the basic word order tendencies that combine into the complex patterns in Vorfeld occupation that we can observe in spoken Dutch. In addition, we have seen evidence of the influence of communicative success on production, as formalized by bidirectional OT. Although the investigations presented in the preceding chapters deal with a specific construction in a specific language, the potential scope of the methods and analyses is wider than just the Dutch Vorfeld. For instance, all Germanic languages allow for argument fronting of the kind studied in this dissertation; most Germanic languages even have a position that resembles the Dutch Vorfeld. A natural next step would thus be to systematically investigate argument fronting in the other Germanic languages and to compare the relative influence of the basic word order tendencies that I have considered here. Such a cross-linguistic, comparative approach is also motivated for the investigation of the effect of communicative success on word order variation, since typological predictions are inherent in any OT based analysis, including the bidirectional OT analysis I have defended. It would be an exciting enterprise to try and find corpus evidence that communicative success influences word order variation in other languages, too, using the methods I have employed for Dutch. Finding such evidence would strengthen the claim that a comprehensive description of word order variation must encompass the speaker perspective and the hearer perspective.

Appendix A

List of Abbreviations

This appendix provides an overview of the abbreviations used by the CGN, and used in this dissertation when discussing CGN annotation. The translations/explanations are minimal and rather rough. For a in depth discussion and a more comprehensive list I refer to Hoekstra et al. (2003) and Van Eynde (2003) (both in Dutch). The abbreviations are listed alphabetically per type of concept. The following types are discerned: syntactic categories (phrase labels, POS-tag) and dependencies (grammatical function, discourse function).

A.1 Syntactic categories

Phrase labels

ADVP	adverb phrase
AHI	infinitival clause headed by *aan het* (indicating progressive aspect)
AP	adjective phrase
CONJ	result of conjoining two elements
CP	complementizer phrase
DETP	determiner phrase (complex determiners)
DU	discourse unit
INF	infinitival verb phrase
LIST	as CONJ but without an (overt) conjunction
MWU	multiword unit (mainly full proper names)
NP	noun phrase
OTI	infinitival verb phrase headed by *om te, voor te* (∼'to')

PP	preposition phrase
PPART	verb phrases headed by past or passive participles
PPRES	verb phrases headed by present participles
REL	relative clause
SMAIN	main (finite) clause
SSUB	subordinate (finite) clause
SV1	verb-initial finite clause
SVAN	finite subordinate clause headed by *van* (∼reporting 'like')
TI	infinitival verb phrase with *te* ('to')
WHQ	V2 wh-question
WHREL	free relative clause
WHSUB	subordinate clause headed by a wh-word

Part-of-speech tags

Given below are only the major divisions in the CGN Part of speech tag set. See Van Eynde (2003) for a detailed description.

ADJ	*adjectief*: adjectives
BW	*bijwoord*: adverbs
LID	*lidwoord*: articles
N	*naamwoord*: nouns, proper names
SPEC	*speciaal*: reserve, e.g. foreign words, non-speech, unrecognisable, parts of proper names.
TW	*telwoord*: numerals
VG	*voegwoord*: coordinating, subordinating and correlative conjunctions
VNW	*voornaamwoord*: pronouns, quantifiers
VZ	*voorzetsel*: prepositions
WW	*werkwoord*: verbs

A.2 Dependencies

Grammatical Functions

APP	apposition
BODY	body of a clause or VP (i.e., not complementizer, relativizer, question word, etc.)
CMP	complementizer

CNJ	conjunction
CRD	conjoined member
DET	determiner
HD	head
HDF	final, postpositional part of a circumposition (daughter of PP)
LD	locative/directional argument
LP	LIST member
ME	measure
MOD	modifer
MWP	part of multi-word unit (MWU)
OBCOMP	object of comparative
OBJ1	direct object
OBJ2	indirect object
PC	prepositional complement
POBJ1	dummy direct object
PREDC	predicative complement
PREDM	predicative modifier
PRT	particle
RHD	head of relative clause (REL)
SE	obligatory simplex reflexive argument
SU	subject
SUP	dummy subject
SVP	verb particle
VC	verbal complement
WHD	head of a WHREL/WHQ/WHSUB

(Rudimentary) Discourse Functions

DLINK	discourse link
DP	discourse part
NUCL	nucleus
SAT	sattelite (lead and tail as in Ch. 3)
TAG	tag

Appendix B

Examples of Vorfeld Occupants in CGN

Below are examples for each of the dependency paths in Table 4.1, p89. The Vorfeld occupants are given in boldface. For each dependency path the total number of times it occurs in the corpus is given, as well as the proportion of occurrences that lead to a Vorfeld occupant.

(1) SUP 2 731 occ., 67.3% in Vf
d'r was echt een brand ontstaan
EXPL was really a fire come into existence
'A fire had really started.' (Nl-a 252:137)

(2) SU 71 148 occ., 65.0% in Vf
zij loopt altijd met hoge hakken
she walks always with high heels
'She's always wears high heels.' (Nl-a 250:47)

(3) MOD 70 276 occ., 21.4% in Vf
dan sta je zo voor paal
then stand you so much for pole
'At that moment you look like such a fool.' (Nl-a 250:159)

(4) OBJ1 PC SU 86 occ., 15.1% in Vf
daar zijn de meningen over verdeeld
there are the opinions about divided
'The opinions on that are divided.' (Nl-f 7176:49)

(5) OBJ1 VC 8 157 occ., 14.6% in Vf
dat heb ik wel 'ns verteld misschien
that have I AFF PART told perhaps
'I might have told that already.' (Nl-l 250:140)

(6) OBJ1 18 411 occ., 14.3% in Vf
dat zegt ie altijd
that says he.RED always
'That's what he always says.' (Nl-a 250:139)

(7) OBJ1 PC BODY VC 140 occ., 13.6% in Vf
daar heb je van af te blijven
there have you of off to stay
'You're not allowed to touch that.' (Nl-a 938:162)

(8) OBJ1 VC VC 774 occ., 13.4% in Vf
die rails zou je kunnen ophangen
those rails would you can hang up
'You could hang up those rails.' (Nl-a 254:479)

(9) OBJ1 PC VC 1 428 occ., 13.2% in Vf
daar hadden dus mensen op ingeschreven
there had PART people on signed up
'People had signed up for that.' (Nl-a 391:50)

(10) OBJ1 PC PREDC 444 occ., 13.1% in Vf
daar ben je heel goed in
there are you very good in
'You're very good at that.' (Nl-a 415:191)

(11) OBJ1 PC OBJ1 259 occ., 12.9% in Vf
daar heb ik geen plek meer voor
there have I no place anymore for
'I have no place left for that.' (Nl-a 303:160)

(12) LD 5 653 occ., 11.8% in Vf
naar Bouillon gaan ze nu deze zomer
to Bouillon go they now this summer
'They're now going to Bouillon this summer.' (Vl-b 400169:505)

(13) OBJ1 PC 2 412 occ., 10.7% in Vf
daar blijkt 't al wel uit
there appears it PART AFF from
'That already makes it clear.' (Nl-f 7233:14)

(14) OBJ1 PC VC VC 136 occ., 10.3% in Vf
daar hebben we ook aan zitten denken
there have we also of sit think
'We've also been thinking of that.' (Nl-a 816:100)

(15) OBJ1 BODY VC 486 occ., 10.1% in Vf
't rampgebied zijn ze nu aan 't ruimen
the disaster area are they now clear.PROG
'They're clearing the disaster area now.' (Nl-l 7566:59)

(16) POBJ1 220 occ., 9.5% in Vf
dat zeiden ze ook van ja fantastisch boek
that said they also QUOT yeah great book
'That's what they said, (like) "yeah, great book."' (Nl-a 568:74)

(17) OBJ1 LD 3 551 occ., 8.4% in Vf
daar hangen ballen in van dertig jaar oud
there hang balls in of thirty years old
'It's got thirty year old (christmas) balls in it.' (Nl-a 273:14)

(18) PREDM 1 498 occ., 5.8% in Vf
ijskoud geven we God ook nog de schuld van wat wij verkeerd doen
ice cold give we God too the blame of what we wrong do
'Without shame, we even blame God for what we do wrong.' (Nl-m 298:64)

(19) OBJ2 VC 441 occ., 5.7% in Vf
die moeten we ook water geven morgen
those must we also water give tomorrow
'We should water those tomorrow, too.' (Nl-a 594:67)

(20) OBJ1 MOD VC VC 265 occ., 5.7% in Vf
daar zijn Frank en ik toen mee gaan eten
there are Frank and I then with go eat
'Then Frank and I went out to eat with him/her.' (Nl-c 8023:253)

(21) LD VC 1 868 occ., 4.7% in Vf
hier is het gebeurd
here is it happened
'This is where it happened.' (Nl-a 260:249)

(22) OBJ1 LD VC 13 064 occ., 4.5% in Vf
daar vond hij dit wel in passen
there found he this AFF in fit
'He thought this would be appropriate.' (Nl-a 374:179)

(23) MOD VC 6 374 occ., 4.1% in Vf
zo moeten we 't doen
so must we it do
'This is how we should do it.' (Nl-a 265:216)

(24) OBJ1 MOD OBJ1 814 occ., 3.9% in Vf
daar heb ik een berekening voor
there have I a calculation for
'I have a way of calculating that.' (Nl-a 321:89)

(25) OBJ2 592 occ., 3.9% in Vf
die zeg ik gedag
DEM say I goodbye
'I tell her/him goodbye.' (Nl-a 389:58)

(26) OBJ1 PREDC 1 083 occ., 3.9% in Vf
dat raak je niet kwijt
that COP you not rid
'You don't lose that.' (Nl-a 505:97)

(27) OBJ1 MOD VC 2 113 occ., 3.7% in Vf
hier heb ik ontzettend veel van geleerd
here have I very much of learnt
'I have learnt a great deal from this.' (Nl-a 374:208)

(28) TAG OBJ1 618 occ., 3.6% in Vf
nou zegt die zuster dat is niet gebruikelijk
now says that sister that is not usual
"'Well," said the nurse, "that's unusual."' (Nl-a 254:250)

(29) OBJ1 MOD SU 611 occ., 3.6% in Vf
daar stond een stuk van in Trouw een tijd geleden
there stood an article of in Trouw a while ago
'There was an article by him/her in Trouw a while ago.' (Nl-a 593:234)

(30) MOD VC VC 697 occ., 3.4% in Vf
de twaalfde had ie af moeten zijn
the twelfth had heRED finished must be
'It should have been finished by the twelfth.' (Nl-a 959:86)

(31) OBJ1 MOD OBJ1 VC 470 occ., 3.4% in Vf
daar heeft u geen exemplaren meer van bewaard
there have you no copies more of save
'You didn't save any copies of that.' (Nl-j 7280:11)

(32) PC VC 1 582 occ., 3.1% in Vf
over jou gaan we praten
about you will we talk
'We will talk about you.' (Nl-a 446:112)

(33) PREDC VC 1 233 occ., 2.8% in Vf
zoiets moet 't dan zijn
something like that must it then be
'It has to be something like that.' (Nl-a 610:352)

(34) PREDC 17 290 occ., 2.7% in Vf
zo erg is 't niet
so bad is it not
'It's not that bad.' (Nl-a 490:69)

(35) OBJ1 MOD 5 467 occ., 2.6% in Vf
daar werken we dan wel omheen
there work we then AFF around
'We'll work around that, then.' (Nl-a 392:114)

(36) PC 2 628 occ., 2.4% in Vf
niet alleen van brood alleen leeft de mens
not only of bread alone lives the man
'Man does not live on bread alone.' (Nl-m 271:1)

Bibliography

Ackema, Peter and Ad Neeleman. 2000. Absolute ungrammaticality. In Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer, editors, *Optimality Theory: Phonology, Syntax, and Acquisition*. Oxford University Press, Oxford, pages 279–301.

Agresti, Alan. 1996. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Inc., New York, NY.

Agresti, Alan. 2002. *Categorical Data Analysis*. John Wiley & Sons, Inc., New York, NY, second edition.

Aissen, Judith. 1999. Markedness and subject choice in Optimality Theory. *Natural Language and Linguistic Theory*, 17:673–711.

Aissen, Judith. 2003. Differential object marking: iconicity vs economy. *Natural Language and Linguistic Theory*, 21:435–83.

Altmann, Hans. 1981. *Formen der 'Herausstellung' im Deutschen. Rechtsversetzung, Linksversetzung, Freies Thema und verwandte Konstruktionen*. Niemeyer, Tübingen.

Anttila, Arto. 1997. Deriving variation from grammar. In Roeland van Hout Frans Hinskens and Leo Wetzels, editors, *Variation, Change and Phonological Theory*. John Benjamins, Amsterdam/Philadelphia, pages 35–68. Also: ROA-63.

Anttila, Arto and Vivienne Fong. 2000. The partitive constraint in Optimality Theory. *Journal of Semantics*, 17(2):281–314.

Arnold, Jennifer. 1998. *Reference Form and Discourse Patterns*. Ph.D. thesis, Stanford.

Arnold, Jennifer. 2006. Reference resolution: both given and new can be expected. Extended abstract of an invited talk given at the Conference on intersentential pronominal reference in child and adult language. ZAS Berlin. 1–2 December 2006.

Arnold, Jennifer, Thomas Wasow, Ash Asudeh, and Peter Alrenga. 2004. Avoiding attachment ambiguities: the role of constituent ordering. *Journal of Memory and Language*, 51(1):55–70.

Asudeh, Ash. 2001. Linking, optionality and ambiguity in Marathi. In Peter Sells, editor, *Formal and Empirical Issues in Optimality Theoretic syntax*, volume 5 of *Studies in Constraint based Lexicalism*. CSLI Publications, Stanford.

Baayen, Harald. forth. *Analyzing Linguistic Data. A Practical Introduction to Statistics*. Cambridge University Press.

Beaver, David. 2004. The optimization of discourse anaphora. *Linguistics and Philosophy*, 27(1):3–56.

Beaver, David, Itamar Francez, and Dmitry Levinson. 2006. Bad subject: (non-)canonicality and NP distribution in existentials. In Effi Georgala and Jonathan Howell, editors, *Proceedings of Semantics and Linguistic Theory XV*. CLC Publications, Ithaca, NY.

Beaver, David and Hanjung Lee. 2003. Form-meaning asymmetries and bidirectional optimization. In Jennifer Spenader, Anders Eriksson, and Östen Dahl, editors, *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pages 138–48. University of Stockholm.

Beaver, David and Hanjung Lee. 2004. Input-output mismatches in OT. In Reinhard Blutner and Henk Zeevat, editors, *Optimality Theory and Pragmatics*. Palgrave Macmillan, Hampshire, pages 112–53.

Bech, Gunnar. 1952. Über das niederländische Adverbialpronomen *er*. In *Traveaux du Cercle Linguistique de Copenhague*, volume 8, pages 5–32. Consulted electronic version, 2002, to be found at `http://www.dbnl.org/tekst/bech001uber01/`.

van der Beek, Leonoor. 2005. *Topics in Dutch Corpus Based Syntax*. Ph.D. thesis, Rijks*universiteit* Groningen.

van der Beek, Leonoor and Gerlof Bouma. 2004. The role of the lexicon in Optimality Theoretic syntax. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG'04 Conference*, Stanford. CSLI Publications.

van der Beek, Leonoor, Gosse Bouma, Robert Malouf, and Gertjan van Noord. 2002. The Alpino dependency treebank. In Mariët Theune, Anton Nijholt, and Hendri Hondorp, editors, *Computational Linguistics in the Netherlands 2001*. Rodopi.

Behaghel, Otto. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, 25:110–42.

den Besten, Hans. 1985. The Ergative Hypothesis and free word order in Dutch and German. In J. Torman, editor, *Studies in German Grammar*, Studies in Gernerative Grammar 21. Foris, Dordrecht, pages 23–64.

Birner, Betty and Gregory Ward. 1998. *Information Status and Noncanonical Word Order in English*, volume 40 of *Studies in language companion series*. John Benjamins, Amsterdam/Philadelphia.

Bíró, Tamás. 2006. *Finding the Right Words: Implementing Optimality Theory with Simulated Annealing*. Ph.D. thesis, Rijks*universiteit* Groningen.

Bloom, Douglas. 1999. Case syncretism and word order freezing in the Russian language. Master's thesis, Stanford.

Blutner, Reinhard. 2000. Some aspects of optimality in natural language interpretation. *Journal of Semantics*, 17:189–216.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 21*, pages 43–58.

Boersma, Paul and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry*, 32:45–86.

Bosch, Peter, Graham Katz, and Carla Umbach. 2007. The non-subject bias of German demonstrative pronouns. In Monika Schwarz-Friesel, Manfred Consten, and Mareile Knees, editors, *Anaphors in Text: Cognitive, Formal and Applied Approaches to Anaphoric Reference*. John Benjamins, Amsterdam, pages 145–64.

Bouma, Gosse. 2000. Argument realization and Dutch R-pronouns: solving Bech's problem without movement or deletion. In Ronnie Cann, Claire Grover, and Philip Miller, editors, *Grammatical Interfaces in Head-driven Phrase Structure Grammar*. CSLI Publications.

Bouma, Gosse. 2004. Treebank evidence for the analysis of PP-fronting. In Sandra Kübler, Joakim Nivre, Erhard Hinrichs, and Holger Wunsch, editors, *Third Workshop on Treebanks and Linguistic Theories*, pages 15–26.

Bouma, Gosse, Petra Hendriks, and Jack Hoeksema. 2007. Focus particles inside prepositional phrases: a comparison of Dutch, English and German. *Journal of Comparitive Germanic Linguistics*, 10(1).

Bouma, Gosse and Geert Kloosterman. 2002. Querying dependency treebanks in XML. In *Proceedings of the Third international conference on Language Resources and Evaluation (LREC)*, Gran Canaria.

Bouma, Gosse and Geert Kloosterman. 2007. Mining syntactically annotated corpora using XQuery. In *Proceedings of the Linguistic Annotation Workshop (The LAW, ACL 07)*, Prague.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Braun, Bettina and D. Robert Ladd. 2003. Prosodic correlates of contrastive and non-contrastive themes in German. In *Proceedings of the 8th European Conference on Speech Communication and Technology*, pages 789–792, Geneva.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irenen Krämer, and Joost Zwarts, editors, *Cognitive Foundations of Interpretation*, Amsterdam. Edita/KNAW.

Buchwald, Adam, Oren Schwartz, Amanda Seidl, and Paul Smolensky. 2003. Recoverability Optimality Theory: discourse anaphora in a bidirectional framework. In *Proceedings of EdiLOG 02*.

Büring, Daniel. 2001. What do definites do that indefinites definitely don't? In Caroline Féry and Wolfgang Sternefeld, editors, *Audiatur Vox Sapientiae: A Festschrift for Arnim von Stechow*, number 52 in Studia Grammatica. Akademie Verlag, pages 70–100.

Büring, Daniel. 2003. On d-trees, beans, and B-accents. *Linguistics & Philosophy*, 26(5):511–45.

Büring, Daniel and Katharina Hartman. 2001. The syntax and semantics of focus-sensitive particles in German. *Natural Language and Linguistic Theory*, 19:229–81.

Cardinaletti, Anna and Michal Starke. 1996. Deficient pronouns: a view from Germanic. In H. Thráinsson, S.D. Epstein, and S. Peter, editors, *Studies in Comparative Germanic Syntax II*. Kluwer, Dordrecht, pages 21–65.

Cedergren, Henrietta and David Sankoff. 1974. Variable rules: performance as a statistical reflection of competence. *Language*, 50:333–55.

CGN. 2004. Corpus Gesproken Nederlands, v1.0. Electronic Resource. See: `http://lands.let.ru.nl/cgn/home.htm`.

Comrie, Bernard. 2000. Pragmatic binding: demonstratives as anaphors in Dutch. In Matthew Juge and Jeri Moxley, editors, *Proceedings of the Twenty-Third Annual Meeting of the Berkeley Linguistics Society*, pages 50–61, Berkeley. BLS.

van Craenenbroeck, Jeroen and Liliane Haegeman. 2007. The derivation of subject-initial v2. *Linguistic Inquiry*, 38(1):167–78.

Dahl, Östen and Kari Fraurud. 1996. Animacy in grammar and discourse. In *Reference and Referent Accessibility*. John Benjamins, Amsterdam, pages 47–65.

van der Does, Jaap and Helen de Hoop. 1998. Type-shifting and scrambled definites. *Journal of Semantics*, 15(4):393–417.

Dryer, Matthew. 1995. Frequency and pragmatically unmarked word order. In Pamela Downing and Michael Noonan, editors, *Word Order in Discourse*. John Benjamins, Amsterdam, pages 105–37.

van Eynde, Frank. 1999. Major and minor pronouns in Dutch. In G. Bouma, E. Hinrichs, G.-J.M. Kruijff, and R. Oehrle, editors, *Constraints and Resources in Natural Language Syntax and Semantics*. CSLI Publications, pages 137–52.

van Eynde, Frank, 2003. *Protocol voor POS tagging en lemmatisering*.

van Eynde, Frank, Jakub Zavrel, and Walter Daelemans. 2000. Lemmatisation and morphosyntactic annotation for the Spoken Dutch Corpus. In Paola Monachesi, editor, *Computational Linguistics in the Netherlands 1999*, pages 53–62, Utrecht.

Faarlund, Jan Terje, Svein Lie, and Kjell Ivar Vannebo. 1997. *Norsk Referansegrammatikk*. Universitetsforlaget, Oslo.

Fanselow, Gisbert. 2000. Does constituent length predict German word order in the Middle Field? In J. Bayer and C. Römer, editors, *Von der Philologie zur Grammatiktheorie. Peter Suchsland zum 65. Geburtstag*. Niemeyer, Tübingen, pages 63–78.

Fanselow, Gisbert and Caroline Féry. 2002. Ineffability in grammar. In Gisbert Fanselow and Caroline Féry, editors, *Resolving Conflicts in Grammars: Optimality Theory in Syntax, Morphology, and Phonology*, volume 11 of *Linguistische Bereichte Sonderhefte*. Helmut Buske Verlag, Hamburg.

Féry, Caroline. 2006. The prosody of topicalization. In Kerstin Schwabe and Susanne Winkler, editors, *On Information Structure, Meaning and Form*. Benjamins, Amsterdam.

Flack, Kathryn. 2007. Ambiguity avoidance as contrast preservation: case and word order freezing in Japanese. In Leah Bateman, Michael O'Keefe, Ehren Reilly, and Adam Werle, editors, *UMass Occasional Papers in Linguistics 32: Papers in Optimality Theory III*. Booksurge Publishing, pages 57–89. Longer version available as ROA 748-0605.

Fleischer, Jürg. 2002. Preposition stranding in German dialects. In Sjef Barbiers, Leonie Cornips, and Susanne van der Kleij, editors, *Syntactic Microvariation. Online Proceedings – Workshop on Syntactic Microvariation*, pages 116–51, Amsterdam. Meertens Instituut. `http://www.meertens.knaw.nl/projecten/sand/synmic/pdf/fleischer.pdf`.

Flemming, Edward. 1995. *Auditory Representations in Phonology*. Ph.D. thesis, University of California, Los Angeles.

Frascarelli, Mara and Roland Hinterhölzl. 2007. Types of topics in German and Italian. In Kertin Schwabe and Susanne Winkler, editors, *On Information Structure, Meaning and Form*. John Benjamins, Amsterdam, pages 87–116.

Frey, Werner. 2005. Pragmatic properties of certain German and English left peripheral constructions. *Linguistics*, 43(1):89–129.

Frey, Werner. 2006. Contrast and movement to the German prefield. In *The Architecture of Focus*, Studies in Generative Grammar 82. Mouton de Gruyter, Berlin, pages 235–64.

Gärtner, Hans-Martin and Markus Steinbach. 2003. What do reduced pronominals reveal about the syntax of Dutch and German? Part 2: fronting. *Linguistische Bereichte*, 196:459–90.

Givón, Talmy. 1983. Topic continuity in discourse: An introduction. In Talmy Givón, editor, *Topic Continuity in Discourse: A Quantitative Crosslanguage Study*. John Benjamins, Amsterdam, pages 5–41.

Givón, Talmy. 1988. The pragmatics of word order: predictability, importance and attention. In M. Hammond, E. Moravczik, and J. Wirth, editors, *Studies in Syntactic Typology*. John Benjamins, Amsterdam, pages 243–85.

Grimshaw, Jane. 1997. Projection, heads, and optimality. *Linguistic Inquiry*, 28(3):373–422.

Grimshaw, Jane and Vieri Samek-Lodovici. 1998. Optimal subjects and subject universals. In *Is the Best Good Enough*. MIT Press, Cambridge, MA, pages 193–219.

Grohmann, Kleanthes. 2003. *Prolific Domains. On the Anti-Locality of Movement Dependencies*. John Benjamins, Amsterdam.

Gundel, Jeanette. 1974. *The Role of Topic and Comment in Linguistic Theory*. Ph.D. thesis, University of Texas at Austin.

Gundel, Jeanette. 1988. Universals of topic-comment structure. In M. Hammond, E. Moravczik, and J. Wirth, editors, *Studies in Syntactic Typology*. John Benjamins, Amsterdam, pages 209–39.

Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij, and M.C. van den Toorn. 1997. *Algemene Nederlandse Spraakkunst*. Martinus Nijhoff, 2nd edition.

Haider, Hubert. 1993. *Deutsche Syntax, generativ*. Günter Narr.

Hale, Mark and Charles Reiss. 1998. Formal and emperical arguments concerning phonological acquisition. *Linguistic Inquiry*, 29(4):656–83.

Harrell, F.E., K.L Lee, and D.B Mark. 1996. Tutorial in biostatistics, multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–87.

Harrell, Frank. 2003. Design: S functions for biostatistical/epidemiologic modeling, testing, estimation, validation, graphics, and prediction. Software. Programmes available from `biostat.mc.vanderbilt.edu/s/Design.html`.

't Hart, Johan. 1998. Intonation in Dutch. In Daniel Hirst and Albert Di Cristo, editors, *Intonation Systems. A Survey of Twenty Languages*. Cambrige University Press, pages 96–111.

Hawkins, John A. 1994. *A Performance Theory of Order and Constituency*. Cambridge University Press, Cambridge.

Hawkins, John A. 2004. *Efficiency and Complexity in Grammars*. Oxford University Press, Oxford.

Heck, Fabian, Gereon Muller, Ralf Vogel, Silke Fischer, Sten Vikner, and Tanya Schmid. 2002. On the nature of the input in Optimality Theory. *The Linguistic Review*, 19:345–76.

Heinze, Georg and Meinhard Ploner. 2003. Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. *Computer Methods and Programs in Biomedicine*, 71:181–187.

Hendriks, Petra and Helen de Hoop. 2001. Optimality Theoretic semantics. *Linguistics and Philosophy*, 24:1–32.

Herring, Susan. 1990. Information structure as a consequence of word order type. In *Proceedings of the 16th Annual Berkeley Linguistics Society (BLS-16)*, pages 163–74. Berkely Linguistics Society.

Heycock, Caroline and Roberto Zamparelli. 2003. Coordinated bare definites. *Linguistic Inquiry*, 34(3):443–69.

Heylen, Kris. 2005. A quantitative corpus study of German word order variation. In Stephan Kepser and Marga Reis, editors, *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*, pages 241–64, Berlin. Mouton de Gruyter.

Hoeksema, Jack. 2000. Verplichte topicalisatie van kale enkelvouden: de feit is dat-constructie. *TABU*, 30.

Hoekstra, H., M. Moortgat, B. Renmans, M. Schouppe, I. Schuurman, and T. van der Wouden, 2003. *CGN syntactische annotatie*.

Hoekstra, Heleen, Michael Moortgat, Ineke Schuurman, and Ton van der Wouden. 2001. Syntactic annotation for the spoken Dutch corpus project (CGN). In Walter Daelemans, Khalil Sima'an, Jorn Veenstra, and Jakub Zavrel, editors, *Computational Lintuistics in the Netherlands 2000*. Rodopi, Amsterdam, pages 73–87.

Höhle, Tilman N. ms. Topologische felder. Cologne 1983. Electronic version 2003, to be found at `http://www.linguistik.uni-tuebingen.de/hoehle/manuskripte/Topologische_Felder.pdf`.

van Hoof, Hanneke. 1997. Left dislocation and split topics in Brabant Dutch. In Elena Anagnostopoulou, Henk van Riemsdijk, and Frans Zwarts, editors, *Materials on Left Dislocation*, volume 14 of *Linguistik Aktuell*. Benjamins, Amsterdam/Philadelpia, pages 275–306.

van Hoof, Hanneke. 2003. The rise in the rise-fall contour: does it evoke a contrastive topic or a contrastive focus? *Linguistics*, pages 515–63.

de Hoop, Helen and Irene Krämer. 2006. Optimal interpretations of indefinite subjects and objects. *Language Acquisition*, 13:103–23.

de Hoop, Helen. 2003. Optionality and optimality. In *Word Order and Scrambling*. Blackwell Publishers, Oxford, pages 201–17.

Huesken, Nicole. 2001. Mirrorsentences. repetition of inflected verb and subject in spoken Dutch. Master's thesis, Utrecht University, General Linguistics.

Jacobs, Joachim. 2001. The dimensions of topic-comment. *Linguistics*, 39(4):641–81.

Jäger, Gerhard. 2002. Some notes on the formal properties of bidirectional Optimality Theory. *Journal of Logic, Language and Information*, 11(4):427–51.

Jäger, Gerhard. 2004. Learning constraint sub-hierarchies. the Bidirectional Gradual Learning Algorithm. In Reinhard Blutner and Henk Zeevat, editors, *Pragmatics in OT*. Palgrave MacMillan, Hampshire, pages 251–87.

Jakobson, Roman. 1936. Beitrag zur allgemeinen Kasuslehre. Gesamtbedeutungen der russischen Kasus. In *Travaux du Cercle Linguistique de Prague 6*, pages 240–88. Consulted in Word and Language, volume 2 of Selected Writings, 1971, Mouton, Den Haag/Paris, pages 23–72.

Jansen, Frank. 1981. *Syntaktische Konstrukties in Gesproken Taal*. Huis aan de drie grachten, Amsterdam.

Jansen, Frank and Raoul Wijnands. 2004. Doorkruisingen van het links-rechtsprincipe. *Neerlandistiek.nl*.

Kaan, Edith. 1997. *Processing Subject-Object Ambiguities in Dutch*. Ph.D. thesis, University of Groningen.

Kaan, Edith. 1999. Sensitivity to NP-type: processing subject-object ambiguities in Dutch. *Journal of Semantics*, 15(4):335–54.

Kaan, Edith. 2001. Subject-object order ambiguities and the nature of the second NP. *Journal of Psycholinguistic Research*, 30(5):527–45.

Kadmon, Nirit. 2001. *Formal Pragmatics: Semantics, Pragmatics, Presupposition, and Focus*. Blackwell.

Kaiser, Elsi and John Trueswell. 2004. The referential properties of Dutch pronouns and demonstratives: is salience enough? In Jürgen Konradi, Cécile Meier, and Matthias Weisgerber, editors, *Proceedings of 8th Annual Conference of the Gesellschaft für Semantik (SuB 8)*, pages 137–51, Frankfurt-am-Main.

Kathol, Andreas. 2000. *Linear Syntax*. Oxford University Press, Oxford.

Kempen, Gerard and Karin Harbusch. 2004. A corpus study into word order variation in German subordinate clauses: animacy affects linearization independently of grammatical function assignment. In T. Pechmann and C. Habel, editors, *Multidisciplinary Approaches to Language Production*. Mouton De Gruyter, Berlin, pages 173–181.

King, Tracy Holloway. 1995. *Configuring Topic and Focus in Russian*. Dissertations in Linguistics. CLSI Publication, Stanford.

Kruisinga, Etsko. 1938. *Het Nederlands van Nu*. Wereldbibliotheek, Amsterdam. Consulted 2nd edition, with revisions and additional material by Herman Godthelp, 1951, Wereldbibliotheek, Amsterdam.

Kuhn, Jonas. 2003. *Optimality-Theoretic Syntax—A Declarative Approach*. CSLI Publications, Stanford, CA.

Kunkel-Razum, Kathrin and Franziska Münzberg, editors. 2005. *Die Grammatik*, volume 4 of *Duden*. Bibliographisches Institut, Mannheim, 7 edition.

Kuno, Susume. 1980. A note on Tonoike's intra-subjectivization hypothesis *and* A further note on Tonoike's intra-subjectivization hypothesis. In Y. Otsu and A. Farmer, editors, *Theoretical Issues in Japanese Linguistics (MWPL 2)*. MIT Working Papers in Linguistics, pages 149–57, 171–85.

Kurz, Daniela. 2000a. A statistical account on word order variation in German. In *Proceedings of the COLING Workshop on Linguistically Interpreted Corpora*, pages 241–64.

Kurz, Daniela. 2000b. Wortstellungspräferenzen im Deutschen. Master's thesis, Saarland University, Saarbrücken.

Kuthy, Kordula De. 2002. *Discontinuous NPs in German—A Case Study of the Interaction of Syntax, Semantics and Pragmatics*. Studies in Constraint-Based Lexicalism. CSLI Publications, Stanford.

Lamers, Monique. 2001. *Sentence Processing: Using Syntactic, Semantic, and Thematic Information*. Ph.D. thesis, Rijksuniversiteit Groningen.

Lee, Hanjung. 2001a. Markedness and word order freezing. In Peter Sells, editor, *Formal and Emperical Issues in Optimality Theoretic Syntax*, volume 5 of *Studies in Constraint based Lexicalism*. CSLI Publications, Stanford.

Lee, Hanjung. 2001b. *Optimization in Argument Expression and Interpretation: A Unified Approach.* Ph.D. thesis, Stanford University.

Lee, Hanjung. 2002. Crosslinguistic variation in argument expression and intralinguistic freezing effects. In T. Ionin, H. Ko, and A. Nevins, editors, *Proceedings of the 2nd HUMIT Student Conference in Language Research (HUMIT 2001)*. MIT Working Papers in Linguistics, pages 103–23.

Lee, Hanjung. 2004. Minimality in a lexicalist Optimality Theory. In A. Stepanov, G. Fanselow, and R. Vogel, editors, *Minimality Effects in Syntax*. Mouton de Gruyter, Berlin, pages 241–89.

Lee, Hanjung. ms. Bidirectional optimality and ambiguity in argument expression. Stanford, 2000. Extended version of a paper given at LFG2000 at Berkeley. Available from `www.stanford.edu/~hanjung` through Archive.org.

Legendre, Géraldine. 2001. Introduction to Optimality Theory in syntax. In Géraldine Legendre, Jane Grimshaw, and Sten Vikner, editors, *Optimality-Theoretic Syntax*. MIT Press, Cambridge, MA, pages 1–29.

Legendre, Géraldine, Colin Wilson, Paul Smolensky, Kristin Homer, and William Raymond. 1995. Optimality and wh-extraction. In S. J. Beckman, Urbanczyk, and L. Walsh, editors, *University of Massachusetts Occasional Papers in Linguistics 18: Papers in Optimality Theory*. University of Massachusetts, Amherst, MA, pages 607–36. Also as: ROA-85.

Lenerz, Jürgen. 1977. *Zur Abfolge Nominal Satzglieder im Deutschen*. Studien zur deutschen Grammatik 5. TBL Verlag Gunter Narr, Tübingen.

Lenerz, Jürgen. 1994. Pronomenprobleme. In Brigitte Haftka, editor, *Was Determiniert Wortstellungsvariation?* Westdeutscher Verlag, Opladen, pages 161–73.

McCarthy, John J. and Alan Prince. 1994. The emergence of the unmarked: optimality in prosodic morphology. In Mercé González, editor, *Proceedings of the North East Linguistics Society 24*, pages 333–79, Amherst, MA.

McNally, Louisa. 1998. On recent formal analyses of topic. In J. Ginzburg, Z. Khasidashvili, C. Vogel, J.J. Lévy, and E. Vallduví, editors, *The Tbilisi Symposium on Language Logic and Computation: Selected Papers*. CSLI Publications, Stanford, CA, pages 147–60.

Meinunger, André. 2004. On certain adverbials in the German 'Vorfeld' and 'Vorvorfeld'. In *Sprache und Pragmatik, Lunder germanistische Forschungen*, pages 64–78.

Meinunger, André. 2007. About object *es* in the German *vorfeld*. *Linguistic Inquiry*, 38(3):553–63.

Meurers, Walt Detmar. 2005. On the use of electronic corpora for theoretical linguistics. case studies from the syntax of German. *Lingua*, 115(11):1619–39.

Mikkelsen, Line. 2002. Reanalyzing the definiteness effect: evidence from Danish.

Mohanan, K.P. and Tara Mohanan. 1994. Issues in word order in South Asian languages: enriched phrase structure or multidimensionality. In Miriam Butt, Tracy King, and Gillian Ramchand, editors, *Theoretical Perspectives on Word Order in South Asian Languages*. CSLI Publications, Stanford, pages 153–84.

Morimoto, Yukiko. ms. 'crash vs yield': On the conflict asymmetry in syntax and phonology. Manuscript Stanford University, 2000.

Müller, Gereon. 1999. Optimality, markedness, and word order in German. *Linguistics*, 37:777–818.

Müller, Gereon. 2001. Optionality in optimality-theoretic syntax. In Lisa Cheng and Rint Sybesma, editors, *The Second Glot International State-of-the-Article Book*. Mouton, Berlin, pages 289–321.

Müller, Gereon. 2002. Free word order, morphological case, and Sympathy Theory. In Gisbert Fanselow and Caroline Fery, editors, *Resolving Conflicts in Grammars*. Buske, Hamburg, pages 9–48.

Müller, Stefan. 2005. Zur Analyse der scheinbar mehrfachen Vorfeldbesetzung. *Linguistische Berichte*, 203:29–62.

Odijk, Jan. 1998. Topicalization of non-extraposed complements in Dutch. *Natural Language and Linguistic Theory*, 16(1):191–222.

Øvrelid, Lilja. 2004. Disambiguation of syntactic functions in norwegian: modeling variation in word order interpretations conditioned by animacy and definiteness. In Fred Karlsson, editor, *Proceedings of the 20th Scandinavian Conference of Linguistics*, Helsinki.

Philippa, Marlies, Frans Debrabandere, and Arend Quak, editors. 2003-9. *Etymologisch Woordenboek van het Nederlands*. Amsterdam University Press. Consulted Web Edition, August 2006.

Pollard, Carl and Ivan Sag. 1994. *Head-driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

Prince, Alan and Paul Smolensky. 1993/2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell. As Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, and Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ, April 1993. Rutgers Optimality Archive 537 version, 2002.

Prince, Ellen. 1998. On the limits of syntax, with reference to Left-dislocation and Topicalization. In Peter Culicover and Louisa McNally, editors, *Syntax and semantics*, volume 29. Academic Press, New York, pages 281–302.

R Development Core Team, 2006. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. See: `http://www.r-project.org`.

Reinhart, Tanya. 1982. *Pragmatics and Linguistics: An Analysis of Sentence Topics*. Indiana University Linguistics Club, Bloomington, Indiana.

van Riemsdijk, Henk. 1978. *A Case Study in Syntactic Markedness: The Binding Nature of Prepositional Phrases*. Floris, Dordrecht.

van Riemsdijk, Henk and Frans Zwarts. 1997. Left dislocation in Dutch and the status of copying rules. In Elena Anagnostopoulou, Henk van Riemsdijk, and Frans Zwarts, editors, *Materials on Left Dislocation*, volume 14 of *Linguistik Aktuell*. Benjamins, Amsterdam/Philadelpia, pages 13–31.

Rietveld, Toni and Roeland van Hout. 1993. *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter, Berlin-New York.

Roberts, Craige. 1996. Information structure in discourse: towards an integrated formal theory of pragmatics. In Jae-Hak Yoon and Andreas Kathol, editors, *OSU Working Papers in Linguistics 49: Papers in Semantics*. Department of Linguistics of Ohio State University, pages 91–136.

Rosenbach, Anette. 2005. Animacy versus weight as determinants of grammatical variation in English. *Language*, 81(3):613–44.

Salverda, Reinier. 2000. On topicalization in Modern Dutch. In Thomas F. Shannon and Johann P. Snapper, editors, *The Berkeley Conference on Dutch Linguistics 1997: Dutch Linguistics at the Millenium*. University Press of America, Lanham, MD, pages 93–111.

Samek-Lodovici, Vieri. 1996. *Constraints on Subjects. An Optimality Theoretic Analysis*. Ph.D. thesis, Rutgers University.

Schelfhout, Carla. 2005. *Intercalatations in Dutch*. Ph.D. thesis, Radboud Universiteit Nijmegen.

Shannon, Thomas F. 2000. On the order or (pro)nominal arguments in Dutch and German. In Thomas F. Shannon and Johann P. Snapper, editors, *The Berkeley Conference on Dutch Linguistics 1997: Dutch linguistics at the millenium*. University Press of America, Lanham, MD, pages 145–95.

SICS. 2005. Sicstus Prolog, v3. Software. See `http://www.sics.se/isl/sicstuswww/site/`.

van der Sijs, Nicoline. 2004. *Taal als Mensenwerk: het ontstaan van het ABN*. Sdu, Den Haag.

Skut, Wojciech, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP)*, pages 88–95, Washington, D.C.

Snider, Neal. 2005. A corpus study of left dislocation and topicalization. PhD Qualifying Paper. Stanford University.

Snider, Neal and Annie Zaenen. 2006. Animacy and syntactic structure: fronted NPs in English. In Miriam Butt, Mary Dalrymple, and Tracy Holloway King, editors, *Intelligent Linguistic Architectures: Variations on Themes by Ron Kaplan*. CSLI Publications, Stanford, CA.

Steedman, Mark. 2000. Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31(4):649–89.

Teleman, Ulf, Staffan Hellberg, and Erik Andersson. 1999. *Svenska Akademiens Grammatik*. Norstedts, Stockholm.

Thrift, Erica. 2003. *Object Drop in the L1 acquisition of Dutch*. Ph.D. thesis, Universiteit van Amsterdam.

TiGer. 2003. Tiger search, v2.1. Software. See `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/`.

Tonoike, Sigeo. 1980. Intra-subjectivization *and* More on intra-subjectivization. In Y. Otsu and A. Farmer, editors, *Theoretical Issues in Japanese Linguistics (MWPL 2)*. MIT Working Papers in Linguistics, pages 136–48, 157–71.

Travis, Lisa. 1984. *Parameters and Effects of Word Order Variation*. Ph.D. thesis, MIT.

Uszkoreit, Hans. 1987. *Word Order and Constituent Order in German*. CLSI Lecture Notes 8. CSLI, Stanford, CA.

Vallduví, Enric and Elisabet Engdahl. 1996. The linguistic realization of information packaging. *Linguistics*, 34:459–519.

Vogel, Ralf. 2004. Remarks on the architecture of Optimality Theoretic syntax grammars. In R. Blutner and H. Zeevat, editors, *Optimality Theory and Pragmatics*. Palgrave Macmillan, Hampshire, pages 211–28.

de Vogelaer, Gunther. 2005. *Subjectsmarkering in de Nederlandse en Friese Dialecten*. Ph.D. thesis, Universiteit Gent.

de Vries, Wobbe. 1911. Dysmelie, opmerkingen over syntaxis. Verhandelingen bij het programma van het Groningsch Gymnasium / Course material of the Groningsch Gymnasium, Groningen.

Ward, Gregory. 1988. *The Semantics and Pragmatics of Preposing*. Outstanding Dissertations in Linguistics. Garland, New York.

Wasow, Tom. 2002. *Postverbal Behavior*. CSLI Publications.

Weber, Andrea and Karin Müller. 2004. Word order variation in German main clauses: a corpus analysis. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*, Geneve.

Weerman, Fred. 1989. *The V2 Conspiracy*. Floris, Dordrecht.

Wilson, Colin. 2001. Bidirectional optimization and the theory of anaphora. In Géraldine Legandre, Jane Grimshaw, and Sten Vikner, editors, *Optimalitiy Theoretic Syntax*. MIT Press, Cambridge, MA.

van der Wouden, Ton. 2002. Partikels: naar een partikelwoordenboek voor het Nederlands. *Nederlandse Taalkunde*, 7(1):20–43.

van der Wouden, Ton, Heleen Hoekstra, Michael Moortgat, Bram Renmans, and Ineke Schuurman. 2002. Syntactische annotatie voor het Corpus Gesproken Nederlands (CGN). *Nederlandse Taalkunde*, 7(4):335–52.

Yamashita, Hirohiko. 2002. Scrambled sentences in Japanese: linguistic properties and motivations for production. *Text*, 22(4):587–633.

Zaenen, Annie. 1997. Contrastive dislocation in Dutch and Icelandic. In Elena Anagnostopoulou, Henk van Riemsdijk, and Frans Zwarts, editors, *Materials on Left Dislocation*, volume 14 of *Linguistik Aktuell*. Benjamins, Amsterdam/Philadelpia, pages 119–50.

Zaenen, Annie, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, Mary C. O'Connor, and Tom Wasow. 2004. Animacy encoding in English: why and how. In *Proceedings of the ACL 2004 Workshop on Discourse Annotation*, pages 118–25, Barcelona.

Zeevat, Henk. 2000. The asymmetry of optimality theoretic syntax and semantics. *Journal of Semantics*, 17(3):243–62.

Zeevat, Henk. 2006. Freezing and marking. *Linguistics*, 44(5):1095–111.

Zeevat, Henk and Gerhard Jäger. 2002. A reinterpretation of syntactic alignment. In Dick de Jongh, Henk Zeevat, and Maria Nilsenova, editors, *Proceedings of the 3rd and 4th International Symposium on Language, Logic and Computation*, Amsterdam. ILLC.

Zerbian, Sabine. 2007. Subject/object-asymmetry in Nothern Sotho. In Kerstin Schwabe and Susanne Winkler, editors, *On Information Structure, Meaning and Form*, Linguistik Aktuell 100. John Benjamins, Amsterdam, pages 323–347.

Zwart, C. Jan-Wouter. 1997. *Morphosyntax of Verb Movement*, volume 39 of *SNLT*. Kluwer Academic Publishers, Dordrecht.

Samenvatting

In een bekende televisiereclame uit het verleden kwamen de volgende regels voor:

(1) a. Koning, keizer, admiraal,
b. Popla kennen ze allemaal!

In zin (1b) staat het lijdend voorwerp *Popla* (een toiletpapiermerk) direct vóór het finiete werkwoord. In de traditionele beschrijvende taalkunde staat deze positie onder andere bekend onder de Duitse term *Vorfeld*.

Het Nederlands kan beschreven worden als een taal met *verb-second, verb-final*: in een bewerende hoofdzin komt het finiete werkwoord op de tweede plaats, en eventuele andere werkwoorden in een cluster tegen het eind van de zin. Afgezien van deze vaste polen laat het Nederlands een zekere variatie toe in de woordvolgorde. Zo is er grote vrijheid in de keuze voor de woordgroep die op de eerste plaats komt, de *Vorfeldbezetter*. De schrijvers van de Popla-reklame hadden bijvoorbeeld de betekenis van (1b) ook uit kunnen drukken met de dichtregel:

(1) b′. Ze kennen Popla allemaal!

In (1b′) wordt het Vorfeld bezet door het onderwerp *ze*. In dit proefschrift probeer ik de vraag te beantwoorden wat de keuze bepaalt tussen twee Vorfeldvarianten, zoals (1b) en (1b′).

Er zijn betrekkelijk weinig grammaticale beperkingen in de keuze voor een Vorfeldbezetter, en de hierboven gegeven varianten illustreren maar twee van de vele mogelijkheden. In dit proefschrift houd ik me echter niet zozeer bezig met de vraag wat er grammaticaal allemaal mogelijk is, maar veeleer met de vraag welke syntactische, discours-semantische en communicatieve overwegingen voor een spreker belangrijk zijn in de vooropplaatsing van een zinsdeel. Een antwoord op deze vraag wordt gegeven aan de hand van een studie van vooropplaatsing van onderwerpen, lijdend voorwerpen en meewerkend voorwerpen in gesproken Nederlands. De bestudeerde Vorfeldvarianten zijn dus als in (2).

(2) a **Ik** heb jou dat verteld.
b **Jou** heb ik dat verteld.
c **Dat** heb ik jou verteld.

Met behulp van een combinatie van theoretische modellering en corpusonderzoek wordt een aantal van de factoren in de keuze tussen (2a), (2b) en (2c) duidelijk gemaakt. Als corpus worden ongeveer zestigduizend zinnen uit het Corpus Gesproken Nederlands (CGN, 2004) gebruikt.

Sprekersvoorkeuren in vooropplaatsing

In de eerste helft van het proefschrift wordt de invloed van drie factoren op vooropplaatsing bestudeerd: de grammaticale functie van een zinsdeel, zijn bepaaldheid (in het geval van naamwoordelijke zinsdelen), en zijn grammaticale complexiteit. Deze factoren, die goed te bestuderen zijn in een syntactisch geannoteerd corpus, zijn bekend uit onderzoek naar woordvolgorde in het Duitse en Nederlandse *Mittelfeld* (tussen het finiete werkwoord in tweede positie en de niet-finiete werkwoorden aan het eind van de zin) en zinsdeelvolgorde in het algemeen in het Duits, Nederlands en Engels (Hoofdstuk 2). Een belangrijke deelvraag in dit onderzoek is dan ook in hoeverre Vorfeldbezetting onderhevig is aan dezelfde globale tendensen als de zinsdeelvolgorde in het Mittelfeld.

1. Grammaticale functie De canonieke volgorde van zinsdelen in het Nederlandse Mittelfeld is onderwerp voor voorwerp, en meewerkend voorwerp voor lijdend voorwerp. Wat betreft Vorfeldbezetting kunnen we het volgende opmaken uit het corpus (Hoofdstuk 4): onderwerpen worden vaker vooropgeplaatst dan voorwerpen, en hoewel het verschil tussen de voorwerpen klein is, lijkt het erop dat meewerkend voorwerpen vaker vooropgeplaatst worden dan lijdend voorwerpen. Hierbij moet echter worden opgemerkt dat er te weinig meewerkend-voorwerpsdata beschikbaar is voor een solide resultaat.

We kunnen de invloed van de grammaticale functie van een zinsdeel op de kans dat het vooropgeplaatst wordt, samenvatten als in (3), waarbij '$\prec$' gelezen wordt als 'neigt sterker naar Vorfeldbezetting'.

(3) *Positieve relatie tussen grammaticale functie en Vorfeldbezetting:*
onderwerp $\prec$ meewerkend voorwerp $\preceq$ lijdend voorwerp

De ordening van grammaticale functies in (3) kennen we van de canonieke zinsdeelvolgorde in het Mittelfeld. We kunnen daarom voorzichtig vaststellen dat Vorfeldbezetting door onderwerpen en voorwerpen een afspiegeling is van zinsdeelvolgorde in het Mittelfeld. Dit maakt het aannemelijk dat grammaticale functie een algemene invloed heeft op woordvolgorde, onafhankelijk van het domein in de zin.

2. Bepaaldheid De relatie tussen bepaaldheid en Vorfeldbezetting is nogal ingewikkeld. Op basis van bestaande resultaten over de zinsdeelvolgorde in het Nederlandse en Duitse Mittelfeld kunnen we volgende bepaaldheidshiërarchie opstellen: voornaamwoorden, bepaalde naamwoordsgroepen, onbepaalde naamwoordsgroepen. In het Mittelfeld bestaat er de neiging om elementen die hoger in deze hiërarchie staan, eerder uit te spreken.

Deze Mittelfeldtendensen zijn gedeeltelijk terug te vinden in de Vorfeldbezettingsdata (Hoofdstuk 4). Bepaalde (niet-voornaamwoordelijke) naamwoordsgroepen worden vaker vooropgeplaatst dan onbepaalde naamwoordsgroepen. In het geval van de voornaamwoordelijke zinsdelen wijkt Vorfeldbezetting echter af van de Mittelfeldvolgorde. Het is bekend dat gereduceerde persoonlijk voornaamwoordelijke voorwerpen niet in het Vorfeld mogen staan. Uit de corpusanalyse blijkt dat de kans dat een gereduceerd persoonlijk voornaamwoord in het Vorfeld gezet wordt, zeer klein is, onafhankelijk van grammaticale functie. In het algemeen is het zelfs zo dat de gemiddelde kans dat een persoonlijk voornaamwoord (gereduceerd of niet, onderwerp of voorwerp) vooropgeplaatst wordt kleiner is dan de kans dat een bepaalde naamwoordsgroep vooropgeplaatst wordt. Dit is niet in lijn met de bepaaldheidshiërarchie. Daar staat echter tegenover dat er een zeer sterke neiging bestaat om aanwijzende voornaamwoorden in het Vorfeld te plaatsen. Deze worden veel frequenter vooropgeplaatst dan bepaalde naamwoordsgroepen.

We kunnen de zojuist geschetste conflicterende tendensen in Vorfeldbezetting samenvatten als (4) en (5). Merk op dat de eerste categorie in (5) het Vorfeld het sterkst vermijdt.

(4) *Positieve relatie tussen bepaaldheid en Vorfeldbezetting:*
voornaamwoorden $\prec$ bepaalde naamwoordsgroepen $\prec$ onbepaalde naamwoordsgroepen

(5) *Negatieve relatie tussen voornaamwoordstype en Vorfeldbezetting*
gereduceerd, persoonlijk $\succ$ vol, persoonlijk $\succ$ aanwijzend

Het negatieve effect op vooropplaatsing van voornaamwoordstype is in het geval van persoonlijke voornaamwoorden sterker dan het positieve effect van de hoge positie op de bepaaldheidshiërarchie.

De trends die samengevat zijn in (4) en (5), zijn terug te vinden in zowel de onderwerpsdata als de voorwerpsdata. De verschillen tussen de graden van bepaaldheid en de voornaamwoordstypen zijn echter groter in de onderwerpsdata.

Een mogelijke verklaring van de negatieve relatie tussen voornaamwoordstype en Vorfeldbezetting wordt gegeven door een universeel woordvolgordeprincipe voorgesteld door (Gundel, 1988), het zogenaamde *first-things-first* principe. Volgens dit principe is er een universele tendens om belangrijke informatie het eerst uit te spreken. Onder belangrijke informatie wordt hier informatie verstaan die nieuw, contrasterend of onverwacht is. Persoonlijke voornaamwoorden, en vooral gereduceerde persoonlijke voornaamwoorden,

worden typisch gebruikt om informatie uit te drukken die in zeer hoge mate voorspelbaar is, en dus niet belangrijk. Als we het Vorfeld als positie voor belangrijk materiaal beschouwen, is het niet verrassend dat voornaamwoorden niet vaak in deze positie gezet worden.

Het feit dat we de invloed van naamwoordelijke bepaaldheid op Vorfeldbezetting uiteen kunnen splitsen in (5) en (4), suggereert dat de bepaaldheidshiërarchie, net als grammaticale functie, een algehele invloed op woordvolgorde heeft, onafhankelijk van het domein in de zin. De differentiatie tussen Vorfeldbezetting en Mittelfeldwoordvolgorde komt dan voort uit het feit dat het first-things-first principe alleen van invloed is in het Vorfeld.

3. Grammaticale complexiteit Een niet ongewone aanname is dat er een algemene tendens bestaat om constituenten die grammaticaal minder complex of korter zijn, eerder uit te spreken dan constituenten die complexer of langer zijn. Deze aanname is al terug te vinden is bij Behaghel (1909). In de Algemene Nederlandse Spraakkunst (Haeseryn et al., 1997) wordt deze aanname het *complexiteitsprincipe* genoemd. Aangezien het Vorfeld een zeer vroege positie in de zin is, wekt het complexiteitsprincipe de verwachting dat Vorfeldmateriaal relatief kort is. Analyse van corpusdata bevestigt dit: zinsdelen in het Vorfeld zijn gemiddeld korter dan zinsdelen die niet in het Vorfeld staan.

We kunnen hiermee echter niet vaststellen dat er een directe invloed van complexiteit op Vorfeldbezetting bestaat. Nadere inspectie van de data laat namelijk zien dat complexiteit voornamelijk een rol speel in de vraag of een zinsdeel aan de *rechter*kant van een zin staat of niet. Het waargenomen verschil in lengte tussen Vorfeldzinsdelen en andere zinsdelen wordt veroorzaakt door het feit dat geen enkele van de Vorfeldzinsdelen aan de rechterkant van een zin staat, maar een deel van de overige zinsdelen wel. Toch kunnen we ook hier een overeenkomst tussen Mittelfeldvolgorde en Vorfeldbezetting zien: ook in het Mittelfeld lijkt er geen algemeen effect van complexiteit op zinsdeelvolgorde te bestaan.

De drie factoren grammaticale functie, bepaaldheid en grammaticale complexiteit spelen dus ieder een rol in de keuze voor een Vorfeldbezetter. Kort gezegd: subjecten, aanwijzende voornaamwoorden en eenvoudige zinsdelen zijn goede Vorfeldbezetters, maar lijdend voorwerpen, onbepaalde naamwoordsgroepen, gereduceerde persoonlijke voornaamwoorden en complexe zinsdelen niet. De ingewikkelde patronen in Vorfeldbezetting die we in het corpus kunnen waarnemen zijn op zijn minst een combinatie van globale tendensen in zinsdeelvolgorde (grammaticale functie en de bepaaldheidshiërarchie), een lokale factor met betrekking tot het Vorfeld (belangrijk materiaal), en een lokale factor met betrekking tot de rechterkant van een zin (grammaticale complexiteit).

Het belang van het toehoordersperspectief

Het onderzoek heeft tot dusver impliciet de invalshoek van de spreker gekozen. Een van de opgaven van de spreker is om zinsdelen te ordenen in een zin, voordat de zin

uitgesproken wordt. Een keuze die de spreker daarbij moet maken is welk zinsdeel in het Vorfeld komt te staan. We weten nu dat zinsdeeleigenschappen als grammaticale functie, bepaaldheid en grammaticale complexiteit een rol spelen in hoe aantrekkelijk een zinsdeel is als Vorfeldbezetter voor een spreker.

In de statistische modellen die gebruikt worden in dit proefschrift nemen we de sprekerspositie zelfs vrij expliciet in: de modellen voorspellen voor een zinsdeel wat de kans is dat het in het Vorfeld terecht komt op basis van de eigenschappen grammaticale functie, bepaaldheid en grammaticale complexiteit.

Er spelen in de keuze voor een Vorfeldbezetter echter ook factoren een rol die niet zo eenvoudig met behulp van eigenschappen van de afzonderlijke zinsdelen te formuleren zijn. Dit wordt duidelijk als we de gevolgen van vooropplaatsing voor de toehoorder in overweging nemen. Een van taken van de toehoorder bij het interpreteren van een zin is om aan elk van de zinsdelen een grammaticale functie toe te wijzen. Aangezien grammaticale functie door de toehoorder afgeleid moet worden uit ander informatie, en aangezien het zinsdeel in het Vorfeld verschillende grammaticale functies kan hebben, kan het zijn dat de toehoorder de verkeerde grammaticale functie toewijst aan de Vorfeldbezetter, en daarmee de zin verkeerd begrijpt.

Voor diverse talen met een vrije woordvolgorde, bijvoorbeeld Duits, Russisch, Hindi en Japans, is beweerd dat woordvolgordevariatie ingeperkt wordt of zelfs uitgesloten is wanneer er niet gegarandeerd kan worden dat een toehoorder de juiste grammaticale functies aan de juiste zinsdelen toewijst. Dit kan bijvoorbeeld optreden in talen met een uitgebreid naamvalssysteem wanneer twee of meer zinsdelen dezelfde naamval of niet van elkaar te onderscheiden naamvallen krijgen. In deze gevallen is de canonieke zinsdeelvolgorde de enige toegelaten volgorde. Dit verlies van de mogelijkheid van woordvolgordevariatie wordt *woordvolgordebevriezing* genoemd. Voor een toehoorder betekent woordvolgordebevriezing dat in specifieke gevallen grammaticale functie direct kan worden afgeleid van de zinsdeelvolgorde. Voor een spreker betekent woordvolgordebevriezing dat er minder vrijheid is in het formuleren van een zin.

Het is op basis van intuïtiedata aannemelijk dat woordvolgordebevriezing ook een rol speelt in Vorfeldbezetting. Neem het voorbeeld in (6).

(6) Jan heeft Piet aan de kant gezet
Voorkeursinterpretatie: Piet is gedumpt door Jan.
Onwaarschijnlijke interpretatie: Jan is gedumpt door Piet.

Het is in principe mogelijk is dat degene die (6) uitspreekt, ervoor gekozen heeft om het lijdend voorwerp op de eerste plaats te zetten. *Jan* zou, wat positie in de zin en wat morfologische eigenschappen betreft, zowel het onderwerp als het lijdend voorwerp van (6) kunnen zijn. Toch is de zin niet duidelijk ambigu en ligt de voorwerpsinitiële interpretatie niet erg voor de hand. Het is wel mogelijk om deze wat meer naar voren

te halen met behulp van context of intonatie. De onderwerpsinitiële interpretatie is echter vanzelf aanwezig. Het lijkt erop dat de beschikbaarheid van de voorwerpsinitiële interpretatie afhankelijk is van de herkenbaarheid van de grammaticale functie van zinsdelen. Ofwel: Vorfeldbezetting in het Nederlands zou wel eens onderhevig kunnen zijn aan woordvolgordebevriezing.

Voortbouwend op eerder werk van onder andere Lee (2001b), stel ik in de tweede helft van het proefschrift dat woordvolgordebevriezing het beste met een grammaticaal model beschreven kan worden dat zowel het sprekers- als het toehoordersperspectief omvat (Hoofdstuk 5). De formalisering van deze verschillende perspectieven is mogelijk in zogenaamde *bidirectionele optimalititeitstheoretische modellen*. In een dergelijk bidirectioneel model worden de keuzemogelijkheden van een spreker begrensd door de vereiste dat een toehoorder de zinsdelen de juiste grammaticale functie toe kan kennen. Hiervoor heeft de toehoorder een reeks aan verschillende informatiebronnen tot zijn beschikking. Toegepast op de keuze voor een Vorfeldbezetter kan het in het model zo zijn dat een voorwerp niet in het Vorfeld gezet kan worden, ookal is het door zijn eigenschappen voor de spreker een zeer aantrekkelijke Vorfeldbezetter, omdat het door de toehoorder niet als voorwerp herkend zou worden, maar als onderwerp.

Het bidirectionele model van Vorfeldbezetting kan getoetst worden met behulp van het corpus door de discrete voorspellingen die model maakt, een kwantitatieve interpretatie te geven (Hoofdstuk 6). Zo voorspelt het model dat een lijdend voorwerp niet vóór het onderwerp komt, en daarmee niet in het Vorfeld, wanneer er geen woordvolgordeonafhankelijke gronden zijn waarop het onderwerp en lijdend voorwerp van elkaar te onderscheiden zijn. In het model interpreteren toehoorders zinsdelen onder andere aan de hand van bepaaldheid en animaatheid. Onderwerpen zijn bij voorkeur bepaald en nog liever voornaamwoordelijk, lijdend voorwerpen bij voorkeur onbepaald. Evenzo verwijzen onderwerpen bij voorkeur naar levende dingen, en lijdend voorwerpen naar levenloze dingen. Onder een kwantitatieve interpretatie van het bidirectionele model kunnen we verwachten dat vooropplaatsing van het lijdend voorwerp frequenter is wanneer het onderwerp en lijdend voorwerp op basis van deze voorkeuren te herkennen zijn (*positieve effecten van relatieve bepaaldheid en animaatheid*). Het positieve effect van relatieve bepaaldheid is inderdaad te vinden als we de zinnen met een onderwerp en een lijdend voorwerp uit het corpus statistisch analyseren. Een analyse van een kleinere verzameling zinnen waarin de zinsdelen voor animaatheid geannoteerd werden, wijst erop dat het positieve effect van relatieve animaatheid ook te vinden is. Onder de kwantitatieve interpretatie van het bidirectionele model is het ook aannemelijk dat vooropplaatsing van het lijdend voorwerp nóg minder frequent is wanneer relatieve bepaaldheid en animaatheid niet alleen niet helpen, maar een toehoorder juist op het verkeerde been zetten (negatief effecten van relatieve bepaaldheid en animaatheid). Deze negatieve effecten konden echter niet vastgesteld worden.

De positieve effecten van relatieve bepaaldheid en animaatheid op vooropplaatsing van lijdend voorwerpen, afgeleid van het bidirectionele model van vooropplaatsing, werden gevonden bovenop de factoren bepaaldheid en grammaticale complexiteit die in de vorige sectie beschreven werden. Samenvattend kunnen we zeggen dat een spreker van het Nederlands minder gebruikt maakt van niet-canonieke woordvolgorde wanneer de kans bestaat dat de toehoorder de zin verkeerd begrijpt. Ik stel daarom vast dat woordvolgordebevriezing daadwerkelijk voorkomt, zij het als tendens, in het gesproken Nederlands, en niet slechts een effect is in geconstrueerde zinnen zonder context zoals (6).

Tot slot

De vraag die in dit proefschrift centraal stond was: 'Waardoor wordt de keuze voor een Vorfeldbezetter bepaald?' Ik heb in het proefschrift laten zien dat uiteenlopende factoren een rol spelen in deze keuze. Welk zinsdeel vooropgeplaatst wordt hangt af van globale woordvolgordetrends (grammaticale functie, bepaaldheid), van lokale factoren die slechts betrekking hebben op een bepaalde positie (belangrijk materiaal in het Vorfeld, complex materiaal aan de rechterkant), en bovendien van de kans dat de resulterende zin juist begrepen wordt (woordvolgordebevriezing). De combinatie van theoretische modellering, grootschalig corpusonderzoek en statistische modellering die in dit proefschrift wordt gebruikt is essentieel om deze factoren te kunnen identificeren en bestuderen in het complexe totaalbeeld van de Vorfeldbezetting.

Groningen Dissertations in Linguistics (GRODIL)

1 Henriëtte de Swart (1991). *Adverbs of Quantification: A Generalized Quantifier Approach.*
2 Eric Hoekstra (1991). *Licensing Conditions on Phrase Structure.*
3 Dicky Gilbers (1992). *Phonological Networks. A Theory of Segment Representation.*
4 Helen de Hoop (1992). *Case Configuration and Noun Phrase Interpretation.*
5 Gosse Bouma (1993). *Nonmonotonicity and Categorial Unification Grammar.*
6 Peter I. Blok (1993). *The Interpretation of Focus.*
7 Roelien Bastiaanse (1993). *Studies in Aphasia.*
8 Bert Bos (1993). *Rapid User Interface Development with the Script Language Gist.*
9 Wim Kosmeijer (1993). *Barriers and Licensing.*
10 Jan-Wouter Zwart (1993). *Dutch Syntax: A Minimalist Approach.*
11 Mark Kas (1993). *Essays on Boolean Functions and Negative Polarity.*
12 Ton van der Wouden (1994). *Negative Contexts.*
13 Joop Houtman (1994). *Coordination and Constituency: A Study in Categorial Grammar.*
14 Petra Hendriks (1995). *Comparatives and Categorial Grammar.*
15 Maarten de Wind (1995). *Inversion in French.*
16 Jelly Julia de Jong (1996). *The Case of Bound Pronouns in Peripheral Romance.*
17 Sjoukje van der Wal (1996). *Negative Polarity Items and Negation: Tandem Acquisition.*
18 Anastasia Giannakidou (1997). *The Landscape of Polarity Items.*
19 Karen Lattewitz (1997). *Adjacency in Dutch and German.*
20 Edith Kaan (1997). *Processing Subject-Object Ambiguities in Dutch.*
21 Henny Klein (1997). *Adverbs of Degree in Dutch.*
22 Leonie Bosveld-de Smet (1998). *On Mass and Plural Quantification: The case of French 'des'/'du'-NPs.*
23 Rita Landeweerd (1998). *Discourse semantics of perspective and temporal structure.*
24 Mettina Veenstra (1998). *Formalizing the Minimalist Program.*
25 Roel Jonkers (1998). *Comprehension and Production of Verbs in aphasic Speakers.*
26 Erik F. Tjong Kim Sang (1998). *Machine Learning of Phonotactics.*
27 Paulien Rijkhoek (1998). *On Degree Phrases and Result Clauses.*
28 Jan de Jong (1999). *Specific Language Impairment in Dutch: Inflectional Morphology and Argument Structure.*
29 H. Wee (1999). *Definite Focus.*
30 Eun-Hee Lee (2000). *Dynamic and Stative Information in Temporal Reasoning: Korean tense and aspect in discourse.*
31 Ivilin P. Stoianov (2001). *Connectionist Lexical Processing.*
32 Klarien van der Linde (2001). *Sonority substitutions.*
33 Monique Lamers (2001). *Sentence processing: using syntactic, semantic, and thematic information.*
34 Shalom Zuckerman (2001). *The Acquisition of "Optional" Movement.*
35 Rob Koeling (2001). *Dialogue-Based Disambiguation: Using Dialogue Status to Improve Speech Understanding.*
36 Esther Ruigendijk (2002). *Case assignment in Agrammatism: a cross-linguistic study.*
37 Tony Mullen (2002). *An Investigation into Compositional Features and Feature Merging for Maximum Entropy-Based Parse Selection.*
38 Nanette Bienfait (2002). *Grammatica-onderwijs aan allochtone jongeren.*
39 Dirk-Bart den Ouden (2002). *Phonology in Aphasia: Syllables and segments in level-specific deficits.*

40 Rienk Withaar (2002). *The Role of the Phonological Loop in Sentence Comprehension.*
41 Kim Sauter (2002). *Transfer and Access to Universal Grammar in Adult Second Language Acquisition.*
42 Laura Sabourin (2003). *Grammatical Gender and Second Language Processing: An ERP Study.*
43 Hein van Schie (2003). *Visual Semantics.*
44 Lilia Schürcks-Grozeva (2003). *Binding and Bulgarian.*
45 Stasinos Konstantopoulos (2003). *Using ILP to Learn Local Linguistic Structures.*
46 Wilbert Heeringa (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance.*
47 Wouter Jansen (2004). *Laryngeal Contrast and Phonetic Voicing: A Laboratory Phonology.*
48 Judith Rispens (2004). *Syntactic and phonological processing in developmental dyslexia.*
49 Danielle Bougaïré (2004). *L'approche communicative des campagnes de sensibilisation en santé publique au Burkina Faso: Les cas de la planification familiale, du sida et de l'excision.*
50 Tanja Gaustad (2004). *Linguistic Knowledge and Word Sense Disambiguation.*
51 Susanne Schoof (2004). *An HPSG Account of Nonfinite Verbal Complements in Latin.*
52 M. Begoña Villada Moirón (2005). *Data-driven identification of fixed expressions and their modifiability.*
53 Robbert Prins (2005). *Finite-State Pre-Processing for Natural Language Analysis.*
54 Leonoor van der Beek (2005) Topics in Corpus-Based Dutch Syntax
55 Keiko Yoshioka (2005). *Linguistic and gestural introduction and tracking of referents in L1 and L2 discourse.*
56 Sible Andringa (2005). *Form-focused instruction and the development of second language proficiency.*
57 Joanneke Prenger (2005). *Taal telt! Een onderzoek naar de rol van taalvaardigheid en tekstbegrip in het realistisch wiskundeonderwijs.*
58 Neslihan Kansu-Yetkiner (2006). *Blood, Shame and Fear: Self-Presentation Strategies of Turkish Women's Talk about their Health and Sexuality.*
59 Mónika Z. Zempléni (2006). *Functional imaging of the hemispheric contribution to language processing.*
60 Maartje Schreuder (2006). *Prosodic Processes in Language and Music.*
61 Hidetoshi Shiraishi (2006). *Topics in Nivkh Phonology.*
62 Tamás Biró (2006). *Finding the Right Words: Implementing Optimality Theory with Simulated Annealing.*
63 Dieuwke de Goede (2006). *Verbs in Spoken Sentence Processing: Unraveling the Activation Pattern of the Matrix Verb.*
64 Eleonora Rossi (2007). *Clitic production in Italian agrammatism.*
65 Holger Hopp (2007). *Ultimate Attainment at the Interfaces in Second Language Acquisition: Grammar and Processing.*
66 Gerlof Bouma (2008). *Starting a Sentence in Dutch: A corpus study of subject- and object-fronting.*

GRODIL
Secretary of the Center for Language and Cognition
P.O. Box 716
9700 AS Groningen
The Netherlands