

# **Sentence patterns in English and Dutch**

**A contrastive corpus analysis**

Published by  
LOT  
Janskerkhof 13  
3512 BL Utrecht  
The Netherlands

phone: +31 30 253 6006  
fax: + 31 30 253 6406  
e-mail: [lot@let.uu.nl](mailto:lot@let.uu.nl)  
<http://www.lotschool.nl>

ISBN: 978-94-6093-023-2  
NUR 616

Copyright © 2010: Lotte Tavecchio. All rights reserved.

VRIJE UNIVERSITEIT

Sentence patterns in English and Dutch  
A contrastive corpus analysis

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan  
de Vrije Universiteit Amsterdam,  
op gezag van de rector magnificus  
prof.dr. L.M. Bouter,  
in het openbaar te verdedigen  
ten overstaan van de promotiecommissie  
van de faculteit der Letteren  
op dinsdag 13 april 2010 om 15.45 uur  
in de aula van de universiteit,  
De Boelelaan 1105

door

Lotte Marije Tavecchio

geboren te Amsterdam

promotoren:

prof.dr. M. Hannay

prof.dr. W.P.M.S. Spooren

**TAALCENTRUM - VU**

**Laat woorden werken.**

VERTALINGEN • TRAININGEN • TEKSTREDACTIE

Dit onderzoek is financieel mogelijk gemaakt door  
het Onderzoeksfonds van het Taalcentrum-VU.

This research project is funded by the Taalcentrum-VU Research Fund.



# Acknowledgements

Although I knew in theory that one day it would be finished, I have to admit that I wondered many times whether that day would ever arrive in practice. But it has arrived. And now I look back on a process in which I have learned so much, worked with very inspirational people and have been surrounded by wonderful friends and family. I would like to take this opportunity to thank them.

First and foremost, I would like to express my gratitude to my daily supervisors, Mike Hannay and Wilbert Spooren. I think we worked really well as a team. Mike, I would like to thank you for arousing my interest in linguistics, which did not just start during my PhD project, but already in the years preceding that, during my studies. Thank you for your critical, but ever valuable feedback, your prompt replies to my many little questions through email at odd hours and your contagious enthusiasm for language, rhetoric and punctuation – you have infected me. Wilbert, I would like to thank you for playing your part in our team of three so well. You have the unique ability to always find a practical solution to any problem. Whenever I got tangled up in my web of data or statistical analyses, you always managed to quickly disentangle it. Thank you both for your constant availability, your confidence in me, and our many inspiring and entertaining meetings. It has always been a very reassuring thought that my final product would not be released into the academic world without being checked by two linguists I esteem so highly.

I would also like to express my utmost gratitude to the Taalcentrum-VU, which has fully funded my PhD project. The contribution of the Taalcentrum-VU has, however, extended far beyond their financial support. A special word of thanks is due to the Managing Director, Rob Doeve, who has always cheered me on from the sideline. As the Taalcentrum-VU PhD student (Taalcentrum-VU AiO), I benefited from many special privileges, of which the invitations to the yearly Christmas lunches and the cards and emails with constant words of encouragement are just a few examples. Without the Taalcentrum-VU this project could not have taken place.

Despite the fact that PhD projects have the reputation of being highly individual journeys, I do not think they ever are. My database would not have consisted of such a large number of texts – analysed with the greatest precision and patience – without the help of my two student assistants, Noortje Bakker and Olga Steenhoek-Kolbasina. Thank you very much for all your hard work. It was a pleasure to work with you both.

I would also like to thank Onno Huber, who has shown a great deal of patience and expertise in converting my database into a searchable corpus, the analyses of which have resulted in this dissertation. A similar contribution has been made by Gerben Mulder, who patiently and enthusiastically assisted me with the statistical analyses.

Furthermore, I would like to take the opportunity to thank my colleagues at the Vrije Universiteit Amsterdam, especially those at the English Department and the Language and Communication Department (Taal & Communicatie), the entire staff at the Taalcentrum-VU and my new colleagues at Amsterdam University College.

After having shared an office with certain colleagues for many years, they can – and in my case most definitely have – become very close and dear friends. Thank you Aleth, Merel and Femke for being such wonderful roommates and such wonderful friends. Aleth, I love the fact that we share our passion for teaching and appreciate your willingness to share your ideas and expertise with me at all times. Merel, your drive, punctuality, never-ending diligence and modesty have really inspired me, and still do. Femke, I do not know where to start. Thank you for keeping a clear head whenever I got stuck, for your patience, for our wonderful talks, for our days at the library, for our weeks in Rotterdam and Paris, and – last but not least – for formatting the layout of my dissertation (Ladies and Gentlemen, all credit goes to Femke Straatsma!). I just want to thank you for being such a good friend in every respect.

Working on a PhD dissertation for many years in a row requires patience not only from me, but also from my friends and family. I have realised that I am truly blessed with so many good and dear friends – too many to be able to address individually in this short text. A few of these do, however, deserve to be mentioned by name. Dear Rifka, Suzanne, Rosa, Anita, Janneke, Tessa, Nina and Maarten: thank you. Dear Fabien, thanks for your friendship and indirect contribution to Chapter 3, which was written in your flat in Paris. Thank you Ferdinand and Manon for your friendship and for making me your housesitter: at least half of the pages in this book were written in your house. Opa and Oma Senf, thank you for your care and concern, and for the wonderful home-cooked meals. Carla and Merijn, thanks for showing your friendship and support not just in good times, but also in rocky times. Gary, Nanette, Adriana and Daniella, thanks for our wonderful friendship that is not hindered by distance, for the constant words of encouragement and for



your patience with me. Rémy, the 'gast' (dude) I met at the library, who has become a dear friend. I would really like you to know that our working together has not only inspired me, but has helped me finish my dissertation in a fun way. Dear Floor, thank you for being such a true friend I can really rely and build on. You're a rock. Dear Eline, you have become a very dear friend in a short period of time: thanks for everything.

Dear Hilde, we are a dream team. Thank you for your friendship, for your endless patience and your ability to always listen to me and provide me with valuable feedback. Thanks for your support in difficult times and thanks for all the fun stuff we have done together. I have very fond memories of the Sundays at my house, the nights we worked in the public library (OBA) and our week on Schiermonnikoog. A friend in need is a friend indeed.

Dear Jochanan, you have not only been a very important part of my life for many wonderful years, but have also played an important role in my dissertation process. I would like to thank you for really believing in me: you have always given me the feeling that you thought I was capable of just about anything. You always had complete faith in my abilities, often when I myself did not have this faith. Thanks for your support, your faith and your love for so many years.

I would like to address a special word of thanks to my brother and sister, Wouter and Gusta. Dear Wout, the remark I keep getting from my friends basically sums it all up: I wish I had a brother like that! Well, I do, and I am very well aware of that. You are the best brother a sister could wish for. You have not only had complete faith in me and my abilities, but have been supportive in many other ways as well. As one of very many examples, thanks for providing a break away from the dissertation every year on our trips to Sydney, Rome, New York, Barcelona... and the list keeps going on. Dear Gus you are not just my sister, but you are my very best friend. You are the person I can *always* turn to, who always understands me, who I can cry with and laugh with. You have been a real support these past few years. Thank you both.

A final word of thanks goes to my parents, Louis and Marjan. I would like to thank you for your endless support and faith in me. For the patience you have had these past years and for always, always being there for me, day and night. You truly are wonderful parents and I could never, ever have finished this dissertation without you.

Lotte Tavecchio,  
Amsterdam, January 2010



# Contents

<b>Introduction</b>	<b>15</b>
<hr/>	
<b>1.Sentences in English and Dutch</b>	<b>19</b>
<hr/>	
<b>1.1 Introduction</b>	<b>19</b>
<b>1.2 The sentence: a definition</b>	<b>19</b>
<b>1.3 Sentencing</b>	<b>22</b>
1.3.1 Decision: what and how much information to put into a sentence	23
1.3.2 Decision: how to order information within a sentence and determine the informational status of a unit	24
1.3.3 Decision: what syntactic form should different information units of a sentence take	25
1.3.4 Decision: does the sentence suit the style of the genre to which it belongs?	30
1.3.5 Summary	32
<b>1.4 Sentencing in English and Dutch: a contrastive analysis</b>	<b>33</b>
1.4.1 Aspects of the English vs. the Dutch language system	33
1.4.2 Aspects of the English vs. the Dutch writing culture	35
<b>1.5 Conclusion</b>	<b>37</b>
<b>2.Discourse segmentation</b>	<b>39</b>
<hr/>	
<b>2.1 Introduction</b>	<b>39</b>
<b>2.2 The continuum of spoken and written language</b>	<b>41</b>
<b>2.3 Overview of approaches to discourse segmentation</b>	<b>43</b>
2.3.1 Discourse segmentation: from theory to data	44
2.3.2 Discourse segmentation: from data to theory	61
2.3.3 Discourse segmentation: the need for a new definition	79
<b>2.4 Sentence Information Units (SIUs): the role of punctuation</b>	<b>80</b>
2.4.1 Grammars of English and Dutch on punctuation	81
2.4.2 Variation in the use and application of punctuation marks	85
2.4.3 Use of punctuation in determining SIU boundaries: borderline cases, exceptions and problems	87
<b>2.5 Sentence Information Units (SIUs): determining hierarchical status</b>	<b>95</b>
2.5.1 Determining the nuclear status of a Sentence Information Unit	96
2.5.2 Problems in determining nuclearity	98
2.5.3 Determining satellite status and type of satellite: prepended, appended and interpolated	103
<b>2.6 Conclusion</b>	<b>110</b>

<b>3. Grammatical categorisation</b>	<b>113</b>
<b>3.1 Introduction</b>	<b>113</b>
<b>3.2 Quirk et al. as a basis for grammatical classification</b>	<b>115</b>
<b>3.3 General categorisation issues</b>	<b>116</b>
3.3.1 The cline of clausiness	116
3.3.2 Form vs. function	127
<b>3.4 Categorisation issues in combining relations</b>	<b>133</b>
3.4.1 Coordination vs. subordination	134
3.4.2 Apposition	145
<b>3.5 Genre-specific categorisation issues</b>	<b>154</b>
3.5.1 Fragments	154
3.5.2 Discourse markers, vocatives, tags	158
<b>3.6 Conclusion</b>	<b>160</b>
<b>4. Methodology</b>	<b>162</b>
<b>4.1 Introduction</b>	<b>162</b>
<b>4.2 A corpus-based study</b>	<b>163</b>
<b>4.3 Corpus design and composition</b>	<b>164</b>
4.3.1 General concerns in corpus design	164
4.3.2 Genre-specific design and compilation decisions	168
<b>4.4 Corpus annotation</b>	<b>175</b>
4.4.1 Annotation practice	176
4.4.2 Accuracy and reliability of annotation	183
<b>4.5 Statistical analysis</b>	<b>185</b>
<b>4.6 Conclusion</b>	<b>185</b>
<b>5. Sentencing patterns in English and Dutch</b>	<b>188</b>
<b>5.1 Introduction</b>	<b>188</b>
<b>5.2 General information on the corpus and sentences</b>	<b>188</b>
<b>5.3 Main sentence patterns</b>	<b>191</b>
5.3.1 Frequencies of main sentence patterns	192
5.3.2 Conclusion	194
<b>5.4 The C-pattern</b>	<b>195</b>
5.4.1 Grammatical realisation of subpattern C – uncoordinated, single unit	196
5.4.2 Grammatical realisation of subpattern C – coordinated nuclei	199
5.4.3 Summary	201
<b>5.5 The XC pattern</b>	<b>201</b>
5.5.1 The AC subpattern	203
5.5.2 Summary	213
<b>5.6 The CX pattern</b>	<b>215</b>
5.6.1 The CD subpattern	216
5.6.2 The CDE subpattern	229
5.6.3 The CDEF+ subpattern	243
5.6.4 Summary	244

<b>5.7</b>	<b>The XCX pattern</b>	<b>246</b>
5.7.1	The ACD subpattern	249
5.7.2	Summary	259
<b>5.8</b>	<b>Conclusion</b>	<b>260</b>
<b>6. Complex beginnings</b>		<b>274</b>
<hr/>		
<b>6.1</b>	<b>Introduction</b>	<b>274</b>
<b>6.2</b>	<b>Overview and exemplification of complex beginnings</b>	<b>275</b>
<b>6.3</b>	<b>ABC(X) and AIC(X) patterns</b>	<b>279</b>
6.3.1	Grammatical realisation of A and 1 in A1C(X)	279
6.3.2	Grammatical realisation of A and B in ABC(X)	285
<b>6.4</b>	<b>The A1AC(X) pattern</b>	<b>289</b>
6.4.1	Grammatical realisation of A and 1 in the A1AC(X) pattern	290
<b>6.5</b>	<b>Three elements in sentence-initial position</b>	<b>293</b>
<b>6.6</b>	<b>Conclusion</b>	<b>298</b>
<b>7. Sentence patterns and punctuation</b>		<b>302</b>
<hr/>		
<b>7.1</b>	<b>Introduction</b>	<b>302</b>
<b>7.2</b>	<b>Colons, semi-colons and dashes</b>	<b>303</b>
<b>7.3</b>	<b>The use of colons</b>	<b>305</b>
7.3.1	Type and status of units linked by colon	305
7.3.2	Grammatical realisation of discourse units surrounding colons	306
<b>7.4</b>	<b>The use of semi-colons</b>	<b>314</b>
7.4.1	Position of semi-colon between discourse units	314
7.4.2	Grammatical realisation of discourse units surrounding semi-colons	320
<b>7.5</b>	<b>The use of dashes</b>	<b>325</b>
7.5.1	Position of dash between discourse units	325
7.5.2	Grammatical realisation of discourse units surrounding dashes	327
<b>7.6</b>	<b>Comma splice</b>	<b>334</b>
<b>7.7</b>	<b>Conclusion</b>	<b>337</b>
<b>8. Interruptions</b>		<b>340</b>
<hr/>		
<b>8.1</b>	<b>Introduction</b>	<b>340</b>
<b>8.2</b>	<b>Overall frequencies of sentences with one and two interruptions</b>	<b>341</b>
<b>8.3</b>	<b>Interruptions in the C-pattern</b>	<b>345</b>
<b>8.4</b>	<b>Interruptions in the XC pattern</b>	<b>360</b>
<b>8.5</b>	<b>Interruptions in the CX pattern</b>	<b>368</b>
<b>8.6</b>	<b>Interruptions in XCX pattern</b>	<b>376</b>
<b>8.7</b>	<b>Interruptions and punctuation</b>	<b>380</b>
<b>8.8</b>	<b>A characterisation of the 6 most frequent subpatterns</b>	<b>382</b>
<b>8.9</b>	<b>Conclusion</b>	<b>391</b>

<b>9.Discussion</b>	<b>394</b>
<b>9.1 Introduction</b>	<b>394</b>
<b>9.2 Main findings - languages</b>	<b>395</b>
9.2.1 Sentence length	395
9.2.2 Main sentence patterns	397
9.2.3 Beginnings of sentences	399
9.2.4 Interruptions	403
9.2.5 Ends of sentences	407
9.2.6 Summary	410
<b>9.3 Main findings – languages * genres</b>	<b>412</b>
9.3.1 Academic prose	412
9.3.2 Newspaper articles	423
9.3.3 Short stories	437
9.3.4 Public information leaflets	448
<b>9.4 Conclusion</b>	<b>459</b>
<b>10.Conclusion</b>	<b>468</b>
<b>10.1 Introduction</b>	<b>468</b>
<b>10.2 Main findings</b>	<b>468</b>
<b>10.3 Limitations</b>	<b>473</b>
10.3.1 Methodological limitations	473
10.3.2 Limitations in relation to the interpretation of results	475
<b>10.4 Future research</b>	<b>476</b>
<b>10.5 Practical implications</b>	<b>477</b>
<b>References</b>	<b>480</b>
<b>Samenvatting in het Nederlands (Summary in Dutch)</b>	<b>489</b>
<b>Curriculum Vitae</b>	<b>499</b>
<b>Appendices: see enclosed cd-rom</b>	

# Introduction

When reading a Dutch or an English newspaper, one may come across sentences such as the following:

- (1) Het gevolg: vrijwel alle commerciële radio- en tv-stations zijn in meerderheid in handen van buitenlandse uitgevers gekomen. <s1990, newspaper articles><sup>1</sup>

(The result: almost all commercial TV and radio stations are for the majority in the hands of foreign publishers.)

- (2) But Sir Philip, making his first important adjudication since succeeding Elizabeth Filkin, has expressed dismay in his report that both sides in the battle - Mr Duncan Smith and his ex-chief of staff, Vanessa Gearson - resorted to expensive lawyers to advance their case, something past commissioners have managed to avoid. <s1607, newspaper articles>

The first example is taken from the Dutch national newspaper *de Telegraaf* and the second example from the English national newspaper *the Guardian*. Dutch readers of the first sentence may be familiar with this use of the colon and simply associate it with a device that is commonly used by writers to create an effective and powerful sentence. The same applies to English readers of the second sentence: they may be familiar with a fairly lengthy sentence that is interrupted a few times. English readers of the translation of (1) may, however, consider such use of the colon stylistically inappropriate, as the part that precedes the colon constitutes a sentence fragment and not a full clause. Dutch readers of (2), on the other hand, may perceive this as a complex sentence and perhaps even consider its structure too complex for a newspaper text. The question whether certain aspects of sentence structure are indeed more characteristic of one language than another is the focus of the present study. Its main aim is to investigate how sentences are typically structured in English and Dutch in four different genres and to what extent the languages differ from each other in this respect.

---

<sup>1</sup> All example sentences given in this book are taken from the corpus that was especially designed for the present study. All sentences have received a unique identification number, which is provided between square brackets after each example, together with the information about which of the four genres the sentence is taken from.

A similar question was addressed in an earlier small-scale contrastive corpus study by Hannay (1997), in which he compared English and Dutch sentence structure in two text types, newspaper editorials and fundraising letters. His main motivation for comparing these languages at the sentence level was that he considers the sentence 'relevant as a domain of textual analysis, since as an orthographical and rhetorical unit it in fact shares many qualities with higher units of texts' (p. 232). In constructing a sentence, a writer has to make a number of decisions about, for instance, what and how much information to put in a sentence; he has to determine what information contains the main message and what provides background information, and he has to decide what grammatical form these different pieces of information should take (cf. Chapter 1 for more elaborate account of these decisions). With respect to examples (1) and (2) above, this would mean that the respective writers of these sentences have had to decide how they wanted to present the information, thereby making decisions about sentence length, syntactic form, punctuation marks, flow of the sentence, and so on. Hannay (1997) was interested in whether this process of sentence construction is dependent on the language in which one writes. Specifically, similar to how language and writing are seen as cultural phenomena in the field of contrastive rhetoric (Kaplan 1966), where each language is seen as having rhetorical conventions that are unique to that particular language (Connor 1996: 5), Hannay expected that the process of sentence construction is also influenced by similar rhetorical conventions present at the level of sentence structure in English and Dutch. On the basis of his study, he concluded that rhetorical conventions are indeed present at the sentence level and characterised Dutch as having a 'chopping, prominence-promoting style' in which certain devices are employed to make individual messages more prominent, and English as having a more combining style, in which various pieces of information are combined using a wide range of subordinating devices (1997: 235).

The present study assigns similar relevance to the status of the sentence in a text and adopts a view of the sentence presented more recently by Siepmann, Gallanher, Hannay and Mackenzie (2008), in which they describe it as 'a segment of text comprising one or more units of information which together constitute a specific and relatively independent step in fulfilling your communicative aim at the level of the paragraph' (p. 86, see Chapter 1 for a more detailed account of the present definition of what constitutes a sentence). To analyse sentence structure in English and Dutch, the present study will use Hannay's (1997) small-scale corpus study as a reference point. It will, however, further elaborate his analytical model



and incorporate not only more texts, and thus more sentences, but also more text types.

In a time in which English plays a more and more prominent role in Dutch society, with the effect that a wide range of Dutch people with various professional backgrounds are more often exposed to English, a systematic comparison between these two languages could prove useful both for these people and for the English language professionals who are faced with an increased demand for extensive and detailed knowledge of the differences between these languages. Despite the fact that there are some contrastive studies of English and Dutch (e.g. Aarts & Wekker 1987; Mackenzie 1997; de Moor 1998; Hannay & Mackenzie 2009), the general tendency in these studies is to compare the languages at a more qualitative level, in which the claims are often based on introspection or intuition. In her contrastive corpus analysis of clause linking in English, Dutch and French, Cosme comments that ‘the few contrastive studies that are corpus-based use corpora only as sources of examples, but fail to provide quantitative information’, such as frequency counts (2007: 3). She also notes that the vast majority of contrastive studies for the languages that she compares deal with ‘lexical, semantic or micro-level grammatical phenomena’ and that few ‘have dared “venture” into the domain of contrastive stylistics’ (*ibid*). For this reason she presents a more systematic contrastive corpus analysis of information packaging and clause linking in English, Dutch and French editorials in quality newspapers (2007).

The aim of the present study is to compare English and Dutch sentence structure in a systematic way on the basis of a self-compiled and manually annotated corpus that consists of nearly 17,000 sentences in total, divided across two languages and four genres, namely academic prose, newspaper articles, short stories and public information leaflets. The main research aim is to gain insight into what the main sentence patterns are in English and Dutch; to establish to what extent these can be related to the particular linguistic systems of the language systems involved or to their writing cultures, or to the interaction between linguistic system and writing culture, and, lastly, to establish to what extent genre influences sentence structure.

### **Outline of the study**

Chapter 1 first provides a definition of what is considered to constitute a sentence in this study. It then continues to explain what is involved in constructing sentences

– the act of sentencing – and makes suggestions about how this process might be influenced by whether one constructs a sentence in English or in Dutch.

Chapter 2 describes how sentences have been analysed at one of the two levels of analysis: the level of discourse structure. It presents an overview of different approaches to defining what constitutes a unit in discourse, provides the operational definition of a discourse unit and explains how this definition has been applied to the analysis of sentences. In outlining the annotation process, the chapter presents the main problems encountered and explains how these have been dealt with in order to achieve consistent annotation.

Chapter 3 describes how sentences have been analysed at the second level of analysis: the grammatical categorisation of discourse units. It provides an overview of common issues in grammatical categorisation and focuses on how the main difficulties encountered in this study with respect to the categorisation process have been dealt with in order to achieve consistent classification.

Chapter 4 provides more information about the method adopted in this study: corpus research. It describes the corpus design and compilation process, the annotation procedure and the statistical analyses.

Chapters 5 through 8 present the results of this study. Chapter 5 constitutes the main results chapter and provides an overview of the most frequent sentence patterns in English and Dutch in the four different genres. Chapter 6 takes a closer look at the beginnings of sentences in English in Dutch, Chapter 7 at how a number of punctuation marks signal relations between various discourse units in the two languages, and Chapter 8 takes a closer look at the sentence patterns formed by interpolated satellites.

Chapter 9 presents a discussion of the main findings of this study, making an attempt to relate these to either the linguistic systems of the two languages involved or to their writing cultures.

Chapter 10, finally, concludes the study by not only summarizing its main findings, but also by presenting some limitations of the research carried out, making a number of suggestions for future research and outlining the practical implications of the present study.

# 1. Sentences in English and Dutch

## 1.1 Introduction

To be able to perform a contrastive analysis of sentence patterns in English and Dutch, it is important to first establish what exactly is here understood by the notion of sentence. This chapter will provide a definition of this notion and describe in more detail what types of decisions are involved in constructing sentences. It will then continue by explaining how this sentence construction process – or sentencings – may be influenced by the language in which the sentence is written.

## 1.2 The sentence: a definition

When asked to give a definition of what constitutes a sentence, the straightforward answer would be, at least in written language, that it is something that starts with a capital letter and ends with a full stop, a question mark, an exclamation mark or any other sentence-final punctuation mark. This would mean that the following example consists of three sentences:

- (1) De stress draagt dan niet bij tot betere prestaties. Integendeel. Uiteindelijk leidt deze situatie tot overspannenheid of burn-out. <s9978, s9979, s9980, leaflets>

(The stress does not lead to better results. On the contrary. Eventually this situation leads to stress or burn-out.)

Some, however, may have doubts about whether the second sentence does indeed constitute a sentence or whether it should, for instance, be seen as belonging to the sentence that follows it. This doubt may be raised by the fact that it takes the form of a phrase and the notion of sentence may be associated with a unit that consists of at least a few obligatory elements, such as a subject and a verb. An overview of just how many different approaches there are to defining what constitutes a sentence has been presented by de Beaugrande (1999), who has

noted that a large number of influential linguists ‘have shared a symptomatic eagerness to put the “sentence first” by assuming that sentences are and must be there, “given in advance” or “instinctively” prior to any “conscious analysis”, whether or not we can reliably define or observe them’ (p. 6). And indeed, when looking at a number of contemporary grammars of English and Dutch, it quickly becomes clear that they all acknowledge the complexity involved in defining sentences. For instance, in the *Comprehensive Grammar of English*, Quirk, Greenbaum, Leech and Swartvik (1985) note that ‘[t]he sentence is an indeterminate unit in the sense that it is often difficult to decide, particularly in spoken language, where one sentence ends and another begins’ (p. 47). In line with de Beaugrande’s observations, similar remarks can be found in a number of other grammars and reference books (eg. Haeseryn et al. 1997: 1086; Huddleston & Pullum 2002: 44-45; Downing & Locke 2006; Biber et al. 1999: 202).

Part of the complexity in defining sentences arises when adopting a structural approach, marked by de Beaugrande as ‘the most widely established’ (1999: 27), in which the question is what constitutes a grammatical sentence or what elements need to be present for a sentence to be considered grammatical. Although determining grammaticality represents another area in linguistics about which no consensus has been reached (cf. Quirk et al. 1985: 47), a criterion that can be used in determining what counts as a grammatical sentence is, for instance, the notion of syntactic completeness (cf. Matthews 1981: 29). In this view a sentence can be considered grammatical if it contains all the necessary elements that make it syntactically whole. However, determining what these elements are and, more importantly, whether the same criteria should be set for sentences produced in the written language as in the spoken language poses yet a new set of questions to which there are no straightforward answers (Quirk et al. 1985: 47; Biber et al. 1999: 202, 1069ff; Carter & McCarthy 2006: 486).

One of the ways in which some grammars appear to have dealt with the ‘indeterminate’ nature of the sentence is by introducing the notion of clause alongside it, which is ‘in many ways a more clearly-defined unit than the sentence’ (Quirk et al. 1985: 47, but see also Biber et al. 1999: 50; Huddleston & Pullum 2002: 44; Halliday & Matthiessen 2004: 371; Downing & Locke 2006: 274). The sentence is still considered the highest linguistic unit, but the focus has shifted from defining sentences to defining clauses. Although there is some variation between definitions, in general terms, the relation between the sentence and clause is that the former is the highest-ranking linguistic unit that forms the top of a hierarchy that is made up of sentences, clauses, phrases, words and morphemes. The sentence is then

defined in relation to the clause, as a unit that consists of one or more clauses (Quirk et al. 1985: 47; Huddleston & Pullum 2002: 44-45; Carter & McCarthy 2006: 486; Downing & Locke 2006: 272). Others also choose to just replace the term sentence by independent clause as a means of avoiding the definition problem of sentences (Biber et al. 1999: 202). Clauses, on the other hand, are typically classified as one of two types, either independent clauses or dependent clauses, including incomplete clauses. Clause status is determined on the basis of the extent to which a clause contains its necessary clause elements, which range from central elements to peripheral elements. The distinction between central and peripheral elements is relative rather than absolute and these elements may therefore be classified differently in different grammars (cf. Quirk et al. 1985: 50; Carter & McCarthy 2006: 487; Downing & Locke 2006: 275, but see also Chapter 3, Section 3.3.1 on the 'cline of clausiness'). With this shift in focus from sentence to clause, the clause has become a unit that is described in terms of its grammatical structure, whereas the sentence is typically referred to as an orthographic unit (cf. Halliday & Matthiessen 2004: 8).

Without attempting to resolve the issue of defining what constitutes a sentence, the present study will simply adopt a particular approach. Following Siepmann, Gallagher, Hannay and Mackenzie (2008), it approaches the notion of sentence from a discourse-rhetorical perspective. On the one hand, it is an orthographic unit, which means that it is identifiable by its capital letter at the start and the sentence-final punctuation mark at the end. It should be noted that this approach is possible in the present study, as it is restricted to the analysis of sentences in the written language. On the other hand, the sentence is seen as a rhetorical unit, which means, to use Siepmann et al.'s words, that it is 'a segment of text comprising one or more units of information which together constitute a specific and relatively independent step in fulfilling [one's] communicative aim at the level of the paragraph' (2008: 86). The orthographic definition thus defines the sentence in terms of its form, whereas the rhetorical definition defines it in terms of its status with respect to one's communicative intentions (*ibid*).

With respect to example (1) above, this is thus analysed in the context of this study as consisting of three orthographic-rhetorical sentences. As this example illustrates, different sentences can be realised grammatically in different ways. Whereas the first and the last sentence take the form of an independent clause, the sentence in between these two clauses is realised as a clause fragment or phrase. The orthographic-rhetorical sentence is seen as reflecting the choices a writer has made in the way in which he wants to package information and present

this to the reader. For instance, with respect to (1), the decision to present the conjunct *integendeel* as a separate orthographic sentence is seen as a deliberate one. In this case, precisely because there is a mismatch between the orthographic sentence and the grammatical sentence, in that the latter lacks the necessary sentence elements to make it syntactically whole, the writer achieves the effect of underlining the contrast that is already captured by the function and meaning of the conjunct *integendeel*.

The following section will take a closer look at what decisions are involved in constructing sentences.

### 1.3 Sentencing

The process of constructing orthographic-rhetorical sentences involves making decisions about how to package information in linguistic units. This process will be referred to as 'sentencing', a term coined by Hannay in his small-scale contrastive corpus analysis of Dutch and English (1997: 232, see also Siepmann et al. 2008: 92). Instead of gaining insight into the process of sentencing by, for instance, trying to establish what occurs during online sentence construction, the present study will make an attempt to reconstruct this process by using the end product, the sentence itself, as the starting point. In other words, on the basis of an analysis of the completed sentence, assumptions will be made about the types of decisions that were involved in its construction. It will be assumed that the writer will have had to make decisions about at least the following aspects (cf. Hannay 1997: 232; Siepmann et al. 2008: 92-96):

- What and how much information to put into the sentence (1.3.1)
- How to order information within a sentence and determine the informational status of a unit (1.3.2)
- What syntactic form different elements of a sentence should take (1.3.3)
- Whether a particular sentence suits the style of the genre to which it belongs (1.3.4)

The following sections will describe in more detail what is involved in making each of these decisions.

### 1.3.1 Decision: what and how much information to put into a sentence

In using the sentence as a means of conveying information to the reader, the writer first has to make a decision about what information he wants to put into the sentence. The type and amount of information determines its length and its general degree of complexity. Although this decision is, at least to a certain extent, influenced by the genre in which he writes (see 1.3.4 below), the writer can choose to package the information in sentences of various lengths, while being aware of the different purposes that, for instance, short or long sentences serve and the different effects they can achieve.

Information about these different effects can, for instance, be gained from various writing manuals, which provide information about short and long sentences and generally advise readers to vary their sentence length (see, for example, Brooks & Warren 1979: 240; Kane 1988: 171; Onrust et al. 1993: 161; Nederhoed 2000: 350; Renkema 2005: 87, 90; Permentier 2003: 148; Burger & de Jong 2009: 131; Hannay & Mackenzie 2009: 104ff). Short sentences, on the one hand, are typically associated with a simple style that requires little processing effort on the part of the reader. However, if a text predominantly contains short sentences, this can be perceived as having a so-called segregating style, where the information contained in separate sentences seems unrelated and may require the reader to spend his processing effort on linking the various information units to each other (Onrust et al. 1993: 161; Burger & de Jong 2009: 131). At the same time, the segregating style can also be used deliberately to achieve a particular rhetorical effect, by forcing the reader to pause after each sentence, and thus drawing attention to each separate unit, as illustrated by *integendeel* in example (1) above, but also by example (2) below (taken from Onrust et al. 1993: 162, see also Kane 1988: 120; Verhagen 1991: 83):

(2) I came. I saw. I conquered.

Long sentences, on the other hand, are often characterised as carrying the risk of containing too much information and a structure that is too complex. The effect of this may be that not only the reader, but also the writer might get lost in the sentence (Burger & de Jong 2009: 133). At the same time, writing manuals point to the fact that long sentences should not be avoided, as long as they are well structured and techniques are mastered that can be applied in managing long sentences (cf. Williams 1990: 135; Renkema 2005: 79; Burger & de Jong 2009: 132ff).

Because sentences of different lengths can achieve different effects and thus serve different purposes, writers are advised to adjust the length of the sentence to the effect they want to achieve, to the type of text to which the sentence belongs and to the readership of the text. This means that writers have to be aware of the effects that sentences with different lengths can have and how they can use this knowledge effectively in constructing sentences. Determining what and how much information to put into the sentence thus constitutes one of the main decisions that writers have to make in constructing sentences.

### **1.3.2 Decision: how to order information within a sentence and determine the informational status of a unit**

A writer not only has to determine what and how much information to put into a sentence, but also how to structure and classify different pieces of information. He has to determine what information constitutes the central message and what information provides less central information, and how he can arrange these different units of information in such a way to best achieve his communicative goal.

In the process of sentencing, i.e. packaging information into linguistic units, in those cases in which there is more than one information unit<sup>2</sup>, the writer has to decide how he wants to order these information units within the sentence. This involves determining the informational value of each of the units and the relation between them. Specifically, he has to determine which information unit constitutes the central or nuclear message and which provides background or satellite information. The informational status of a unit is not solely dependent on its position in the sentence, but is also determined by its grammatical realisation and semantic content. The following example illustrates how a writer can choose to present different pieces of information in a particular way to achieve a certain rhetorical effect.

- (3) 1. Alma's echtgenoot ging niet elke dag meer naar zijn werk. <s16653, short stories>

---

<sup>2</sup> The notion of information unit will be further elaborated on in Chapter 2, which will provide an overview of various approaches to defining what constitutes a unit of information in discourse. It will also provide a definition of the unit of analysis that has been used in the present study in order to analyse the information structure of sentences in English and Dutch



2. Hij was gespannen en bleef liever thuis. <s16654, short stories>  
 3. En Alma richtte, vanaf toen, al haar aandacht op haar dochtertje en het huishouden.  
 <s16655, short stories>

- (1. Alma's husband no longer went to his work every day.  
 2. He was tense and rather stayed at home.  
 3. And Alma focused, from then on, all her attention on her little girl and the housework.)

In (3.3), it could be argued that the decision to place the information about time, *vanaf toen* (*from then on*), not at the start of the sentence, but more towards the middle and presenting it as a separate punctuation unit by surrounding it by commas reflects a deliberate choice on the part of the writer. Specifically, precisely because it is surrounded by commas, it interrupts the flow of the sentence, which has the effect of placing prominence on information that may usually or typically be associated with background information. This illustrates just one of the many types of choices writers have in determining the order of information within a sentence. In creating sentences, a writer thus has to be aware of the functions and informational value of different information units and how different ways of ordering these can lead to different effects.

### 1.3.3 Decision: what syntactic form should different information units of a sentence take

In constructing sentences, the decisions a writer has to make with respect to information structuring do not occur independently of the decisions about the syntactic form that an information unit takes. For instance, if a writer wants to provide background information and he can choose between putting this information in an independent clause or a dependent clause, chances are that he will go for the dependent clause, as 'the information in a subordinate clause is often placed in the background with respect to the superordinate clause' (Quirk et al. 1985: 919). This means that a writer has to make a decision about the syntactic form of an information unit and that he has to be aware of the various effects that different syntactic forms can have, where he has to select those that best fit his communicative goal.

In building sentences, a writer can basically choose to package his information in a simple sentence, a compound sentence or a complex sentence.

Whereas simple sentences typically consist of one information unit, compound and complex sentences consist of more than one information unit, which means that the writer has to determine how the various information units are related to each other. These can be related to each other in two main ways: the information units can be paratactically related, in which case the related units are of equal status, or they can be hypotactically related, in which case the units are of unequal status (Downing & Locke 2002: 280-281)<sup>3</sup>. This means that in constructing sentences writers not only need to be aware of the different ways in which they can combine information units, but also of the variety of grammatical constructions and combinations that they can choose from and the effects that these different constructions and combinations have.

### Simple sentences

In those cases in which a sentence consists of one information unit, this unit typically takes the syntactic form of an independent clause containing one subject and verb, although it was established in Section 1.2 above that simple sentences can also be realised by something other than independent clauses. Consider example (4), which consists of four simple sentences of which the first two are realised as independent clauses and the last two as a fragment and a subordinate clause respectively.<sup>4</sup>

- (4) It seems certain that the Madrid massacre was carried out by Al-Qaeda or some similar fanatical group. That brings the threat even closer to Britain. And not just geographically. For there is little doubt that early results from Spain's general election show that many voters think their government increased the danger from terrorism by supporting the war in Iraq. <s236-239, newspaper articles>

This example serves to illustrate that orthographic sentences do not always need to be realised as independent clauses, but can also be realised as clause fragments or

---

<sup>3</sup> See Chapter 3 for more information on hypotactic and paratactic relations and the gradient relation between them (see also Cosme 2007 for an extensive overview).

<sup>4</sup> Even though *for* can be seen as being on the gradient between coordinators and subordinators, it is here analysed as a subordinator, in line with Quirk et al.'s analysis of this conjunction (1985: 921, note; 927).

subordinate clauses. Determining what syntactic form a sentence takes thus involves another decision on the part of the writer.

When looking at the type of advice that is usually given in relation to this decision, it becomes clear that writing manuals typically advise against the use of sentences that are grammatically incomplete. The reasons provided are, for instance, that sentence fragments are associated with more informal writing styles (Kane 1996: 137, 173; Anson & Schwegler 1998: 291), or that they require readers to use unnecessary processing effort in establishing the connection of these fragments to the surrounding context (Anson & Schwegler 1998: 290). However, some also point to the rhetorical effect that these incomplete sentences can have when used sparingly, precisely because they present a break from the standard pattern and serve as an effective way of varying sentence constructions or creating emphasis (Kane 1988: 139, 179; Anson & Schwegler 1998: 291; Tiggeler 2006: 194). In determining the grammatical realisation of the nuclear message, writers should thus be aware that simple sentences can be syntactically realised in different ways and that different realisations have different effects. In constructing simple sentences, a writer has to determine what syntactic realisation best suits his communicative goal.

### **Compound sentences**

Compound sentences typically consist of two or more clauses that are paratactically related. The essence of a paratactic relation is that the units that are joined are of equal grammatical status. In constructing a compound sentence, the writer has to make a decision about how explicitly he wants to link the units. For instance, a choice that presents itself in the case of coordination, which constitutes a type of parataxis, is whether to link the clauses or phrases syndetically, i.e. by means of coordinating conjunctions such as *and*, *or* and *but*, or asyndetically, i.e. without the use of coordinating conjunctions. In opting for asyndetic coordination, the writer can, for instance, choose to link two clauses of equal grammatical status by means of a semi-colon, a punctuation mark that is frequently used to link asyndetically linked units (cf. Quirk 1985: 1622; Huddleston & Pullum 2002: 1742-1743). An example of asyndetic coordination by means of a semi-colon is presented in (5) below.

- (5) Haiti has few natural resources; its economy is mainly agricultural. <s1229, newspaper articles>

With respect to the use and construction of compound sentences, the advice that can be found in various writing manuals is for writers to find the right balance in the frequency with which they are used. On the one hand, writers are advised not to overuse them, as this can lead to a stylistically weak and boring style, in which one message is just added to the next (Onrust et al. 1996: 165; Anson & Schwegler 1998: 384). On the other hand, writers are made aware of the rhetorical effect that can be achieved by the use of compound sentences, as sporadic excessive use can, for instance, reinforce tediousness or repetitiveness of particular situations (Onrust et al. 1996: 163-165). Knowledge of the various ways in which information units can be combined through parataxis thus forms another aspect of constructing effective sentences.

### **Complex sentences**

In creating complex sentences, writers combine ideas or information units that are of unequal grammatical status. These information units are, in other words, hypotactically related (cf. Downing & Locke 2002: 280-281). A complex sentence generally contains one central information unit that is modified by one or more information units that provide less central, and typically more background information. It is the writer's task not only to arrange the information units in such a way that the reader can distinguish between central and less central information, but also to make use of those linguistic structures that reflect the informational status of a unit. This means that the writer needs to be aware of the basic principles of information structure, of the wide array of linguistic structures that are available, and particularly of how these two interact.

Section 1.3.2 briefly described that in constructing sentences, a writer needs to be aware of the informational status of the various units in the sentence and make decisions about the order in which he wants to present these. With respect to the grammatical realisation of the various units in a sentence, he needs to be aware of the fact that central information typically takes the form of an independent clause and background information typically that of a dependent clause or phrase. Dependent clauses come in different types: finite, non-finite and verbless clauses. Within these three main types, a further subdivision can, for example, be made into adverbial clauses and relative clauses. The following examples serve to illustrate how the interaction between the position of the information unit in the sentence, its hierarchical status and its grammatical realisation can yield different effects.

- (6) Toen Pronk een paar dagen later tijdens een kenningsmakingsgesprek met Verdonk het woord niet wilde terugnemen, escaleerde de situatie en wees Verdonk hem woedend de deur. <s2776, newspaper articles>

(A few days later during an introductory meeting with Verdonk when Pronk did not want to take the word back, the situation escalated and a livid Verdonk showed him the door.)

- (7) Mr Blair, who had talks with UN Secretary General Kofi Annan last night, has a critical role in bringing the two sides together. <s257, newspaper articles>
- (8) Both major opposition parties are now refusing to cooperate, casting serious doubt over the credibility of the inquiry. <s578, newspaper articles>
- (9) Nu het betalingssysteem van de zorg op de schop gaat, is het van belang dat het nieuwe systeem doorzichtiger wordt, zodat de prestaties en de kwaliteit van de zorginstellingen beter kunnen worden gemeten en beter kunnen worden vergeleken. <s2190, newspaper articles>

(Now that the health care payment system is for the chop, it is important that the new system is more transparent, so that the results and quality of the health care institutions can be better measured and compared.)

Sentence (6) presents an example of a situation in which a dependent clause occurs in sentence-initial position and precedes the main clause. In this example, the information contained in the dependent clause has the function of setting the scene for the main message that is yet to come. Sentence (7) contains an example of a situation in which the main clause is interrupted by a dependent clause. This interruption can, for instance, have the function of providing additional or background information about the subject of the main clause, as it does here, and can place emphasis on the surrounding words precisely because it interrupts the flow of information (Kane 1988: 135; Siepmann et al. 2008: 168ff; Hannay & Mackenzie 2009: 100). Sentence (8) contains an example of a situation in which an independent clause is followed by a dependent clause in sentence-final position. This position is typically reserved for information that receives the main focus and the function of clauses in this position can be to elaborate on or to clarify what has been said in the clause that precedes it (Onrust et al. 1993: 173; Hannay & Mackenzie 2009: 113ff). In this particular example the dependent clause describes the consequence of what was said in the preceding independent clause. Last,

sentence (9) presents an example of a situation in which the independent clause is both preceded and followed by independent clauses, with the sentence-initial clause fulfilling a scene-setting function and the sentence-final clause providing a reason for the information contained in the preceding independent clause.

In writing manuals, attention is often drawn to the effect that complex sentences can have on readability and the writer is advised to take great care in composing them. On the one hand, writers are made aware of the advantages that can be gained from the clear structuring of information in terms of processing ease for the reader, while at the same time they are advised to guard against putting too much information in one sentence, as this can also lead to processing difficulties (Onrust et al. 1993: 166-168; Anson & Schwegler 1998: 387; Nederhoed 2000: 347; Permentier 2003: 150; Burger & de Jong 2009: 132ff). In writing manuals that contain more extensive sections on constructing complex sentences, writers are often presented with various coordinating and subordinating strategies (eg. Fowler & Aaron 2010: 389ff).

In the construction of complex sentences writers thus need sound knowledge of the principles of information structuring and the linguistic possibilities that the language system offers, and how these interact to achieve their intended communicative effect.

### **1.3.4 Decision: does the sentence suit the style of the genre to which it belongs?**

One further aspect of sentence construction that writers need to be aware of is whether a particular sentence suits the genre to which the text belongs. Without presenting an overview of the wide variety of ways in which genre can and has been defined, the present study will adopt Biber et al.'s (1999) approach to defining genres, or registers, as they refer to them. They explain that they do not distinguish one register from another on the basis of linguistic features, but on the basis of 'situational characteristics', such as 'mode, interactiveness, domain, communicative purpose, and topic' (1999: 15). In this view, registers constitute 'situationally defined varieties' (p. 4). With respect to, for instance, newspaper editorials, these can be characterised as constituting non-interactive, published written texts about a wide variety of topics with the aim of expressing informed opinions about news items, published in newspapers (p. 15).

As for the decisions that a writer has to make when constructing sentences, these will thus also be influenced by the characteristics of the particular genre in which he writes. For instance, when writing a public information leaflet, a writer has to be aware of the intended readership of the text and its purpose. As this typically constitutes a wide and varied readership, writers are advised to present the information in such a way to make it accessible to a broad readership. Suggestions put forward to increase the accessibility and readability are, for instance, to keep sentences rather short, to pose direct questions, to use personal pronouns, subheadings and lists (cf. van den Boomen & van der Lans 1991: 103, 114-115; Woerkum & Kuiper 1995: 128ff; Huigen 2004: 27ff ; Hopster & Tiggeler 2007). As Biber et al. explain, '[t]he situational characteristics that define registers have direct functional correlates, and, as a result, there are usually important differences in the use of grammatical features among registers' (1999: 15). With this in mind, consider the following sentences:

- (10) While it is clearly suggested that autobiographical memories of childhood are, on the whole, subject to some form of contamination across the life-span, Ross and Conway (1986) and Baddeley (1990), although sceptical about the efficacy of retrospective studies, have conceded that, in their view, most people's recall of past events remains accurate across time.  
<s5423, academic prose>
- (11) Theo, zo heette hij. Nu de achternaam nog. <s13636, s13637, short stories>  
(Theo, that was his name. Now for his surname.)
- (12) This leaflet is about:  
 - the health effects of exposure to loud noise;  
 - your legal duties to protect the hearing of your workers;  
 - how to assess and control noise;  
 - how to choose quieter equipment and machinery;  
 - different methods of hearing;  
 - health surveillance <s6783, leaflets>

Example (10) is taken from an academic journal article, example (11) from a short story, and example (12) from a public information leaflet. Example (10) can be characterised as a long and complex sentence, the complexity of which is not only related to its length, but also to the interrupted information structure and high proportion of hypotactic relations. Example (11), on the other hand, can be characterised by its fragmentary style, reflected in the grammar by the use of

incomplete clauses. Last, example (12) can be characterised by the particular form in which the information is presented, i.e. as a bullet point list. These differences in structure and style between these sentences are reflective of ‘the fundamental influence of register on grammatical choice’ (Biber et al. 1999: 24). As Biber et al. explain it, ‘[w]hen speakers switch between registers, they are doing very different things with language, using language for different purposes, and producing language under different circumstances’ (1999: 21).

In short, although writers might not be aware of the precise effect of genre distinctions on linguistic structure, they need some awareness of the conventions of a particular text type and how those relate to sentence structure, but also how these conventions relate to the communicative purpose of a particular text type (also see Onrust et al. 1993: 169).

### **1.3.5 Summary**

This section has provided some insight into the types of decision a writer needs to make in sentencing, i.e. the process of packaging information in linguistic units to construct orthographic-rhetorical sentences. This process involves determining what information to put into a sentence; how much information to put into a sentence; how to order the information in a sentence and determine the hierarchical status of the various information units; what linguistic form the various units should take; and whether a sentence suits the genre to which it belongs. It can be assumed, or at least suggested, that a number of these decisions are to a certain extent influenced by the language in which the sentence is written, similar to how they are dependent on the genre in which they are written. The next section will look into the ways in which a particular language, in this case English or Dutch, and a language writing culture can influence sentencing.



## **1.4 Sentencing in English and Dutch: a contrastive analysis**

It is the main aim of this study to uncover what the most frequent sentence patterns are in English and Dutch, to compare and contrast them and to establish to what extent genre influences sentence structure. Since the act of sentencing involves making linguistic and rhetorical decisions about sentence structure, it is expected that this process is, on the one hand, influenced by the linguistic system in which the sentence is produced, i.e. either the English or Dutch linguistic system, and, on the other hand, by the language culture in which the sentence is produced, i.e. either the English or Dutch writing culture. Both aspects will be further elaborated on in the sections below.

### **1.4.1 Aspects of the English vs. the Dutch language system**

Despite the fact that the list is not very extensive, there have been a number of studies in which English and Dutch have been compared (eg. Aarts & Wekker 1993; Mackenzie 1997; de Moor 1998; Cosme 2007). However, the majority of these focus on 'lexical, semantic or micro-level grammatical phenomena', as Cosme also observed (2007: 3), such as the use of tenses, the make up the noun phrase and the use of the modal auxiliaries (cf. Aarts & Wekker 1993). As such contrastive analyses mainly concern linguistic differences at the phrase or clause level and not necessarily at the level of sentencing or clause combining, they will not be considered in great detail in the present section. Instead, this section will be restricted to discussing those syntactic differences that could affect sentence construction at the level of sentencing or clause combining.

One of these differences, which mainly affects the beginning of sentences, concerns the fact that Dutch, unlike English, can be characterised as a verb-second language, which means that the finite verb is typically placed in second position and that there is a general rule that no more than one element is allowed in sentence-initial position (Haeseryn et al. 1997: 1261; Smits 2002: 22). This means that examples such as the following, in which the sentence starts with two adverbials that precede the subject, are probably either non-existent or rare in Dutch.

- (13) Similarly, in the actual crisis of May 1940, it was Amery rather than Salisbury who occupied the centre-stage, as even Witherell's narrative tends to confirm. <s4348, academic prose>

In her study on complex beginnings in native and learner English, Smits (2002) indeed found that while Dutch does not categorically reject sentences with adverbial clusters in sentence-initial position, the pattern is much more common in English (p. 168-169). Although this concerns a difference between the linguistic systems of English and Dutch, it is interesting to note that Smits also found a frequency difference in the occurrence of complex beginnings between the different genres that she incorporated in her study, with the English academic prose genre showing the highest frequency (2002: 54, 110). As she only included one Dutch text type, i.e. newspaper texts, it would be interesting to see what an analysis of complex beginnings in the four different genres included in the corpus designed for this study yields. In the analysis of sentencing patterns in English and Dutch, differences can thus be expected in the make-up of the beginning of sentences.

Another rather well-known difference between English and Dutch concerns the occurrence of non-finite clauses. Various studies have observed – often without providing the quantitative data to support the claims – that these are notably more frequent in English than in Dutch (cf. Aarts & Wekker 1987: 301; De Moor 1998: 309; Cosme 2007: 279-280; Hannay & Mackenzie 2009: 93). This means that a sentence such as the following, in which the subordinate clause in sentence-final position takes the form of a non-finite clause, presents an example of a particular clause type that is more common in English than in Dutch.

- (14) Both major opposition parties are now refusing to cooperate, casting serious doubt over the credibility of the inquiry. <s578, newspaper articles>

With respect to sentencing patterns, it can thus be expected that a difference can be found in the form that subordinate clauses take in the two respective languages, with English showing a higher frequency of non-finite clauses.

Although differences between the use and occurrence of adverbial clusters in sentence-initial position and non-finite clauses are associated with differences in the linguistic systems of English and Dutch, both still constitute differences of degree instead of absolute differences. In his small-scale contrastive corpus analysis of English and Dutch sentencing patterns, Hannay analyses the

frequency and use of various syntactic structures and reaches a similar conclusion by stating that '[w]hile Dutch and English writers make use of the same syntactic devices when formulating messages and constructing sentences, there are significant differences in the frequency with which certain devices occur' (1997: 249). With this in mind, it would be interesting to see if insight can be gained into the factors that influence this frequency. As text type or genre could be one of the explanatory variables, the particular composition of the corpus designed for the present study, which includes four different genres, may provide more insight into the precise nature of the frequency differences between English and Dutch.

#### **1.4.2 Aspects of the English vs. the Dutch writing culture**

Besides differences in the linguistic systems of English and Dutch that may lead to differences in sentencing patterns, there may also be relevant differences in the writing cultures of English and Dutch. At the risk of using a concept that can be approached and defined in many different ways, it will first be established what is here meant by the notion of writing culture. The term writing culture has been used within the context of second language writing, and more specifically, within the context of contrastive rhetoric (Kaplan 1966). In contrastive rhetoric, language and writing are seen as cultural phenomena and each language is seen as having rhetorical conventions that are unique to that particular language (Connor 1996: 5, also see Connor, Nagelhout & Rozycki 2008 for the current state of contrastive rhetoric, or 'intercultural rhetoric' (p. 4)). This means that it is assumed that the way in which ideas are packaged into linguistic units is culturally determined or, put differently, that writing is seen as an activity that is embedded in a culture. Although there are many different methods of gaining insight into a particular writing culture, the method that will be adopted in the present study is by relating findings in sentencing patterns to the guidelines provided in writing handbooks and style guides of English and Dutch. These are then considered to reflect the writing culture of a particular language community at least to a certain extent.

In an earlier corpus-based, contrastive analysis of sentencing patterns in English and Dutch argumentative texts, Hannay (1997) related his findings of differences in sentence patterns to differences in writing styles – or rhetorical conventions – of the two languages. He found that, on the whole, Dutch and English make use of the same syntactic devices when constructing sentences, but that there were significant frequency differences in the occurrence of certain

devices in the particular text types under consideration (1997: 249). The relatively high frequency of sentence fragments, colon climaxing structures (see example (1) above), interrogatives and relatively short sentences led to the characterisation of written Dutch as having a so-called chopping style. This is a style in which the use and occurrence of particular structures serves to make individual messages more prominent. It was suggested that this style may be reflective of the fact that written Dutch is closer to the spoken language than written English (1997: 234, see also Hannay & Mackenzie 2009: 219). As for written English, it was the relatively high frequency of coordinated and subordinated clauses and high frequency of long sentences that lead to its characterisation as a language that can be described as having a so-called combining style. Whether these differences in style are indeed characteristic of the two languages or just of the particular text type under consideration is a question that has also been posed by Cosme (2007) in her contrastive analysis of clause combining in English, Dutch and French. She, however, also restricted her analysis to one text type, i.e. newspaper editorials. One way to determine whether the differences in style are indeed representative of the languages under consideration would be to expand the number of text types in the analysis. It is for this reason that the present study will analyse sentencing patterns in four different genres – academic prose, newspaper articles, short stories and public information leaflets.

Hannay thus relates his findings in writing style to the rhetorical conventions present in the Dutch and English writing cultures respectively. When comparing, for example, the guidelines that a random selection of style guides in either language give on the subject of sentence length, it becomes clear that the Dutch ones more often focus on restricting the length of the sentence, in order to increase its readability (cf. Permentier 2003: 141, 150; Hermans 2006: 49; Tiggeler 2006: 190). A random selection of English style guides, on the other hand, shows that sentence length actually is not explicitly addressed (eg. Sinclair 1992; Ritter 2002; Peters 2004; Hicks 2009). In the exceptional cases in which reference is made to length, the focus tends to be more on the ways in which long sentences can be structured, using various coordinating and subordinating devices (eg. Williams 1990: 135). It could be possible that such explicit differences in advice may also lead to differences in sentence structure. More specifically, in Section 1.3.1 it was explained that determining what type of information and how much information should be included in a sentence constitutes one of the main decisions in sentence construction, a decision that affects other decisions, such as the order of different information units in a sentence and the syntactic form of these units. If Dutch

writers receive explicit instructions about sentence length, sometimes even expressed in the maximum number of words that a sentence should consist of (eg. Lamers 1986: 124-131; Donkers & Willems 2002: 184-186), and English writers typically do not receive such explicit advice, it could be hypothesised that this type of advice affects the way writers go about composing sentences. This would mean that a number of the decisions would then thus not be exclusively determined by the language in which they write, but also the language culture.

The present study will try to establish what the main sentencing patterns are in English and Dutch; it will try to establish to what extent these can be related to the particular linguistic systems of these languages or to the writing cultures of these languages, or to a combination of linguistic system and writing culture; and it will try to establish to what extent genre influences sentence structure. It will do so using a self-compiled corpus of English and Dutch texts that belong to the following four genres: academic prose, newspaper articles, short stories and public information leaflets.

## **1.5 Conclusion**

The main aim of this chapter was to provide insight into the types of decisions a writer makes when composing a sentence. In other words, the aim was to explain what is involved in the act of sentencing. However, before the process of sentencing could be described, it first had to be established what constitutes a sentence. In the present study the sentence is seen as constituting both an orthographic unit and a rhetorical unit. This means that a sentence is considered to reflect the choices the writer has made in its composition in order to achieve his communicative goal. It was then explained that these choices can be either determined by the linguistic system or the writing culture, or by a combination of both.

It is the main aim of the present study to identify the main sentencing patterns in English and Dutch, to compare and contrast them and to determine to what extent sentence structure can be related to the linguistic system, the writing culture or the interaction between these two. To analyse sentences in a systematic way, a corpus has been compiled of English and Dutch sentences that were taken from four different genres. The following chapters will describe how all sentences have been analysed at both the level of discourse (Chapter 2) and grammar (Chapter 3) in order to be able to carry out the sentencing analysis.



## 2. Discourse segmentation

### 2.1 Introduction

Chapter 1 described the main aim of the present study as gaining insight into sentencing patterns in the English and Dutch written language in order to see to what extent these two related languages show either similarities or differences in this respect and to see to what extent the four genres included in this study influence sentence structure. In a reconstruction of the sentencing process, a description was given of how a writer needs to make a number of decisions about, for instance, what information he wants to put in a sentence; what information constitutes the central message and what information provides supporting or background information; about the syntactic form that various pieces of information take; and about how a writer has to determine whether a particular style also suits the genre to which a sentence belongs.

Gaining insight into sentencing patterns thus involves gaining insight into the ways in which writers structure and package information in sentences. To be able to analyse this information packaging process, one of the key issues that has to be resolved first is determining what constitutes a message or a unit of information. A rather standard way of analysing sentences is to take a syntactic approach and to segment them into syntactic units, in which case the clause is typically taken to correspond to a unit of information or message. However, an example such as (1) below shows that discourse, be it spoken or written, is not always produced in clauses, but that pieces of information can also be smaller or bigger than the clause.

- (1)
1. How `bout that?
  2. You up for that, Rhianne?
  3. You mean - go out tonight?
  4. Yeah, why not?
  5. No time like the present, eh Rena?
  6. Right Carole. <s12164- s12169, short stories>

Example (1) is taken from an English short story and consists of six orthographic sentences. An analysis of these sentences quickly shows that none of them can be

analysed as constituting or consisting of syntactic clauses, only of clause fragments or non-clausal units. This means that a syntactic approach in which the clause is seen as constituting the basic unit of analysis would not be applicable to this particular example. Moreover, in addition to the question of what constitutes the basic unit of analysis and whether or not this corresponds to a syntactic clause, a further question concerns determining how different pieces of information within the orthographic sentence as a whole are related to each other. For instance, what piece of information should be taken to constitute the central message in (1.4): the affirmative 'yeah' to the question posed (1.3), 'go out tonight?', or should 'yeah' be seen as a response form that precedes the central message, realised as the question 'why not'? An analysis of an example such as (1) thus not only illustrates that a syntactic approach to the analysis of sentencing patterns may not work very well for sentences that consist of various pieces of information that are difficult to classify syntactically, such as 'you mean', 'yeah' and 'right Carole', but also that it may not provide sufficient information about the hierarchical status of the information units. This means that in order to analyse the sentencing patterns of (1) the unit of analysis or information unit needs to be specified and identified in terms other than syntax alone.

In fact, determining what constitutes the basic unit of analysis or the basic unit of information in discourse has proven to be far from straightforward. It is therefore the main aim of the present chapter to gain insight into the problematic nature of discourse segmentation with the practical purpose of determining what approach is most suitable for the analysis of the discourse structure of sentences. As the present study has been restricted to the analysis of written language, the chapter will start by classifying the four genres included in the corpus in relation to the traditional distinction between spoken and written language and describe how this distinction affects the process of discourse segmentation. It will then provide an overview of various approaches to discourse segmentation, in which special emphasis will be placed on identifying and evaluating the type and applicability of the definitions and criteria that are provided to consistently identify information units in the analysis of discourse. The chapter will end by presenting the approach to discourse segmentation that has been developed for this study and describe how this has been applied to the analysis of a corpus of English and Dutch written texts.



## 2.2 The continuum of spoken and written language

An analysis of various approaches to discourse segmentation shows that they can roughly be divided into two main groups: approaches that focus on the analysis of spoken discourse and those that focus on the analysis of written discourse. As the present study has been restricted to the analysis of written discourse, it may seem logical to also limit the overview of approaches to those that deal with the analysis of written discourse. However, a closer analysis of the four genres included in the corpus shows that these do not form a uniform group in terms of discourse structure. On the one hand, academic prose, newspaper articles, short stories and public information leaflets can all be considered to belong to the written mode of discourse, simply because they constitute printed text. On the other hand, some of these genres contain a range of linguistic features that are more typically associated with spoken discourse. An analysis of the discourse structure of sentences in these four genres is thus inherently related to the question to what extent spoken and written discourse differ from each other and to whether the discourse produced in either modality calls for a different type of analysis.

Studies that have compared and contrasted the two modalities point to the different functions that speech and writing have and to how these diverging functions affect the linguistic realisation of text produced in either modality (cf. Chafe 1982; Tannen 1982; Chafe & Danielewicz 1987; Halliday 1989, 1994; Nunberg 1990; Biber 1995; Biber et al. 1999). In very general terms, speech can be characterised by the fact that it is produced online, with the words and grammatical construction being composed 'on the spot' (Biber et al. 1999: 9), which characterises it as faster and evanescent (Chafe & Danielewicz 1987: 9, 13). It is characterised as being personal and interactive (Biber et al. 1999: 16), involved (Biber 1995: 161; Chafe & Danielewicz 1987: 23), and containing situation-dependent reference and non-abstract content (Biber 1995: 161).

Writing, on the other hand, can, in very general terms, be characterised as being informative, explicit, abstract (Biber 1995: 163), as a planned activity that can be edited and revised (Chafe & Danielewicz 1987: 8-9, 22; Biber et al. 1999: 9), as slower than speaking (Chafe 1982: 37; Chafe & Danielewicz 1987: 8), as more integrated, i.e. allowing for more linguistic devices that integrate information into one unit (Chafe 1982: 45), and as less or non-interactive and more detached (Chafe 1982: 45; Chafe & Danielewicz 1987: 23; Biber et al. 1999: 16).

These are just broad characterisations of the two modes and Biber rightly points out that such broad generalizations 'do not adequately describe the details of the relations between speech and writing' (1995: 47). He emphasises that these characteristics cannot be applied to all spoken and written genres and suggests that there is no absolute difference between the two modes (1995: 163, see also Tannen 1982; Chafe & Danielewicz 1987: 5, Halliday 1989: 32 for a similar view). Instead of presenting the differences as a clear dichotomy and as two discrete poles, he suggests describing the differences as continuous, as continuums of variation (1995: 9, 22). He characterises a genre as the *extent* to which it belongs to the spoken or written mode (1995: 162-163, italics in original). Certain types of discourse can then be described in terms of being more or less typical for a particular mode (1995: 161). For instance, Biber places academic prose at the far end of the written continuum and conversation at the far end of the spoken continuum, by categorising the former as 'typical writing' and the latter as 'typical speech' (p. 164).

When applying this notion of a continuum to the four genres included in this study, the academic prose genre and the short stories genre, containing considerable sections of simulated dialogue, take up different positions on the continuum. This is in line with Biber et al.'s classification of the fiction genre as being 'intermediate' between spoken and written language, also because of the conversational dialogue (1999: 16). These two genres can also be distinguished from each other in terms of purpose of the text types, with the academic prose having a more 'informational focus', whereas fiction is associated with 'pleasure reading' (*ibid*). With respect to the other two genres included in this study, newspaper articles and public information leaflets, these could then be placed in between academic prose and short stories. Unlike the academic prose genre, but similar to the short stories genre, these genres have a diverse and wide readership, and unlike the short stories genre, but similar to the academic prose genre they have an informational purpose. One further characteristic that distinguishes academic prose from especially the newspaper and leaflets genres is its global and specialist nature, written for an international audience (*ibid*). As Biber et al. point out '[e]ven a casual inspection of texts from different registers reveals extensive linguistic differences' (1999: 9), which can best be systematically identified not on the basis of an analysis of individual text, but on the basis of corpus analysis (p. 11).

Moreover, in addition to the different situational and linguistic differences between the different genres on the spoken-written continuum, some also point to a difference between the two modes in cognitive or psychological terms, which

could affect the type and size of information units (eg. Chafe & Danielewicz 1987; Biber 1995). Biber, for instance, characterises these differences in terms of other 'cognitive constraints' that are placed on speakers and writers (1995: 160ff). Specifically, speaking is seen as having certain processing constraints, because it is produced and comprehended in real-time, which sets a 'cognitive ceiling' for the level of syntactic and lexical complexity (p. 163). Written language, on the other hand, shows a higher frequency of 'literate features', such as informational density, careful word choice, explicit reference and the use of abstract information, than spoken language (1995: 163). A similar view is presented by Chafe and Danielewicz (1987: 14-16), who also link this to the type and size of information units in either modality, where the unit in spoken discourse is typically smaller in size and contains less information than the unit produced in written discourse (1987: 14-15).

With respect to the analysis of the discourse structure of the sentences included in this study, it could benefit from insights gained from both spoken and written approaches to discourse segmentation, as the sentences are taken from genres that take up different positions on the continuum of spoken and written language.

## **2.3 Overview of approaches to discourse segmentation**

This section will present an overview of various approaches to discourse segmentation. These were categorised in two ways. First, a distinction was made between approaches that can be classified as describing discourse segmentation from a more theoretical perspective as opposed to those that approach the issue from a more practical and applied perspective. Second, a distinction was made between approaches that describe the process of discourse segmentation in the spoken language and those that focus on the written language.

The main aim of this section is twofold. On the one hand, it serves to provide insight into the complexity of discourse segmentation, by showing and explaining how a wide number of approaches have tackled this issue. On the other hand, it serves to identify the criteria put forward to identify basic units in discourse consistently. This latter aim relates to the aim of this chapter of finding or developing an approach to discourse segmentation that can be applied consistently to the analysis of data.

### 2.3.1 Discourse segmentation: from theory to data

A number of the theoretical approaches discussed in this section can be divided into two groups: those that focus on what constitutes a basic unit in conversation and those that focus on what constitutes a basic unit in the written genres. The approaches that do not make this distinction as explicitly will be presented as if they do, by describing in separate sections how they approach discourse segmentation in spoken language and how they approach this in written language. The main motivation for this division is to achieve clarity by grouping the different views on what constitutes a basic unit of discourse in conversation and what constitutes a basic unit of discourse in the written language.

#### Theoretical approaches to the analysis of spoken data

In their search for an approach to the analysis and segmentation of actual spoken data, Foster, Tonkyn and Wigglesworth (2000) present a fairly extensive overview of different models. They explain that their search led to ‘a plethora of definitions of units of analysis, a paucity of examples, and a worrying tendency for researchers to avoid the issue altogether by drawing a veil of silence over their methods’ (2000: 357). They group the various definitions into three broad categories that take a semantic, intonational or syntactic approach. Examples of semantic units are *proposition* (Sato 1988), *the c-unit* (Pica et al. 1989) and *the idea unit* (Kroll 1977). Foster et al. argue that segmenting discourse on the basis of semantic criteria ‘will tend to be extremely hard for the analyst to work with reliably’ (2000: 358), as it has to be determined where one idea unit or proposition ends and where the next one begins. The second category they identify concerns units that have mainly been determined on the basis of intonational features, examples of which are *the tone unit* or *phonemic clause* (Crystal & Davy 1975) and *the intonation unit* (Chafe 1994). As segmentation on the basis of intonation is mainly determined by a rise or fall in pitch and by pauses, Foster et al. argue that these criteria cannot be easily applied in analysing the speech of non-native speakers, for instance, since the pauses they produce may not always indicate unit boundaries (2000: 359). They therefore only consider intonation a useful criterion to identify unit boundaries when used in tandem with other criteria, such as syntactic ones (2000: 359). The third category they identify concerns units that have mainly been determined on the basis of syntactic criteria, with a rather popular one being that of the *T-unit* (cf. Hunt 1966, as cited in Foster et al. 2000: 360). This has been adopted by others in various ways, but can generally be defined as constituting a clause with or without

subordinate or embedded clauses attached to it. Because some considered the T-unit inadequate for analysing spoken language, many have used the *C-unit* instead, originally described by Loban (1966) to deal with the elliptical nature of the spoken language (Loban 1966), but, again, modified and extended by many others (Foster et al. 2000: 361). Foster et al.'s overview aptly illustrates that there are a great number of approaches that have identified the need for a basic unit of discourse in the analysis of spoken language and that have made an attempt at defining and describing this basic unit. The following paragraphs will describe a selection of approaches in more detail, namely the ones put forward by Chafe (1994), Halliday and Matthiessen (2004), Langacker (2001), Steen (2005) and Hannay and Kroon (2005).

First, in his analysis of spoken data, Chafe (1994) observed that the flow of speech is not continuous, but that it is produced in spurts. These spurts are considered basic functional segmentations of discourse and are labelled 'intonation units'. Intonation units play an important role in both the production and comprehension of language and are seen as verbalizations of the information that is active in the speaker's mind at their onset (1994: 62-63). Possible features of these intonation units are changes in pitch, changes in duration (which is perceived as the shortening or lengthening of syllables or words), changes in loudness, pausing, changes in voice quality and sometimes changes of turn (1994: 58). An analysis of one or more of these features helps in identifying intonation units and in distinguishing them from each other. Because not all intonation units have a similar make-up or function, Chafe distinguishes between three main types: 1) substantive intonation units, 2) regulatory intonation units, and 3) fragmentary intonation units. The first type, the substantive intonation unit, conveys substantive ideas of events, states or referents. In the samples analysed by Chafe, it appears that they have an average length of four words and are typically realised as clauses in around 60% of the cases (1994: 65). The second type, the regulatory intonation unit, regulates the interaction or information flow. This type usually consists of only one word and is typically realised linguistically as a unit that has the function of a discourse marker (1994: 64). Moreover, the regulatory intonation units can be further subdivided into units that have a textual function, an interactional function, a cognitive function and a validational function (1994: 63-64). Even though these are presented as distinct functions, it is explained that the distinctions between them may not always be clear-cut (1994: 64). The third type, the fragmentary intonation unit, constitutes truncated or unfinished units. As they have not been completed, nothing can be said about their average size or linguistic

realisation. The following piece of dialogue illustrates the various types of intonation units and their functions (taken from Chafe 1994: 63-64):

- |     |  |                       |
|-----|--|-----------------------|
| (2) | 1. .... Well,                                    | (regulatory: textual) |
|     | 2. isn't she healthy?                            | (substantive)         |
|     | 3. ....Mhm,                                      | (regulatory:          |
|     | interactional)                                   |                       |
|     | 4. ....I mean she                                | (fragmentary)         |
|     | 5. I know she has                                | (fragmentary)         |
|     | 6. More or less                                  | (substantive)         |
|     | 7. ... She has [something with her] gallbladder, | (substantive)         |
|     | 8. [gallbladder and, ]                           | (substantive)         |
|     | 9. .... heart trouble and,                       | (substantive)         |
|     | 10. [back problems.]                             | (substantive)         |
|     | 11. [She has heart ] trouble,                    | (substantive)         |
|     | 13. ... Her she has an enlarged heart.           | (substantive)         |

Even though the list of features and varying functions of the different types of intonation unit can help in identifying and distinguishing them from each other, Chafe acknowledges that consistently segmenting speech into intonation units is a skill that needs to be practised and guided by an experienced transcriber (1994: 62). He further acknowledges that his ideas about the different types and functions of intonation units would benefit from more extensive and varied analysis of spoken language data (1994: 70).

Showing overlap with Chafe's notion of the intonation unit is Halliday and Matthiessen's definition of the information unit (2004). In Halliday and Matthiessen's view, grammar manages the flow of discourse by means of two related systems, one being the system of the clause and the other the system of information. The information unit is the unit of the system of information, where information is seen as the tension between what is already known or predicted and what is new or unpredictable (Halliday & Matthiessen 2004: 88-89). The information unit is made up of two functions, the New and the Given, and in an idealised form each unit consists of both a given and a new element, with the given element preceding the new element. In unmarked cases, the information unit is parallel to the clause and in marked cases the information unit can be smaller or bigger than the clause (2004: 88). The notion of information unit is not restricted to either the spoken or written language, but can be used to describe the discourse structure of both modalities. In speech, the information unit is realised as a tone group that may be rising or mixed, and contains a foot that receives tonic

prominence, which is the element that carries the information focus. An analysis of the characteristic features of a tone group could then help in identifying and distinguishing between different information units. Whereas Chafe makes a clear distinction between different types of intonation unit and identifies three main types, Halliday and Matthiessen identify only one type, but do distinguish between unmarked or default conditions, in which an information unit is, for instance, co-extensive with a clause, and marked conditions, in which an information unit is bigger or smaller than the clause.

Also related to Chafe's notion of the intonation unit is Langacker's concept of 'a window of attention' (2001: 154). In an attempt to capture the relation between discourse and cognitive grammar, Langacker describes the basic unit of discourse in relation to what he labels 'attention framing', which is a level of organisation where a single window of attention is to be identified with Chafe's substantive intonation unit. At a conceptual level these frames 'consist of information fully active in the mind at one time; grammatically they tend to coincide with single clauses' (2001: 154), similar to the typical or unmarked linguistic realisations of Chafe's intonation unit and Halliday's information unit. They can, however, consist of only a part of a clause that is stretched out into more than one window or of more than one clause (2001: 154). In Langacker's view, the function of these windows of attention is to update the Current Discourse Space (CDS), the mental space shared by the speaker and hearer as a basis for communication. Attention frames are explained as being mapped onto or imposed on structures, which are to a large extent arranged hierarchically. These frames can, however, be mapped onto structures in different ways. Example (3) represents two different framings for an *if...then* sentence (taken from Langacker 2001: 156):

- (3)      a. If she said it, then it's true.  
           b. If she said it then it's true.

In (3a) the expression is parsed into two attention units and in (3b) it is 'squeezed' into a single attention unit (p. 156). Langacker explains that although the meaning is the same in either framing and although the expression has the same conceptual content, there is a subtle semantic contrast. In the case of (3a), by giving each clause separate unit status – or 'dwelling on each clause individually' (p. 157) – it enhances its cognitive salience. Sentence (3b), on the other hand, represents a case of phonological and conceptual compression, where more has to be squeezed into a single and limited set of processing time (2001: 157-158). He then emphasises

that presenting the expression as one or two attention units is a way of construing a different meaning and that the possibility of presenting ideas in different ways, by using various framing options, may constitute a conventional linguistic pattern. One particular framing may then be unmarked or prototypical, but others that occur may also be familiar or 'conventionally sanctioned'. And although there is certainly room for flexibility in how expressions are framed, it is partially shaped by convention, 'a matter of constrained freedom rather than unbridled license' (2001: 162). This seems to emphasise that a speaker can choose to present his thoughts in various ways, using one or more units and using various linguistic expressions, but is guided by conventions. In contrast with Chafe, Langacker focuses on only one type of intonation unit or attention frame – the one that coincides with Chafe's substantive intonation unit – and does not explicitly distinguish between different types of attention frames, although he does distinguish between unmarked and marked cases. And where Chafe provides a list of features of the various types of intonation units that help in identifying them and distinguishing them from others, Langacker does not provide such a detailed list. He only states that attention frames are typically realised as clauses. Moreover, he does explicitly state that his main aim is to present his theory and not to test it using actual discourse data or detailed analyses (2001: 144). He encourages other scholars to pursue this aim.

Steen (2005) appears to have accepted Langacker's indirect invitation to investigate the relationship between cognitive grammar and discourse more systematically, as he presents a theory of basic discourse units that takes Langacker's proposal as a starting point. He underlines the relevance of identifying basic discourse units by explaining that they are 'an important tool for language users when they have to break continuous text and talk up into equivalent segments for cognitive processing' (2005: 284). Similar to Langacker's approach, Steen's basic discourse unit – or Basic Discourse Act (BDA) – consists of an intonation unit, a clause and a proposition. However, Steen adds another defining value of BDAs, which is their communicative role, typically manifested by an illocutionary act. He presents his model as a multi-dimensional approach in that it analyses discourse from a material, linguistic, conceptual and communicative dimension. Furthermore, he does not explicitly restrict his discussion to the analysis of spoken language, in which the unit of analysis is the intonation unit, but also incorporates the analysis of written language, in which the unit of analysis is the punctuation unit (see section on analysis of written language below). With respect to the types of basic discourse acts, Steen distinguishes between two main types: typical or less typical BDAs. Typical BDAs consist of one illocutionary act, one



proposition, one clause and one substantive intonation unit or punctuation unit. Less typical BDAs, on the other hand, can be characterised by the fact that they either have more than one salient value on one or more dimensions or zero salient values on one or more dimensions (2005: 302). The illocutionary act, however, forms an exception to this rule, as this one cannot be removed from a BDA, nor can one BDA consist of more than one illocutionary act. In other words, illocutionary acts appear to play an important role in identifying the basic units in discourse. Moreover, Steen emphasises that in describing BDAs, reference should always be made to all four dimensions and that these dimensions are independent of each other (2005: 290), as reference to only one or two dimensions cannot capture the full complexity of the function of BDAs (2005: 290, 297). To illustrate what he means by this, let us consider Steen's analysis of the same *if...then* sentence that was presented above and repeated below for convenience as (4).

- (4)      a. If she said it, then it's true.  
           b. If she said it then it's true.

In his analysis, (4b) can be regarded as a basic discourse unit because it contains one complex proposition, one illocutionary act and one intonation unit (2005: 297). However, because it consists of two clauses instead of one on the material dimension, it could be classified as a 'less typical' basic discourse act. Steen points to the added value of taking a multi-dimensional approach by contrasting it with a mono-dimensional approach such as Mann and Thompson's Rhetorical Structure Theory (1988, see Section 2.3.1 below), in which (4b) could be analysed as consisting of two units only because it consists of two clauses on the material dimension. By taking all four dimensions into account the 'discourse unity' of (4b) as one whole becomes clear (2005: 297).

In addition to the distinction between basic and less basic discourse acts, Steen also gives examples of non-basic discourse acts. Consider in this respect (5), taken from Steen (2005: 296):

- (5)      On the desk, he found an important-looking document.

In Steen's analysis, (5) as a whole is analysed as one less typical basic discourse unit, because it has one proposition, one clause, one illocution, but two intonation units. The two intonation units are then labelled non-basic discourse units, because they are a part of a (less) basic discourse unit and the first intonation unit, 'on the desk',

is dependent on the complete unit (2005: 296, 303). This means that these non-basic units are only units on one dimension of discourse, the material dimension. In order to be classified as basic discourse units, they have to be considered units on the other dimensions as well. Moreover, non-basic units are typically realised by non-salient categories for each of the four dimensions. This means, for example, that the unit is realised as a dependent clause or phrase on the lexico-grammatical dimension instead of an independent clause, or that it is realised as a regulatory intonation unit instead of a substantive intonation unit on the material dimension.

Steen thus appears to present a sort of continuum when it comes to defining basic discourse units, where typical basic discourse units consist of one salient, independent value on each of the four dimensions, less typical units either contain more or less of these features that are more or less dependent on others, and non-typical units contain zero salient values. An analysis of the presence or lack of salient values can then be used in identifying these units in discourse. Moreover, the distinction between typical and less typical could also be used to describe certain features in spoken language, such as answers to questions, phrases of greeting or perhaps even gestures (2005: 300). With respect to the analysis of real discourse data, Steen's approach looks promising in the sense that it has a flexible design as it allows for different manifestations of units in language. His ideas of typical and less typical discourse units are, however, mainly based on theoretical assumptions. Whether his proposal can be applied to real language data and whether the same proposal would work for the analysis of both spoken and written data is a question that Steen raises himself and answers by stating that this is an empirical question that needs to be decided with reference to large-scale corpus research (2005: 308).

In comparison to the other models discussed so far in this section, Steen's model distinguishes itself by taking a multi-dimensional approach to discourse analysis. At the same time, it shows overlap with the other approaches in that he makes a distinction between basic and non-basic unit, which is similar to the distinctions presented between unmarked and marked, typical and non-typical or Chafe's distinction between the different types of intonation unit.

Another theoretical approach to discourse segmentation is presented by Hannay and Kroon (2005), who present it within the framework of Functional Discourse Grammar (Hengeveld 2004). Their aim is to shed more light on the relation between discourse and grammar, not restricting themselves to either the analysis of spoken or written discourse. In their approach, the basic unit of discourse or basic discourse act is seen as each single communicative step that a

language producer takes in order to achieve his communicative aim (2005: 95). Hannay and Kroon see these acts as elements of discourse planning, which means that they are essentially cognitive in nature, and explain that discourse is planned at both a conceptual level and a strategic level. Because discourse is planned at these separate levels, Hannay and Kroon also make a distinction between the units of analysis at both levels. At the conceptual level, they refer to the basic units of discourse as ideas. At the strategic level, where ideas have been turned into steps ready to be executed by the speaker, they refer to the basic units of discourse as acts (2005: 103). The acts are then further subdivided into two main types, substantive acts and regulatory acts, following Chafe's distinction between substantive intonation units and regulatory intonation units (1994). To illustrate how their model is applied to the analysis of discourse, consider example (6), taken from Hannay and Kroon (2005: 100):

(6) Last year I was robbed | by a four year old.

In Hannay and Kroon's analysis, (6) consists of one idea at the conceptual level that has been presented in two steps, or acts, at the strategic level, because (6) consists of two intonation units. As an example of their subdivision of acts into substantive and regulatory acts, consider example (7) (2005: 22, taken from Dik 1997: 387):

(7) Well, Ladies and Gentlemen, shall we start the game?

This example is analysed as consisting of two regulatory acts and one substantive act. The regulatory acts can be further specified by describing their functions in discourse. In this example *well* is said to have a text organizational function and *Ladies and Gentlemen* a hearing involving function, again following Chafe's classification of functions of regulatory intonation units.

With respect to the relations between different substantive acts, Hannay and Kroon propose a further distinction between main acts and subordinate acts. Following Mann and Thompson (1988), a similar system of rhetorical relations between substantive acts is assumed. Consider in this respect (8), which consists of two substantive acts, with the second act being related to the first one by the rhetorical relation of justification (2005: 106, taken from Stenström 1994: 44):

(8) Wednesday is quite a good day | cos I don't teach at all on it.

As for the regulatory acts, although these could be considered subordinate to the discourse unit to which they are related, they cannot be described in terms of rhetorical relations because they mainly specify what the rhetorical relation is between two substantive acts and are therefore described in terms of 'management relations' (2005: 106).

With respect to the linguistic realisation of discourse acts, Hannay and Kroon suggest that these are realised as intonation units (IU) in spoken language and punctuation units (PU) in written language. Their definition of intonation units corresponds to Chafe's definition of intonation units. Punctuation units, on the other hand, are defined as 'any stretch of language occurring between the members of any given pair of correlative punctuation marks, regardless of the possible occurrence of interpolated punctuation marks' (2005: 107). By excluding so-called interpolated punctuation marks, serial punctuation marks that separate items on a list, such as *red, yellow and blue*, are not taken to signal unit boundaries. With respect to the relation between intonation units and punctuation units, Hannay and Kroon follow Chafe (1988, 1994), who suggests looking at punctuation units as if they reflect intonation units, taking differences in punctuation style into consideration. The difference between intonation units and punctuation units, in their view, is that the former typically reflect a single focus of consciousness and the latter 'extended foci of consciousness', which often makes them bigger than their spoken language intonational counterparts (2005: 108). In line with Biber (1988, see 2.2 above), they state that this difference in length could be seen as reflecting a difference between the spoken and written system, in which attention is drawn to the online-production capacity of the former system and the planning capacity of the latter system, causing writers to perform relatively more 'content-rich' acts (2005: 25). Hannay and Kroon's ideas about punctuation units will be described in more detail in the section on theoretical approaches to the analysis of written discourse below.

In order to support their claim that in oral English discourse the strategic organisation of discourse is more strongly reflected in the prosodic than the syntactic structure, they present cases in which there is a mismatch between syntactic units on the one hand and intonation units on the other hand. For instance, in (9) two syntactic units are presented as one intonation unit (2005: 109):

(9) So England lost their victory did they.

Instead of identifying the tag as a separate syntactic unit and therefore granting it separate unit status, Hannay and Kroon point out that this tag can be distinguished from other types of tag because it is a case of positive polarity and does not have the function of seeking response from the listener. Instead, its function is simply to draw a conclusion from the previous discourse (2005: 109). Because of this function and because the tag is integrated into one intonation unit, example (9) is taken to consist of one act, realised as one intonation unit. The opposite situation could, however, also occur, where one syntactic unit is presented as two intonation units, as in example (10) (2005: 110, taken from Brazil 1995: 185):

(10) This old lady | told her about her daughter.

In their analysis of this example, they suggest that the first intonation unit could, on the one hand, be considered as a regulatory act that has a preliminary announcement function, or 'launching' function. On the other hand, it could also be considered as being part of the substantive intonation unit, in which case it functions as a subact which has a presentative pragmatic function (2005: 110ff). The analysis of (10) is, however, not considered to be unproblematic and is identified to represent one of the most extreme cases of mismatch between discourse structure and syntactic structure (2005: 112).

Hannay and Kroon conclude their proposal by suggesting that intonation units and punctuation units should be taken more seriously in segmenting discourse structure. In line with the other approaches described in this section, their main objective is to present a plausible theory of discourse segmentation, which they support with a number of examples, but which has not yet been applied to the analysis of real language data. Similar to the other approaches, they also make a distinction between different types of act – substantive and regulatory acts – and explain that these can have different functions in discourse. The main criterion put forward for the identification of these acts in discourse is an analysis of their linguistic realisation, which are intonation units in the case of spoken discourse and punctuation units in the case of written discourse.

In addition to providing insight into the wide variety of approaches there are to the segmentation of spoken discourse, this section has focused on the defining features of these basic units of analysis and on the criteria presented to distinguish them from each other. What becomes clear is that many approaches conceive the basic unit of spoken discourse as some sort of intonation unit, where distinctions are made between different types of unit and between different

functions that units can have. Most approaches agree that the typical or unmarked linguistic realisation of an intonation unit is a clause, with less typical or more marked linguistic realisations being something other than a clause. Some of these approaches, such as Hannay and Kroon's model, explicitly address the hierarchical relation between different types of unit, distinguishing between superordinate and subordinate units, and others seem to imply a similar distinction, by using labels such as substantive and regulatory or independent and dependent. What these models have in common is that they approach the subject of discourse segmentation from a theoretical perspective, supported by means of examples. This distinguishes them from the approaches that will be discussed in Section 2.3.2, which have all been applied to the analysis of actual spoken data.

### **Theoretical approaches to the analysis of written data**

In exploring the relationship between spoken and written language, Chafe (1994) not only identifies the basic unit of analysis of the spoken language, the intonation unit, but also identifies a similar unit for the analysis of written language, the punctuation unit (1988). In his view, punctuation represents the prosody of written language just as intonation represents the prosody of spoken language. Punctuation is seen as making the covert prosody of written language more overt and as providing insight into the writer's prosodic intent (1988: 397). Using punctuation effectively is considered a skill that can be acquired, similar to learning how to use certain devices, such as pitch and hesitations, in speech effectively (1988: 397ff). Using punctuation effectively or ineffectively is seen as the extent to which a writer has learned to listen to his inner voice, which gives him information about the prosodic units – or punctuation units – in writing. Following the prosody of written language intuitively is, however, sometimes prevented by certain 'disturbing factors', such as the level of skill in punctuation, the influence of current fashions in punctuation and the influence of certain rules for punctuating, not all of which are prosodically motivated, but may be placed at grammatical boundaries or be completely arbitrary (1988: 400-401, 410). As an example of a conflict between a grammatical rule and prosody, Chafe provides the following example, in which the subject is separated from the verb by means of a comma (1988: 404):

- (11) Two cups of Quaker 100% Natural Cereal mixed with a little of this and a little of that, make the best cookies you've tasted in years.

Chafe argues that even though it constitutes a grammatical rule that subjects should not be separated from verbs in English, instances of cases in which this happens can be found, probably because writers are inclined to separate a long subject from a verb because this marks a natural intonation break (1988: 404). In order to establish to what extent punctuation reflects the prosody of written language, Chafe carried out a small-scale reading aloud experiment and a repunctuating experiment. These experiments show that punctuation units are typically longer than intonation units, a result he relates to differences in processing constraints between spoken and written language (see 2.2 above). The experiments also show that the length of the punctuation unit is dependent on the genre in which it occurs. For instance, academic texts typically contain longer punctuation units than advertisements. In taking punctuation seriously in marking the prosody of written language, Chafe also addresses the issue of variation in punctuation (p. 399ff). He explains that there is room for variation and that this is mainly influenced by varying styles of punctuation, some of which may be a reflection of the time period in which a text was produced. This variation means that readers might have to rely more on syntax or language itself to segment a text prosodically. Punctuation, in this respect, is then also considered a means to increase transparency in the intentions of the writer. Even though Chafe acknowledges that there are a number of disturbing factors that may make it difficult to identify punctuation units in writing or that may make it difficult to always see punctuation as reflecting the prosody of written language, his main conclusion is that on the whole punctuation does reflect the prosody of the written language and provides insight into the writer's prosodic intent.

In line with Chafe's view, Hannay and Kroon (2005) make a similar distinction between the units of spoken discourse, intonation units, and the units of written discourse, punctuation units. Their description of intonation units was presented in the section on the spoken approaches to discourse above and the current section will focus on their definition of the punctuation unit in more detail. As was already announced in the introduction to this section, the description of both types of units has been separated for the sake of clarity, even though the basic theoretical foundations are similar. In relation to Chafe's punctuation unit, Hannay and Kroon's account of punctuation units is more detailed, as they provide a clear definition of what these units entail and distinguish between different types of unit. Punctuation units are defined as 'any stretch of language occurring between the members of any given pair of correlative punctuation marks, regardless of the possible occurrence of interpolated punctuation marks' (2005:

107). As for the latter type of punctuation mark, it means that the use of serial punctuation marks, as in *red, yellow and blue*, is not taken to mark unit boundaries (*ibid*). Furthermore, they identify four main types of punctuation unit: core units on the one hand, and prepended, appended and interpolated units on the other hand (2005: 113). These labels imply that punctuation units can be related to each other hierarchically, with core units expressing nuclear information and prepended, appended and interpolated units expressing satellite information. This distinction is related to the one Hannay and Kroon make between intonation units that can execute a substantive act or a regulatory act (see above). Punctuation units can also execute a substantive act or a regulatory act. The following sentences, taken from Hannay and Kroon (2005: 113), give examples of each of the four types of punctuation units, executing different types of act:

- (12) However, cases like this are few and far between.
- (13) They visit many schools, sometimes in an official car. (Quirk et al. 1985: 912n)
- (14) John, and Sally too, writes extremely well. (Quirk et al. 1985: 976)

Example (12) is analysed as consisting of two punctuation units, separated by a comma. The first unit, realised as the conjunct *however*, constitutes a prepended punctuation unit that executes a regulatory act and precedes the core unit that executes a substantive act (2005: 113). Example (13) also consists of two punctuation units: a core unit that is followed by an appended unit, both of which execute substantive acts. Finally, example (14) also consists of two punctuation units, but here the core unit is interrupted by an interpolated unit.

Similar to Chafe, Hannay and Kroon discuss a number of cases of punctuational variation. They relate these cases of variation, such as the use of a punctuation mark between coordinated clauses or in marking a clause as a restrictive or non-restrictive one, to the choice that writers have between presenting information as one or two separate units, and thus to performing one or two discourse acts (2005: 114-115). This is in line with Chafe's view of punctuation as reflecting a writer's prosodic intent.

As for the identification of punctuation units in text, Hannay and Kroon do provide examples of the different types of punctuation units they have identified, but do not describe in any great detail how a punctuation unit can be identified consistently in the analysis of written text. Specifically, they do not provide clear criteria to identify or distinguish between different types of punctuation units and



the acts they can execute. Their proposal is thereby mainly theoretical in nature and serves the purpose of making clear why punctuation should be taken seriously in signalling the boundaries of discourse acts. In fact, they even suggest that it would be a 'valuable enterprise to take punctuation out of the realm of stylistics and to develop a discourse theory of punctuation' (2005: 115).<sup>5</sup>

Similar to Chafe (1988) and Hannay and Kroon (2005), Steen (2005) does not restrict his discussion of basic discourse acts – discussed in detail in the section on theoretical approaches to the analysis of spoken language above – to either the analysis of spoken or written discourse. In his multidimensional approach to discourse analysis, he defines basic discourse acts with reference to the four dimensions of illocution, proposition, clause and intonation or punctuation unit, and distinguishes between basic and less basic discourse acts. However, where Chafe (1988) and Hannay and Kroon (2005) really make a case for seeing punctuation as signalling boundaries between discourse units, Steen does not state this as explicitly, but does follow them in this respect. He also points out that the notion of punctuation unit has been less well researched than the notion of intonation unit (2005: 293) and he addresses the question to what extent text and talk should be seen as two varieties of discourse or to what extent they should be seen as different modes that each need their own approach to analysis. He merely raises this question, however, and does not answer it (2005: 309).

A well-known example of an approach that does not necessarily problematise the notion or identification of a basic unit of discourse, but instead concentrates on the rhetorical relations that can exist between different units in texts is Mann and Thompson's (1988) proposal of Rhetorical Structure Theory (RST). Their approach can be classified as essentially a descriptive theory that presents a method for text analysis that characterises the structure of texts in terms of the rhetorical relations that hold between parts of the text. The reason for including their approach to the analysis of written text in this overview is because they argue that relations between text spans are hierarchically structured, making a distinction between nuclear and satellite text spans (1988: 245). The identification of a hierarchical structure of texts has been adopted by many approaches to the

---

<sup>5</sup> Even though Hannay and Kroon do not apply their notion of the punctuation unit to the analysis of text, it should be noted that Hannay (1997) has introduced the notion of punctuation unit in an earlier study, in which he not only provides identification criteria for punctuation units, but also applies these to the analysis of written text. This approach will be described in more detail in Section 2.3.2 below.

analysis of written text, such as Hannay and Kroon (2005) above, who make a distinction between punctuation units that express nuclear information and those that express satellite information. To distinguish between a nucleus and a satellite, Mann and Thompson explain that either term refers to the extent to which the ideas presented are more or less dependent for their meaning on the other unit, more or less suitable for substitution and more or less essential to the writer's purpose, with nuclei typically being independent for their meaning on other units, less suitable for substitution and more essential to the writer's purpose (1988: 266). Mann and Thompson explain that the first step in analysing a text is by dividing it into units and they identify clauses as the basic units of analysis, with the exception of clausal subjects and complements and relative clauses (1988: 248), and with the addition that not all propositions are clausally expressed (1988: 259). They distinguish between structural relations, or assertions, and relational assertions and explain that the former typically take the form of a clause, whereas this need not apply to the latter (1988: 259). It is remarkable that they do not present the identification of basic units in discourse as problematic, but merely state that these essentially correspond to clauses, and focus on the identification of the different relations that can exist between different units of text instead (1988: 248). Examples of these relations are circumstance, antithesis and concession, evidence and justify, and relations of cause (1988: 250). Consider the following extract, which presents an example of an evidence relation between a nuclear unit and two satellite units:

- (15)     1. The program as published for calendar year 1980 really works.  
           2. In only a few minutes, I entered all the figures from my 1980 tax return  
           and got a result which agreed with my hand calculations to the penny.

Example (15.2) is analysed as being in an 'evidence' relation with the nuclear unit in (1). It is provided to increase the reader's belief in the claim expressed in unit (1) (1988: 251). However, as Mann and Thompson's approach is mainly descriptive of the relations that exist between text spans and does not necessarily question the notion of what constitutes a basic unit of discourse or how a distinction can be made between a nuclear unit and a satellite unit, their theory will not be described in more detail here, but instead the focus will be on presenting two approaches that have applied the basic concepts of RST to the analysis of written text and addressed the question of what constitutes a basic unit of discourse in doing so. The first is Dale's (1991) theory on the role of punctuation in signalling rhetorical

relations, which will be discussed in more detail in the present section, and the second is Carlson and Marcu's (2001) application of RST to the annotation of a corpus of written texts, which will be discussed in more detail in Section 2.3.2 below.

As punctuation has been identified by several authors as potentially playing an important role in marking unit boundaries in the written language, it is interesting in this respect to make reference to Dale's study (1991), in which he relates the use of particular punctuation marks to the signalling of rhetorical relations between different units in discourse. Dale himself takes his inspiration for assigning such an important role to punctuation in written language from Nunberg's (1990) study of the linguistics of punctuation and argues that punctuation markers, as well as lexical markers and graphical markers, act as signals of discourse structure, where he focuses on the semantic aspects of punctuation marks (1991: 111). Similar to Chafe (1988) and Hannay and Kroon (2005), Dale considers punctuation decisions to be closely related to the chunking of information. Moreover, he points out that most theories of discourse structure take the clause as the basic unit of analysis and argues that a theory of discourse should also operate above or below the level of the clause (1991: 118). He provides the following example, (16), to show that even though the non-restrictive clause is classified as a minor clause, this in itself does not provide a reason to discard it from the analysis, as it can be described as being in some sort of rhetorical relation with the clause it interrupts (1991: 117):

(16) Knox, which is C4, is enroute to Sasebo.

Moreover, in relation to discourse structure, Dale argues that punctuation marks play at least some role in the interpretation of certain rhetorical relations. He presents the following sentences – taken from Nunberg (1990) – to prove his point (1991: 114):

(17) He reported the decision: we were forbidden to speak with the chairman directly.

(18) He reported the decision; we were forbidden to speak with the chairman directly.

(19) He reported the decision – we were forbidden to speak with the chairman directly.

Dale argues that the choice for different punctuation marks is not a random one: it provides insight into the rhetorical relations that hold between the punctuation units. He characterises these differences in terms of RST. For instance, in example (17) the unit introduced by the colon can be interpreted as an *elaboration* on the unit that precedes it, giving the content of what the decision involved. In (18), on the other hand, the unit introduced by the semi-colon can be considered a *cause*, which gives the reason for why the decision was reported. Example (19), however, can express either of these rhetorical relations. These examples are provided to illustrate that punctuation marks 'at least play some role' in the interpretation of particular rhetorical relations, but that there is not a straightforward one-to-one mapping (1991: 116).

In an attempt to determine what the role of punctuation is in indicating discourse structure, Dale presents a taxonomy of the specific functions performed by punctuation marks. He bases this taxonomy on constraints indicated by style guides on the use of punctuation and suggests that punctuation marks can indicate the degree of rhetorical balance, aggregation and particular rhetorical relations. The function of rhetorical balance relates to how different punctuation marks give information about the relative 'weight' of units connected by the marks in terms of what unit functions as a nucleus and what unit functions as a satellite (p. 118). The function of aggregation gives information about how closely units are related and indicates different degrees of relatedness. This implies some sort of hierarchy of the various punctuation marks. Last, a function of punctuation marks is also to indicate what semantic or rhetorical relations hold between different units, with colons, for example, being used to either emphasise something or to illustrate, amplify or explain something, as in the following example (1991: 118-120):

- (20) Many of the policemen held additional jobs: thirteen of them, for example, doubled as cab drivers.

However, Dale acknowledges that his categorisation of the functions of punctuation marks is very 'impressionistic' and needs further work to determine whether punctuation can indeed be seen as signalling discourse structure (p. 120). In addition to the other theories presented in this section, it should be noted that by providing the taxonomy of punctuation functions, Dale not only focuses on the role of punctuation units in marking chunks in discourse, as Chafe (1988), Hannay and Kroon (2005) and Steen (2005) have also done, but also distinguishes between

the specific role that different punctuation marks can play in indicating certain rhetorical relations.

This overview of theoretical approaches to the analysis of written discourse has focused on those theories that have linked the occurrence and placement of punctuation units in text to representing the boundaries between the basic units of written discourse – the punctuation units. Even though a number of approaches have identified the punctuation unit as the basic unit of discourse, they differ from each other in the extent to which they provide a clear definition of punctuation units and identify different types of unit. Hannay and Kroon (2005) provide the most detailed account in this respect, in that they not only distinguish between punctuation units with a nuclear or satellite function, but also between prepended, appended and interpolated punctuation units. An approach like Dale's (1991) has also related the use of a particular punctuation mark to the rhetorical relation it expresses. The insights gained from these approaches will be used as a basis for the development of the discourse segmentation system designed for this study, to be described in more detail in sections 2.4 and 2.5 below.

### **2.3.2 Discourse segmentation: from data to theory**

In contrast to Section 2.3.1, the present section will present a selection of approaches to discourse segmentation that concentrate on the application of a theory of discourse segmentation to the actual analysis of data. What these approaches have in common is that they all acknowledge that in order to achieve consistent analysis, it is not only necessary to focus on providing a clear definition of what constitutes a basic unit of analysis, but also on presenting clear segmentation criteria. They can be distinguished from the theoretical-oriented approaches in that they pay explicit attention to problem cases in discourse segmentation and the solutions put forward in dealing with such problems.

This section will first describe three applied approaches to the analysis of spoken data, followed by a description of two applied approaches to the analysis of written data. Because the present study can also be classified as an applied approach to the study of written data, the description of the approaches presented below will contain a fair amount of detail and will focus on the problems encountered and the solutions put forward in order to see whether these can be useful for the analysis of the texts contained in this study.

### **Applied approaches to the analysis of spoken data**

In their approach to the segmentation of spoken data, Foster, Tonkyn and Wigglesworth (2000) explain that it is their main aim to not only develop a clear definition of a unit of analysis, but also to present the criteria on the basis of which such a unit can be identified and applied consistently in the analysis of wide range of spoken data. In their analysis of the many definitions that have been put forward by others, they remark that the main problems with these units is that they either look similar, but are defined in different ways; that they are not defined at all or that they are defined in a way that is too simple to be used with real spoken data (2000: 357). Moreover, they criticise other approaches for not providing clear insight into how a definition of a unit has been applied to the analysis of spoken data, as this would not only give an honest account of how difficult it is to segment spoken data consistently, but the shared experience could also be of benefit to other researchers who have a similar task (2000: 365).

This is why they introduce a new unit of segmentation, the Analysis of Speech Unit (AS-unit), which they present as ‘an accessible standard unit of analysis, explicit and exemplified, which is psycholinguistically valid, and which can be applied reliably to a wide range of oral data’ (2000: 365). The AS-unit is mainly a syntactic unit, but uses intonation and pause phenomena to deal with awkward cases (2000: 365-366). It is based on the T-unit, as described by Hunt (1966, see 2.3.1 above), which is defined in different ways, but basically constitutes a unit that has the form of a clause with all subordinate and embedded clauses attached to it. Even though Foster et al. take this unit as the starting point, they elaborate this to deal with the features of spoken data (2000: 365). The main segmentation criterion they put forward is that this unit constitutes a single speaker’s utterance that consists of an independent clause or a sub-clausal unit, where they define the former as ‘minimally a clause including a finite verb’ and the latter as consisting of ‘either one or more phrases which can be elaborated to a full clause by means of recovery of ellipted elements from the context of the discourse or situation or a minor utterance’ (2000: 365-366). As instances of these various linguistic forms, they provide the following examples (2000: 366):

- (21) That’s right
- (22) 1. How long you stay here  
2. Three months
- (23) Yes

Example (21) takes the form of an independent clause; in (22) both AS-units take the form of a sub-clausal unit, and (23) takes the form of a minor utterance. Moreover, because AS-units consist of an independent clause or sub-clausal unit, together with any subordinate clause associated with either (p. 365), example (24) below consists of one AS-unit. The subordinate clause is classified as a clause because it consists of a finite or non-finite verb plus at least one other clause element, such as subject, object, complement or adverbial (2000: 366).

- (24) I serves in a organization government organization in Bangladesh which is called er department of agricultural extension.

As for cases that often present problems for analysis, they identify units that contain various instances of coordination or subordination, and explain that these require clear specification. For coordination, they give the example of coordinated verb phrases and identify clear conditions for determining where unit boundaries lie. Their condition states that these will normally be considered to belong to the same AS-unit, unless the first unit is 'marked by falling or rising intonation and is followed by a pause of at least 0.5 seconds', as in the following example (2000: 366-367):

- (25) and they pinned er a notice to his front telling everybody what he had done (0.5) and marched him around the streets with a gun at his back.

Example (25) has thus been analysed as consisting of two AS-units, because the verb phrases *pinned* and *marched* are separated by a clearly noticeable pause. For subordinate clauses with an adverbial function, they base their segmentation criteria on the position of the adverbial clause in the unit. Initial and medial clauses are identified as usually not causing problems, whereas final adverbial clauses are. Consider the following examples (2000: 367-368):

- (26) When I was in the university er I have specialized in this er subject

- (27) I can understand when I read scientific English

Both (26) and (27) are analysed as each consisting of one AS-unit. The motivation given for (26) is that it is 'clear where [it] belong[s]', by which Foster et al. appear to mean that it is clear that the subordinate clause is a part of the same AS-unit (2000: 367). Example (27) is analysed as one AS-unit because the final subordinate

clause is part of the same tone unit as the clause that precedes it. Thus in the case of problematic final adverbial clauses, intonation is again used as a segmentation criterion.

In addition to these problem cases, Foster et al. also identify other well-known problem areas in the analysis of spoken data, such as false starts, repetitions and self-corrections and interruptions, and set clear segmentation criteria (2000: 368). These will not be described in more detail here, as they are characteristic of the type of spoken data that does not occur in the corpus compiled for this study and is therefore not relevant for the present study.

This account of Foster et al.'s AS-unit does not only clearly mark the contrast with theoretical approaches to discourse segmentation, but, according to them, also with other applied approaches to the analysis of spoken data, as they focus on presenting a unit of analysis that is accessible, clearly defined and can be easily identified on the basis of a clear set of criteria, without excluding problematic segmentation cases from the discussion. Such an approach not only functions as a useful resource for other researchers – which corresponds to their aim (2000: 371) - but also underlines the need for a clear definition and criteria in the segmentation of real language data, be it spoken or written.

Similar to Foster et al. (2000), Ford and Thompson (1996) and Ford, Fox and Thompson (1996, 2002) are also in search of a clear definition of a unit of analysis that can be applied to the analysis of real language data. In their search, they too point to shortcomings of other approaches that do not clearly define these units and do not apply the definitions to test whether they work (Ford & Thompson 1996: 140; Ford et al. 1996: 434). They also draw attention to the need to work with real language data, as working with 'idealized data' has not 'yielded theories that can be applied to the situated language use' (Ford and Thompson 1996: 136).

As a starting point for a unit of analysis, Ford and Thompson (1996) and Ford et al. (1996) take Sacks, Schegloff and Jefferson's (1974) notion of the Turn Constructional Unit (TCU). However, whereas the TCU is conceived as primarily a syntactic unit, Ford and Thompson (1996) and Fox et al. (1996) argue that intonational, pragmatic and even non-verbal communication features also play an important role in determining unit boundaries. Similar to Foster et al.'s account, they too provide a clear definition and explanation of their unit of analysis and set clear criteria that can be applied to determine whether a unit is syntactically, intonationally or pragmatically complete. As for syntactic completeness, it is explained that an utterance or unit is seen as syntactically complete if 'in its



discourse context, it could be interpreted as a complete clause' (Ford & Thompson 1996: 143). Ford and Thompson also explicitly state that they include elliptical clauses, answers to questions and backchannel responses in this category (1996: 143). Consider the following example, which represents one utterance with a series of syntactic completion points, indicated by forward slashes (1996: 144):

- (28) And his knee was being worn / - okay / wait./  
It was bent / that way

Syntactic completion is evaluated 'incrementally', which means that syntactic completion is 'calculated in terms of its relation with a previous predicate if one is available' (1996: 145). This means, for example, that *that way* in (28) is not claimed to constitute an independent unit by itself, but as 'being a second possible syntactic completion point'. Another type of completion, intonational completion, is defined in terms of constituting a prosodic unit or intonation unit, segmented on similar grounds as, for example, Chafe's intonation unit (Chafe 1994, see 2.3.1). In the following example, the syntactic completion points are again indicated by forward slashes and the intonation unit boundaries by a period (1996: 147):

- (29) Okay / this is what t-the problem is/.

On the one hand, example (29) shows that syntactic completion does not have to be accompanied by intonational completion, as *okay* is identified as a syntactic unit, but not as an intonation unit. On the other hand, it also shows that syntactic completion and intonational completion can converge, as it does at the end of the utterance. For the analysis of such examples in terms of units, syntactic criteria are not sufficient. The third type of completion, pragmatic completion, is the least clearly defined one, and remains 'intuitive and provisional' (1996: 150). However, as it is considered to play an important role for speakers in determining unit boundaries, it would be a 'mistake' to ignore it in the analysis (1996: 150). Pragmatic completion is to be identified in terms of final intonation contour and as a complete conversational action. Consider the following example, which contains instances of all three types of completion, where pragmatic completion is indicated by the 'greater than' sign (1996: 151):

- (30) And he said we'll probably have to put an artificial knee in / in five years/.

Example (30) thus constitutes an instance of a basic unit, as all three completion points converge at the end of the utterance.

In addition to providing clear segmentation criteria, Ford and Thompson have applied these to the quantitative analysis of data and found a high degree of coincidence of the three types of completion (1996: 153). For the segmentation of spoken data into units, it is Ford and Thompson's main point that syntactic, intonational and pragmatic phenomena cluster or converge at unit boundaries and thus that syntactic criteria alone are not sufficient to determine unit boundaries (1996: 171). Moreover, in a similar study in which their main question is also to find an answer to what constitutes the basic unit of analysis in speech, Ford, Fox and Thompson (1996) emphasise that their main concern should in fact not be to find an answer to this question, as this answer only gives a partial account of 'what is actually going on in the interactions we are observing'. Without going into this in any more detail here, what Ford, Fox and Thompson want to underline is that an analysis of conversation needs to acknowledge that many different things are going on at the same time and that all aspects have to be taken into account, including non-verbal behaviour, for an analysis to represent what is going on in actual conversation. By providing insight into the applications of their definitions, both Fox and Thompson (1996) and Ford et al. (1996) are not part of the group of researchers that draw 'a veil of silence' over their methods, to use Foster et al.'s words (2000: 357), which makes their study valuable to others that are struggling with similar questions.

The last approach to be discussed in the current selection is the one put forward by Degand and Simon (2005, see Degand & Simon 2009 for more recent contribution to this discussion). Similar to the other applied approaches discussed above, they also start their discussion by pointing out the lack of consensus in the literature on what constitutes a basic unit of discourse and stress the need for clear segmentation criteria by listing four requirements that a definition of a basic discourse unit should meet (2005: 67). The first of these states that only observable linguistic criteria can be taken into account in determining unit boundaries, as opposed to criteria that are not clearly observable, such as conceptual ideas. By observable linguistic criteria they mean syntax on the one hand and either intonation or punctuation on the other hand. The second requirement is twofold and states that it cannot be assumed that coherence relations exist between all basic units and that not all basic units 'contribute equally to the (conceptual) discourse structure' (2005: 67). The motivation for this requirement lies, on the one hand, in the fact that it is not clear-cut what may be admitted as a coherence

relation and, on the other hand, in that some units may not have the function of contributing to the conceptual discourse structure, but ‘play a role on the level of textual or interactional management’ (2005: 66). The third requirement states that a definition should take into account that ‘(spoken) discourse is an incremental process, but the resulting structure is static’ (2005: 67). Finally, the fourth and last requirement states that discourse segmentation implies that a completed product is being analysed, meaning that the perspective that is taken is one of discourse comprehension and not discourse production (2005: 67). In applying these criteria to develop a definition of a basic unit, Degand and Simon follow Hannay and Kroon’s (2005, see Section 2.3.1 above) basic distinction between conceptual units and strategic units, where strategic units are seen as constituting the linguistic realisations of conceptual units. Also in line with Hannay and Kroon (2005), who borrowed the basic idea from Chafe (1994), Degand and Simon distinguish between three types of strategic unit: fragmentary, substantive and regulatory. Even though they explicitly present the criteria that a sound definition of a basic discourse unit needs to fulfil, they do not provide a detailed definition of their own unit of analysis, the minimal discourse unit (MDU), but borrow Selting’s (2000) view on discourse segmentation and consider them ‘the smallest interactionally relevant complete linguistic unit[s], in a given context, that is constructed with syntactic and prosodic resources within their semantic, pragmatic, activity-type-specific, and sequential (...) context’ (Selting 2000: 477, as cited in Degand & Simon 2005: 65).

In addition to presenting these basic requirements, Degand and Simon also underline the need for explicit segmentation criteria, which should not only help the analyst in segmenting discourse into units, but also in identifying the three different types of unit (2005: 68). Discourse segmentation is based on both syntactic and intonational criteria. Without going into the details of the exact syntactic criteria they apply – as this is based on a specific grammatical theory of dependency grammar – Degand and Simon explain that the segmentations at both levels need to be compared in order to establish where unit boundaries lie. As intonation and syntactic boundaries do not always overlap, decisions have to be made in cases such as the following (2005: 70):

- (31)    mais elle disait que  
           (but she said that)
- tu es quelqu’un de très chouette  
           (you are someone very nice)

This example consists of two intonation units and because the first one ‘projects a syntactic continuation’, the two intonation units are together taken as constituting one MDU, as syntactic criteria override intonational ones in examples like these (*ibid*). The notion of syntactic projection is thus presented as an important segmentation criterion. This means that an example like (32) below is analysed as consisting of two units, not because the second clause, the adverbial clause of reason, constitutes a separate intonation unit, but mainly because this adverbial clause has not been projected previously and because the discourse marker *quoi* ‘leaves place to the hearer to react if she wants to’ (2005: 70):

- (32)      ben ca m’a fait un choc quoi  
               (well it was a shock you know)
- parce que c’est mon plus jeune frère quoi  
               (because it is my youngest brother you know)

Degand and Simon’s approach to discourse segmentation is included in the current selection because they have applied their approach to the analysis of a piece of spontaneous conversation. This analysis showed that in 84% of the cases the syntactic units and intonation units overlap; in 8% of the cases the syntactic unit is broken down into two or more intonation units, and in 7% of the cases one or more syntactic units is condensed into one intonation unit. As for the distinction between different types of unit, Degand and Simon hypothesised that regulatory units would overlap with syntactic elements that in their theory of grammar are classified as adjuncts, such as connectives or conversational markers, or combinations of these. However, they found that such adjuncts do not behave in a uniform way in actual discourse, as speakers can ‘package’ them in various ways, as the following examples show (2005: 72):

- (33)      franchement je vois ma grand-mère elle  
               (frunkly I see my grandmother she)
- (34)      mais il s’est quand meme vachement calmé tu sais  
               (but he’s really calmed down you know)
- franchement euh  
               (frunkly uhm)

In (33) the adjunct *franchement* is part of the one intonation unit and in (34) it is not, but instead constitutes a separate intonation unit. The question arises whether this adjunct should in both examples be analysed as a regulatory unit merely because of its syntactic classification, or whether it should only be analysed as regulatory unit if it constitutes a separate intonation unit. Degand and Simon do not answer this question, but merely raise it and conclude that further analyses are necessary to answer such questions (2005: 72). This example thus clearly shows that in applying segmentation criteria at both the level of intonation and syntax, in disputable cases like these a decision has to be made about which of the two criteria has precedence.

This section has presented three applied approaches to spoken discourse segmentation. What clearly distinguishes these from the theoretical approaches to the analysis of spoken data is their explicit emphasis on presenting sound definitions of units and clear segmentation criteria. The discussion has shown that they all present syntactic criteria as the main segmentation criteria, either supported by intonational criteria (Foster et al. 2000), or used in tandem with intonational criteria (Degand & Simon 2005), or in addition to both intonational and pragmatic criteria (Ford & Thompson 1996). Degand and Simon (2005) are the only ones to identify different types of discourse unit. Moreover, all approaches also emphasise the need for theories to be applied to data, as actual analysis of data will lay bare the problem areas. In line with this requirement, they identify problem areas, discuss these and present and motivate their particular analysis of such cases as practical decisions they have taken in the segmentation process. This section has thus shown that the analysis of actual data sometimes calls for practical decisions. Despite the fact that they are theoretically motivated, it is these practical decisions in particular that distinguishes these approaches from the theoretical ones.

### **Applied approaches to the analysis of written data**

A good example of an applied approach to the discourse segmentation of texts is the one put forward by Carlson, Marcu and Okurowski (2003, but see also Carlson & Marcu 2001), who have applied Mann and Thompson's Rhetorical Structure Theory (1988, see Section 2.3.1 above) to the segmentation and annotation of a large corpus of English newspaper texts. In building one of the first large corpora with discourse level annotation, Carlson et al. (2003) explain that the first step in

characterizing the discourse structure of a text is to segment it into basic units, which means that it first has to be determined what constitutes a basic unit of discourse (Carlson et al 2003: 86). They label these Elementary Discourse Units (EDUs) and describe them as the 'minimal building blocks of a discourse tree', which is based on the idea that RST can be represented as a tree of which the nodes correspond to text spans and the leaves to text fragments that represent the EDUs (2003: 86). In texts, EDUs take the form of clauses, and both lexical and syntactic clues are used to help determine unit boundaries (2003: 87). They do not, however, present this segmentation process as a straightforward matter, but instead describe it as 'extremely difficult' (2003: 86) and explain that 'any decision on how to bracket elementary discourse units necessarily involves some compromises' (2003: 87). They state their main goal as finding 'a balance between granularity of tagging and ability to identify units consistently on a large scale' (2003: 87).

As their approach constitutes an applied approach to discourse segmentation, they do not merely provide a definition of an EDU, but also provide insight into how this can be applied to the analysis of data and present specific segmentation criteria. For instance, although EDUs typically correspond to clauses, in line with Mann and Thompson (1988), they do not consider clauses that function as subjects, objects or complements of a main verb to constitute separate EDUs. This means that neither the subject nor the object in (35), both realised as non-finite clauses, are considered to constitute separate EDUs. Instead, (35) as a whole is seen as constituting one EDU (2003: 87):

(35) Making computers smaller often means sacrificing memory.

Another example of an identification criterion or further specification of their definition concerns the situation in which a clause, or an EDU, is broken up by, for instance, a relative clause, a nominal postmodifier or another type of clause. These interrupting clauses are assigned the status of embedded discourse units, an example of which is provided in (36) (2003: 87):

(36) The Bush Administration, trying to blunt growing demands from Western Europe for a relaxation of controls on exports to the Soviet Bloc, is questioning ...

Example (36) is thus analysed as consisting of two EDUs, in which the non-finite clause that interrupts the main sentence is classified as an embedded EDU. Last, as

a further specification of what exactly constitutes an EDU and how this can be identified, Carlson et al. explain that even though EDUs correspond with clauses, a small number of phrasal EDUs are allowed, ‘provided that the phrase begins with a strong discourse marker, such as *because of*, *in spite of*, *as a result of*, *according to* (2003: 88), as in the following example (taken from Carlson & Marcu 2001: 27):

- (37) But some big brokerage firms said they don’t expect major problems as a result of margin calls.

In (37) the phrase introduced by *as a result of* is thus analysed as a separate EDU. In order to guarantee consistent annotation of their corpus and provide insight into the annotation process, Carlson and Marcu (2001) made a list of discourse markers that do and a list of discourse markers and prepositions that do not introduce EDUs, both of which are provided in an annotation manual that they have made publicly available (Carlson & Marcu 2001). This manual not only gives insight into the problems or ambiguous cases they came across in the segmentation process, but also functions as a resource for other researchers, which makes them an exception to the group of researchers that draw ‘a veil of silence over their methods’, to quote Foster et al.’s words (2000: 357, see above). Although this section cannot provide a full account of the annotation process and the decisions that have been taken by Carlson and Marcu (2001), a few problematic cases are worth mentioning, as these have also been identified by others discussed in the current chapter as presenting segmentation difficulties. One such example concerns the segmentation of coordinated units. Carlson and Marcu state that coordinated sentences and clauses are broken into separate EDUs, but that coordinated verb phrases are not seen as constituting separate units. Consider the following examples, in which the square brackets indicate unit boundaries (2001: 11-14):

- (38) [Inventories are creeping up;] [car inventories are already high,] [and big auto makers are idling plants.]
- (39) [Equipped with cellular phones, laptop computers, calculators and a pack of blank checks,] [they parcel out money] [so that their clients can find temporary living quarters,] [buy food,] [replace lost clothing,] [repair broken water heaters,] [and replaster walls.]
- (40) [Under Superfund, those] [who owned, generated or transported hazardous waste] are liable for its cleanup,] [regardless of whether their actions were legal at the time.]

Example (38) is analysed as consisting of three separate EDUs, because each of the coordinates are realised as clauses. The analyses of (39) and (40) are interesting in the sense that they both contain lists, but in (39) the list is analysed as consisting of separate EDUs, whereas in (40) the list is not seen as consisting of separate EDUs. The motivation for this different approach is captured in the segmentation criteria, which state that the items on the list in (39) should be seen as separate EDUs because they constitute clauses, whereas the items on the list in (40) are to be analysed as verb phrases. A test that can be used to determine whether the coordination involves coordination of verb phrases or of clauses is to see whether the verbs share the same direct object or not. As this is the case in (40), in which the verb phrases *owned*, *generated* or *transported* share the same direct object, *hazardous waste* (2001: 14), this is analysed as constituting one EDU. In relating this analysis of coordinated verb phrases to the one presented by Foster et al. (2000) (see above), it becomes clear that Carlson and Marcu use syntactic criteria to determine unit status, whereas Foster et al. base their decision on intonational grounds. What both applied approaches have in common, however, is that they emphasise that problematic cases like these call for clear criteria if data are to be segmented consistently.

In addition to using lexical and syntactic criteria, similar to suggestions put forward by Chafe (1988), Hannay and Kroon (2005) and Steen (2005), Carlson and Marcu use punctuation as a means to identify unit boundaries (2001: 28). Although they provide a full account of what punctuation marks function as boundary markers in their manual, the present section will only present a few examples, to illustrate how they use punctuation as a segmentation device. For instance, when dashes are used to introduce parenthetical information, this information is classified as constituting an embedded EDU, as in the following example (2001: 29):

(41) [I sell] [- a little]

Another example is presented by the colon in (42) below, which is seen as separating two EDUs from each other (2001: 30):

(42) [Dollar:] [142.85 yen, up 0.95; 1.8415 marks, up 0.0075.]

As it is their aim to consistently apply Mann and Thompson's (1988) Rhetorical Structure Theory to the analysis of texts, in addition to segmenting text



into EDUs, Carlson et al. (2003) and Carlson and Marcu (2001) also annotate the rhetorical relations in texts, which not only involves identifying what rhetorical relation connects different units, but also making a distinction between nuclear and satellite units. Similar to the segmentation process, this part of the annotation procedure also proved difficult (2003: 103). In their annotation manual, Carlson and Marcu (2001: 13) provide a number of cases that presented difficulties both in terms of segmenting sentences into EDUs and determining their hierarchical status. Consider the following examples (2001: 31):

- (43) [The earnings were fine and above expectations... ] [Nevertheless, Salomon's stock fell \$1.125 yesterday...]
- (44) [Although the earnings were fine and above expectations,] [Salomon's stock fell \$1.125 yesterday.]

Even though the semantic content of (43) and (44) is very similar, the EDUs in (43) are both analysed as nuclei, whereas (44) is analysed as consisting of one satellite EDU that is followed by one nucleus EDU (2001: 31). In order to distinguish between nuclei and satellites, they apply two tests. The first is a deletion test, which states that 'when a satellite is deleted, the segment that is left, i.e. the nucleus, can still perform the same function in the text, although it may be somewhat weaker' (2001: 32). The second is a replacement test, which states that 'unlike the nucleus, a satellite can be replaced with different information without altering the function of the segment' (2001: 32). They add that embedded EDUs always have satellite status. The annotation manual thus provides a more detailed account of this part of the analysis. Even though the section on determining hierarchical status is rather limited, it explains that this occurs simultaneously with the assignment of a rhetorical relation, which they do discuss in great detail (2001: 31).

This account of Carlson and Marcu's (2001) and Carlson et al.'s (2003) approach shows that in segmenting written discourse into units and determining their hierarchical status, a great number of decisions have to be made, not all of which follow directly from the theoretical framework that has been adopted to annotate the data, RST in this case, but are to a certain extent also based on practical considerations, in order to achieve consistent annotation. Their annotation model and annotated corpus provide an impressive example of a case

in which discourse has been segmented into units in a consistent way and applied to annotate a large corpus at the level of discourse structure.

Another approach in which a unit of analysis is clearly defined and consistently applied to segment texts is presented by Hannay (1997). This approach is of particular interest to the present study, as it not only incorporates two languages in the analysis, but also the same languages as the present study, i.e. English and Dutch, and has therefore functioned as a basis for the discourse segmentation model developed for this study. In order to achieve his research goal, which is to conduct a contrastive analysis of sentence patterns in English and Dutch, Hannay defines his unit of analysis, the Punctuation Unit (PU), and provides clear segmentation criteria. He presents the Punctuation Unit as a unit that encodes an independent message and that can be seen as the written counterpart of the intonation unit, as introduced by Chafe (1994, see Section 2.3.1 above) (1997: 235).

As it constitutes an applied approach, Hannay clearly defines the segmentation criteria and describes how problematic cases have been dealt with. Similar to various other approaches, such as Foster et al. (2000), Carlson and Marcu (2001), although not explicitly stated, Hannay appears to consider the situation in which a PU is realised as a clause as the default situation, as he presents the cases in which it is either bigger or smaller than the clause as problematic, to which he then pays explicit attention. Again similar to the other approaches, these are cases in which one clause is embedded in another clause or, for instance, when one clause is coordinated with another clause. First, although he acknowledges that determining what exactly constitutes embedding may be problematic, he classifies a clause as being embedded when it does not constitute a message in its own right, but is part of a larger structure – a distinction he makes on the basis of using the context and punctuation (p. 236). This is in line with Mann and Thompson's (1988) distinction between the two types of clauses they identify, i.e. independent clauses, which have unit-status, and clauses that are part of their host clause, such as clausal subjects and complements and restrictive relative clauses, and which do not have unit-status (1988: 248), and is also similar to Foster et al.'s (2000, see above) decision to use an intonation break as a criterion to determine whether, for instance, a final adverbial clause should be seen as constituting a separate unit. As an example of Hannay's distinction between embedded clauses and dependent clauses, the latter of which thus receive discourse unit status, consider (45) and (46) (1997: 235-236):

- (45) The government at the time declared that environmental concerns had forced it to close down the mines. However, it soon became clear that the decision was taken *because the gas industry had bribed senior ministers*.
- (46) Over a hundred young deafblind people like Alan look forward to the other summer in the hope that Sense will be able to offer them a holiday. Their parents keep their fingers crossed too, *because [...] it's usually the only chance they get to relax knowing their children are in safe hands*.

Hannay explains that he classifies the adverbial clause introduced by *because* in the second sentence of (45) as an embedded clause because it does not constitute a message in its own right, but instead is part of another message. This means that this embedded clause plus the independent clause of which it is a part are together taken to constitute one information unit or Punctuation Unit. This is to be contrasted with the situation in (46), in which the adverbial clause introduced by *because* is analysed as constituting an independent message, both because it gives a reason for something that is asserted in the clause that precedes it and because it constitutes a separate PU (1997: 236). This analysis is reminiscent of Langacker's (2001) analysis of example (4) discussed above, in which the adverbial clause of condition and the independent clause that follows it constitute one attention unit when presented as one punctuation unit, and two attention units when presented as two punctuation units. Langacker explains that the speaker can choose to put the information into one or more attention units.

Second, another situation in which the PU may be greater than the clause is when two or more clauses are coordinated. Consider the following examples (1997: 236-237):

- (47) Please add your voice to that of Amnesty's, and together we really can make the world a better place.
- (48) He knows what he wants and gets what he needs.

Example (47) can be analysed as consisting of two clauses that also constitute two separate messages because they contain 'elements which make the two clauses more different', such as different subjects and the fact that they are both relatively lengthy. In cases where two clauses constitute two separate messages, they are more likely to be separated by a punctuation mark, such as the comma (cf. Quirk et al. 1985: 1615ff; Huddleston & Pullum 2002: 1740). In example (48), on the other hand, Hannay sees the two coordinated clauses as being integrated into one

message unit, as both clauses share the same subject (1997: 237). Hannay's analysis of coordinated clauses thus shows some overlap with Foster et al.'s analysis (2000: 233-367, but also see above), who make a similar distinction, but then on intonational grounds, and Carlson and Marcu's analysis (2001: 14, but also see above) who apply syntactic criteria to distinguish between, for example, coordinated verb phrases and clauses. Hannay uses both punctuation and syntactic criteria in analysing coordinated units.

In addition to cases where the punctuation units are greater than the clause, Hannay also addresses the situation in which the punctuation unit is smaller than the clause. In these cases it is not clear whether the punctuation unit should be seen as realizing independent messages (1997: 237). He identifies five problem areas. The first concerns the situation in which subjects or objects are separated from verbs by means of a punctuation mark, as in the following example (1997: 237):

- (49) En een van de eerste beslissingen die u op grond van die analyse moet nemen, is dan 'via welk kanaal communiceer ik?'

(And one of the first decisions that you on the basis of that analysis have to take, is then 'through which medium do I communicate?')

Hannay identifies this situation as being particular to Dutch punctuation practice, in which it is allowed to separate long subjects or long objects from the rest of the sentence (1997: 237). He decides not to interpret this use of punctuation as marking unit boundaries, as it cuts the grammatical structure in two. The example of this use of the comma shows that differences between the punctuation systems of two languages have to be taken into account if punctuation is used as a segmentation criterion to compare sentence structure or discourse structure of two languages.

The second situation Hannay identifies in which the punctuation unit is smaller than the clause is when a grammatical structure is interrupted, such as in the following example (1997: 237):

- (50) Although the story of Janet – a girl who looked about 12 but insisted she was 18 – does spring to mind.

This example is analysed as consisting of two PUs and not three, as the apposition that interrupts the adverbial clause introduced by *although* is seen as constituting

one message and the adverbial clause itself is another message (1997: 237). This analysis shows similarities with Carlson and Marcu's classification of embedded EDUs, who would also identify it as a separate unit both because it interrupts another unit and because it is surrounded by dashes (2001: 20, 29, but also see above). However, in addition to assigning it unit status, Carlson and Marcu also analyse the rhetorical relations between different units, with such units being considered satellites.

The third situation in which the punctuation unit is smaller than the clause is when the comma is used serially to separate items on a list, as in the following example (1997: 238):

- (51) The front line of our caring effort has been further strengthened with more branches, support groups and carers' contacts.

Hannay discounts these PUs from his analysis, as they constitute units below the level of the clause (1997: 238). However, it should be noted that this analysis is more straightforward when the list consists of NPs than, for instance, of VPs, in which case the coordinated units may also be seen as coordinated clauses instead of phrases. As discussed above, this situation has been addressed by Carlson and Marcu (2001: 14), who makes this distinction on the basis of syntactic criteria. As Hannay uses punctuation as a segmentation criterion, the question arises whether he also distinguishes between different types of punctuation marks and whether, for instance, a semi-colon used to separate a list of items would be treated similarly to commas separating a list of items. This question is, in fact, related to the discussion whether different punctuation marks should be considered as performing different discourse functions, which is something that will be discussed in more detail in Section 2.4.1 below.

The fourth situation in which the punctuation unit is smaller than the clause is when connectives, such as *however* and *to start with*, are presented as separate PUs. On the one hand, these are not integral parts of the clause they are associated with, but, on the other hand, it is difficult to see them as expressing separate messages that have 'propositional content' (1997: 238). Hannay explains that they can be seen as performing a function in the management of discourse, a concept he further develops later in the more theoretical-oriented study to discourse segmentation discussed in Section 2.3.1 above (Hannay & Kroon 2005). Moreover, Hannay points out that whereas in English these types of connectives are typically presented as PUs, in Dutch they are usually integrated into the

associated clause. This is why, for the purposes of his analysis, he discounts them from his study.

The fifth and final situation in which the punctuation unit is smaller than the clause concerns the case of adjuncts and disjuncts that occur as PUs. Consider the following examples, in which (52) contains a PU that takes the form of an adjunct and (53) a PU that takes the form of a disjunct (1997: 238):

- (52) The English language has expanded, *over the last 500 years*, to become the everyday speech of over 300 million people across the world.
- (53) Countries that have independent central banks also tend, *significantly*, to have low-inflation economies.

Hannay explains that when occurring in sentence-medial position, these adjuncts and disjuncts can be analysed as elliptical clauses. However, he notes that it is difficult to see them as such when they occur in sentence-initial position, 'since there is no antecedent', which is why he excludes them from his analysis. An additional reason for excluding them is that, again, in Dutch they are usually integrated into the associated clauses, whereas in English they are realised as separate PUs (1997: 239).

This account of Hannay's approach shows that even though he acknowledges that the definition of what exactly constitutes a punctuation unit still needs to be refined, he makes an attempt to provide a clear definition and the segmentation criteria, and also discusses the problematic cases. This illustrates that, similar to the other applied approaches discussed in this section, consistent segmentation of discourse, be it spoken or written, at times calls for practical decisions.

This section has described two applied approaches to the analysis of written discourse in quite some detail, focussing on the situations that present problems in the segmentation process. Similar to the applied approaches to spoken discourse and in contrast with the theoretical approaches, what characterises these approaches is that in addition to providing a theoretical basis of their segmentation criteria, they also acknowledge that consistent segmentation of actual data also calls for practical decisions at times. As for the types of segmentation criteria put forward, Carlson and Marcu (2001) explain that they combine lexical, syntactic and punctuation cues to identify unit boundaries, which shows overlap with Hannay

(1997), who also bases his decisions on both punctuation and syntactic criteria. What distinguishes these approaches is that Carlson and Marcu make explicit reference to the hierarchical structure of language and label units as either nuclei or satellites, which is something Hannay does not address until a later study (Hannay & Kroon 2005, described in Section 2.3.1 above). What makes Hannay's study of particular interest to the present study is that he applies his approach to discourse segmentation to both English and Dutch texts with the purpose of comparing their sentence structure.

### **2.3.3 Discourse segmentation: the need for a new definition**

Sections 2.3.1 and 2.3.2 have provided an overview of a wide variety of theoretical and applied approaches to the segmentation of both spoken and written discourse. The reason for presenting this overview relates to the main aim of this chapter which is to find a unit of analysis that can be applied consistently to texts in order to segment various types of English and Dutch sentences into discourse units to be able to perform a contrastive analysis of the sentencing patterns of these languages.

The search has not resulted in the description of one particular unit that exactly meets all the requirements that the unit that will be applied to the analysis of the sentences in the present study needs to fulfil. These requirements are as follows: the unit of analysis needs to be able to be applied consistently to the segmentation of text; it needs to be able to deal with various types of text, some of which contain features usually associated with spoken language; it needs to be able to deal with texts produced in two different languages, English and Dutch, and it needs to be able to capture the hierarchical structure of discourse. The search has, however, resulted in providing the foundations for the unit that will be defined and applied in the present study: the Sentence Information Unit (SIU).

The Sentence Information Unit is largely based on the Punctuation Unit, as presented by Hannay (1997) and Hannay and Kroon (2005), as it also uses punctuation as a criterion to identify unit boundaries. However, it deviates from Hannay's (1997) earlier definition in two respects, First, it includes certain PUs that he has excluded from his analysis and, second, it also takes into account how different units are related to each other hierarchically. Although the hierarchical analysis is added to the later version of the PU, presented by Hannay and Kroon (2005), this theory has not been applied to the analysis of actual data and is restricted to a description of the structure of English discourse. The hierarchical

analysis of SIUs has been based a combination of the ideas presented by Hannay and Kroon (2005), Carlson and Marcu (2001) and Chafe (1994). In addition to punctuational cues, both syntactic and semantic or contextual segmentation criteria will also be applied in identifying unit boundaries, which form a group of criteria that have been used to varying extents by most of the approaches presented in Sections 2.3.1 and 2.3.2 above.

The SIU can be defined as a unit of written discourse that expresses a single message, which can either be a dependent message or independent message and which can perform different functions in the discourse. An identification of unit boundaries occurs by applying both punctuational and syntactic criteria. The identification of SIUs occurs in two steps, the first of which involves segmenting the text into units and the second step assigning hierarchical status to these units.

As punctuation is used as a criterion in identifying unit boundaries, the following section will argue in detail why punctuation is seen as playing an important role in the organisation of written discourse and how this can be used as a reliable criterion in identifying unit boundaries. Explicit attention will be paid to differences in punctuation practice between English and Dutch and how these will be dealt with consistently. It will then present the segmentation criteria, focusing on the cases that present problems with respect to segmentation and the solutions or decisions put forward as to how these problems are dealt with in the present study to achieve consistent analysis.

After a description of the first step in the segmentation process, the identification of units, a full section (2.5) will be devoted to how the hierarchical status of a SIU can be determined, again focussing on the cases that present the main problems in this respect.

## **2.4 Sentence Information Units (SIUs): the role of punctuation**

As punctuation plays an important role in identifying SIU boundaries, this section will motivate the decision to assign punctuation marks this status and describe how they are used in identifying unit boundaries. The part of Section 2.3.1 above on written approaches to discourse segmentation described a number of approaches that included punctuation to varying extents in representing unit boundaries. Chafe (1988) and Hannay and Kroon (2005) label this unit the Punctuation Unit and argue



that punctuation practice can, at least to a certain extent, be interpreted as reflecting the prosody of the written language. Dale (1991) links the type and use of punctuation marks to the extent to which they reflect various rhetorical relations between different units. In contrast to these more theoretically-oriented approaches to the function of punctuation, Section 2.4.1 will look at how the occurrence of punctuation and punctuation practice is typically seen and described in the more practical approaches to this subject – grammars of English and Dutch. An important question that will be addressed is to what extent punctuation practice is seen as being guided by grammatical rules and to what extent the use of punctuation allows for personal input and variation. Related to this question is determining to what extent punctuation can be characterised by consistent application of punctuation marks on the one hand or by much individual variation on the other hand. This will be dealt with in Section 2.4.2, which will also describe how variation in the application of punctuation has been dealt with in this study. The third and final section in this part, 2.4.3, will give an account of how exactly punctuation is used in determining SIU boundaries, where explicit attention will be paid to differences in the punctuation systems of English and Dutch.

#### **2.4.1 Grammars of English and Dutch on punctuation**

Although there are various grammars of English and Dutch, this section will be limited to presenting the views of some of the main reference grammars of these languages on punctuation. It will not include a discussion or an overview of how punctuation is dealt with in the wide variety of style guides that are available of these two languages, as these can typically be characterised by taking a prescriptive perspective with respect to the use of punctuation.

With respect to Quirk et al.'s (1985) view on punctuation, they explain that punctuation marks serve two main functions in written discourse, namely separation and specification. Punctuation marks are seen as separating linear units from each other, such as words or sentences, or as marking off interpolated units, such as parenthetical clauses. With respect to the relation between punctuation and prosody, they state explicitly that punctuation is governed by grammatical considerations and that there is only little room for personal decisions (1985: 1611). However, at the same time they explain that there is room for 'a great deal of flexibility' and personal taste in the use of the comma, which they identify as the most frequently used punctuation mark together with the period (1985: 1613). Confusingly enough, this would mean that even though punctuation is seen as

being mainly governed by grammatical considerations, this does not necessarily apply to one of the most frequently used punctuation marks (cf. Chafe 1988 for a similar observation). They do, however, qualify their generalisation with respect to the relation between punctuation and prosody by stating that the conventions for punctuation are more strictly adhered to in printed material and less so in personal writing, which means that the latter text type might for that reason contain more inconsistencies in the application of punctuation that would not be admitted in most printed material (1985: 1611). As an example of an area in which the punctuation system allows for personal style or variation, they present coordination, noting that the writer can choose to insert or leave out commas, depending on his rhetorical purpose (p. 1618). Consider the following example in this respect (1985: 1618):

(54) I enjoy tennis but I don't play it often.

Quirk et al. explain that the two coordinated units can also be separated by a comma, depending on the writer's 'rhetorical reasons' (1985: 1618). Although they do not explain what exactly these reasons involve, they presumably mean the extent to which the writer wants to present the information in different units, thereby giving more emphasis to the individual units. This means that they see punctuation practice as being influenced, at least to a certain extent, by the rhetorical intentions or purposes of the writer. Specifically, they do not regard punctuation practice as being solely guided by rules, but also by 'decisions' writers make about how they present text (1985: 1624).

With respect to the differences between different punctuation marks, Quirk et al. present the punctuation system as a hierarchical system, in which the different marks are hierarchically related to each other. The hierarchy represents the extent to which elements are related or separated from each other, with the period indicating sharp separation, followed by the semicolon, the colon, the dash and finally the comma (1985: 1612).

The perspective taken by Quirk et al. (1985) on the relation between grammar and prosody is shared by Huddleston and Pullum (2002), as they also argue that punctuation does not reflect the prosody of spoken language (p. 1728). Instead, punctuation marks are described as giving 'indications of the grammatical structure and/or the meaning of stretches of written text' (2002: 1724). Their use and application is described as being guided to a large extent by 'codified rules' as set out in manuals or guides, although many of the rules for punctuation can also

be considered to be part of the ‘tacit linguistic knowledge’ of competent writers (2002: 1726-1727). As for the main functions of punctuation marks, they distinguish between the following four: indicating boundaries, indicating status, indicating omission and indicating linkage. They add that punctuation marks are often optional and that light and heavy styles of punctuation differ from each other in the extent to which they insert or leave out these optional marks (2002: 1730). In line with Quirk et al., they present a similar hierarchy of punctuation marks and illustrate how different punctuation marks provide insight into the hierarchical structure of the sentence (2002: 1731). Consider the following example, in which the semi-colon is explained to perform a different function than the commas in this sentence (2002: 1742):

(55) The Latin, for example, was not only clear; it was even beautiful.

Huddleston and Pullum explain that the semi-colon in (55) marks a different type of boundary than the internal commas in the first clause of this example. Specifically, it marks a boundary that is higher in the hierarchical structure of the sentence than the one that the commas mark (2002: 1743). The relation between the two clauses linked by the semi-colon is interpreted as one of coordination, which means that the clauses are at the same hierarchical level. Now consider (56) below, in which the relation between the two clauses is not interpreted as one of coordination, but as one in which the second clause elaborates on the first one. This means that the semi-colon now links two clauses that are not at the same hierarchical level (2002: 1742-1743):

(56) The bill was withdrawn; the sponsors felt there was no sufficient support to pass it this session.

Both examples (55) and (56) thus serve to illustrate that punctuation marks not only separate different constituents or punctuation units from each other, but also provide insight into how these units are related to each other hierarchically. This relates to Dale’s (1991) view on punctuation, who links the type of punctuation mark to the type of rhetorical relation it indicates between different units (see Section 2.3.1 above). In line with both Quirk et al. (1985) and Huddleston and Pullum (2002), the present study assumes a similar hierarchy of punctuation marks with different marks indicating different types of relations that can be used to interpret how various Sentence Information Units are related to each other.

With respect to grammars of Dutch on punctuation, it is rather remarkable to note that the most comprehensive grammar of Dutch, *de Algemene Nederlandse Spraakkunst* (1997), does not address the use and application of punctuation in any detail. A rather detailed account is, however, provided by Onrust, Verhagen and Doeve (1993) – technically not a grammar of Dutch – who devote a full chapter to presenting a descriptive account of the use and function of the four main punctuation marks, the comma, semi-colon, colon and period. Their view on the relation between punctuation and prosody can be described as one that does not deny a relationship between the two, but they emphasise that punctuation marks do not merely serve to indicate boundaries as pauses do in speech (1993: 181). Moreover, similar to Quirk et al. (1985) and Huddleston and Pullum (2002), Onrust et al. present a hierarchy of punctuation marks, with periods occurring at the top of the hierarchy and commas at the bottom. Punctuation is considered to play a role in the structuring of information within a paragraph, as it indicates hierarchical and rhetorical relations between the various units (p. 185, 192-193). Similar to Huddleston and Pullum’s account, the main punctuation marks, the comma, semi-colon and colon and period, are shown to perform different functions in the structuring of information within orthographic sentences. Commas, for example, are considered weak boundary markers that do provide clear insight into hierarchical structure and therefore need to be supported by lexical markers or other devices (p. 190). Consider the two versions of (57) below (Onrust et al. 1993: 189-190):

(57a) De minister geniet een zekere populariteit. Die is voornamelijk gebaseerd op zijn vroegere beleid.

(The minister enjoys quite some popularity, this is predominantly based on his former policy-making.)

(57b) De minister geniet een zekere populariteit, die is voornamelijk gebaseerd op zijn vroegere beleid.

(The minister enjoys quite some popularity. This is predominantly based on his former policy-making.)

The comma in (57a) is not strong enough to keep the two sentences together and should therefore either be supported by a coordinator that makes explicit what the relation between the sentences is, or the comma should be replaced by a 'stronger' punctuation mark, such as the semi-colon or the colon, as in (57b) (1993: 1990).

This brief description of a selection of sources on the grammar of punctuation shows that punctuation is seen as being predominantly guided by grammatical rules, but that there is room for personal variation, especially in the case of the comma. Moreover, punctuation marks are considered to play a role in marking the hierarchical structure of units, with different punctuation marks performing a different function and each having a unique position in the hierarchy of punctuation marks.

#### **2.4.2 Variation in the use and application of punctuation marks**

Since punctuation functions as an important criterion in identifying SIU boundaries, it is important to determine to what extent punctuation can be used reliably for these purposes. This leads to the question to what extent punctuation marks are used consistently by writers, as the grammars referred to above explain that there is room for variation, especially in the case of the most frequently used punctuation mark, the comma. This is related to the question how variation in the application of punctuation is to be interpreted. On the one hand, variation can be seen as reflecting a lack of skill on the part of writers to apply punctuation appropriately and consistently, which represents a commonly held belief about the use of punctuation. On the other hand, variation in the application of punctuation can be interpreted as reflecting particular choices and decisions writers make about how they present and package their information in a text (cf. Chafe 1988; Dale 1991; Hannay 1997; Carlson & Marcu 2001; Hannay & Kroon 2005, for similar observations, discussed in sections 2.3.1 and 2.3.2 above). The present study supports the latter interpretation on variation in the application of punctuation and interprets it as reflecting decisions about discourse structure on the part of the writer. However, this interpretation needs to be qualified in certain respects. The first qualification concerns the matter that even though punctuation is considered to reflect the intended discourse structure of the writer, being able to use punctuation effectively is, at least to a certain extent, considered an acquired skill (see also Chafe 1988: 399). However, this does not mean that punctuation is reduced to constituting a set of rules that can be acquired. Instead, it could be argued that we have 'tacit linguistic knowledge' of punctuation rules, as

Huddleston and Pullum suggest, which can be mastered or made explicit by competent writers (2002: 1727). Second, variation in punctuation can also be brought about by other factors than particular decisions on the part of the writer. Specifically, it can be induced by the punctuation system itself, as this allows for variation in particular cases; by the particular time in which a text is written or published; or by the particular text type or genre to which the text belongs. These three types of variation will be described in more detail in the following paragraphs.

Differences in personal style of punctuation are referred to as *light* versus *heavy* punctuation or *open* versus *closed* punctuation (Quirk et al. 1985: 1631; Huddleston & Pullum 2002: 1727; Chafe 1988: 415), which means that a writer can choose in certain situations to either insert a punctuation mark or leave it out. Quirk et al. (1985) identify the comma as the punctuation mark that allows for most variation, especially in the use of adverbials and coordinated units (Quirk et al. 1985: 1615-1619, 1626-1628). Similar to Quirk et al., Huddleston and Pullum also distinguish between light versus heavy punctuation and also identify adverbial phrases and clauses and coordinated units as the areas that allow for most variation in the use of punctuation (2002: 1739, 1746). These observations will be taken into account in determining which uses of punctuation can be interpreted as marking SIU boundaries, with special attention being paid to the use of punctuation in coordinated units and adverbials. Section 2.4.3 below will list the decisions that have been taken in this respect to guarantee consistent segmentation.

Another type of variation is brought about by the time period in which a text is produced or, to use Chafe's words, the punctuation style that is 'fashionable' at a particular time (1988: 416). In his view, the very fact that there is room for variation in especially the use of commas 'opens the doors to fashion' (1988: 416). Other studies have also related the use of punctuation to the time period in which a text is produced and have identified a general change in punctuation that occurred both in English and in Dutch. This involved a change in the function of punctuation, which used to be closer to representing the prosody of speech in writing and has become closer to representing the grammatical structure and prosody of the written language (Chafe 1988; Verhagen 1991; Onrust et al. 1993; and see Jones 1996 for an account of the history of punctuation). The present study will guard against this type of variation by only incorporating texts that are published in the same time period.

The last type of variation identified here is related to variation between text types, where a distinction is made between printed and published texts on the one hand and unpublished texts, such as manuscripts or personal documents, on

the other hand (Quirk et al. 1985: 1611; Huddleston & Pullum 2002: 1726). The main idea is that punctuation in published and printed texts is applied consistently and deliberately, also according to the punctuation trend current at the time of publishing. For this study, in order to guarantee that all punctuation has been applied consistently and deliberately, i.e. reflecting discourse intentions, only published texts will be included in the corpus.

### **2.4.3 Use of punctuation in determining SIU boundaries: borderline cases, exceptions and problems**

In order to use punctuation consistently as a criterion in identifying SIU boundaries, this section will present the guidelines that have been developed to take into account those situations in which punctuation marks do not indicate SIU boundaries, situations in which there are differences in punctuation practice between English and Dutch, and variation in the use of punctuation.

#### **Punctuation marks separating subjects, verbs and objects**

The first example of a situation in which a punctuation mark, typically a comma, is not considered to mark a SIU boundary is when it separates a subject, object or complement clause from the rest of the sentence. This is particularly common in Dutch (cf. Hannay 1997: 237) and much less so in English, which can be explained by the fact that in English it is considered ‘an unacceptable comma’ (Quirk et al. 1985: 1619) that may only be used in certain exceptional situations (Huddleston & Pullum 2002: 1744). Sentence (58) is an example of a Dutch sentence in which the subject is separated from the verb by means of a comma.

(58) Een bedrijf dat meer klanten wil trekken, moet juist de prijzen in toom houden <s2003, newspaper articles>.

(A business that wants to attract more customers, has to keep the prices under control.)

The motivation for not considering this use of punctuation marks to indicate SIU boundaries is twofold. On the one hand, it is motivated by practical considerations in that it concerns a difference in punctuation practice between English and Dutch. On the other hand, because this use of punctuation marks separates central clause

elements (cf. Quirk et al. 1985: 1619), it can be argued that it is difficult to see the resulting punctuation units as constituting independent messages and to determine what the relation is between the different units (cf. Hannay 1997: 237). It is, however, acknowledged that such uses of the comma might, and perhaps should, be allowed in certain circumstances, especially when it prevents misreading or confusion or better reflects the intended discourse structure (cf. Chafe 1988: 404; Hannay & Kroon 2005: 29; Huddleston & Pullum 2002: 1744).

All commas that separate long subjects, objects or complement clauses from the rest of the sentence have received the annotation labels <comma\_subject> or <comma\_NL>. A count of all instances showed that it is indeed far more common in Dutch sentences, as it occurs in 269 sentences in total (3.0%), whereas in English it only occurs in 26 cases (0.3%).

### **Coordination of main clauses**

A situation in which not punctuational, but syntactic criteria are used to determine SIU status and boundaries is when two or more main clauses that have the hierarchical status of nuclei (see 2.5 below on determining hierarchical status of SIU) are coordinated with each other. As determining clause status is not clear-cut, this will be described in more detail in Chapter 3, Section 3.3.1. This means that even though the discourse segmentation process is described and presented separately from the grammatical categorisation process (Chapter 3), in certain situations these interact, as syntactic criteria are used on various occasions in the segmentation process to determine and identify unit boundaries. With respect to the motivation for analysing the coordination of independent clauses in the way proposed here, this is because the syntactic criteria are in this particular situation considered to override punctuational criteria in the sense that independent clauses with nuclear status are considered to have SIU status in informational terms. Consider example (59):

- (59) <Ca>Families must take responsibility for their health<Ca> <Cb>but the Government and food manufacturers also have big role to play<Cb>. <s69, newspaper articles>

This examples consists of two independent clauses, each with a different subject, which are coordinated with each other by means of the coordinator *but*. Despite the fact that these independent clauses are not presented as separate punctuation



units, they are still considered separate SIUs in this study, indicated by the labels Ca and Cb respectively.

### **Serial use of punctuation marks: lists vs. multiple coordination**

Another situation in which the use of punctuation is not considered to indicate SIU boundaries concerns the serial use of the comma or semi-colon at or below the level of the phrase (cf. Chafe 1988; Hannay 1997; Hannay & Kroon 2005, for a similar criterion for identifying Punctuation Units). The following example contains a list of phrases that is separated by serial commas:

- (60) Basal cell cancers usually occur on areas of skin most exposed to the sun such as the head, neck, shoulders and limbs. <s8019, public information leaflets>

Serial use of punctuation can become problematic for discourse segmentation when a distinction has to be made between a list and an instance of multiple coordination (cf. Carslon & Marcu 2001 and Foster et al. 2000 for similar observations). In this study, the distinction between items on a list and coordinated items is made on syntactic grounds, similar to Carlson and Marcu's criteria (2001: 14, see Section 2.3.2 above). Items on a list are not considered to constitute separate SIUs, where a list consists of two items or more, whereas coordinated items are each considered to constitute a separate SIU. The main syntactic criterion used to distinguish between list items and coordinated units is that the item must be considered a clause in order to be qualified as a SIU, where clause status is determined on the basis of the presence of a finite or non-finite verb in addition to at least one other clause element (i.e. subject, object, complement or adverbial) (cf. Foster et al. 2000: 366, and Chapter 3.3.1 for a more detailed definition of a clause). The following examples illustrate how the distinction between lists and coordinated items is made in the present study, with square brackets ( [ ] ) marking unit boundaries:

- (61) [We suggest that very young children's conception of pictures encompasses all of recognizable images, abstract figures, patterns, writing, numbers and perhaps more]. <s5541, academic prose>
- (62) [The relatives want to know why the Red Caps had to give up weapons and ammunition], [why they did not have radios that would have allowed them to summon help] [and why they were not warned that they were going into a danger area]. <s13, newspaper articles>

- (63) [This may be an erroneous perceived generational difference], [but it could be a real effect]. <s5268, academic prose>

Example (61) contains a list of noun phrases that starts with *recognizable images* and ends with *perhaps more*. These NPs are not seen as constituting separate SIUs because this is taken to be a clear instance of the serial use of the comma that separates a list of phrases. Example (62), on the other hand, contains a list of embedded clauses. Precisely because these are classified as clauses, they are each considered to constitute a separate SIU. This example thus contains a series of coordinated SIUs. Finally, example (63) contains two coordinated clauses and is therefore analysed as consisting of two SIUs. However, because it could be argued that the SIUs in (62) are not identical in type or status to the SIUs in (63), the grammatical labels that are added after the segmentation process provide information about the syntactic differences between the coordinated units (see Chapter 3 on the labelling of the grammatical realisation of SIUs).

Thus, even though it is acknowledged that the distinction between lists and multiple coordination is not clear-cut, it has been treated as such by presenting distinct guidelines to distinguish between the two in order to guarantee consistent segmentation into SIUs. These guidelines are particularly relevant for the consistent segmentation of sentences in the texts that contain many instances of lists and coordination, such as the public information leaflets genre included in this corpus.<sup>6</sup>

### **Punctuation used to mark off phrases and embedded clauses**

Related to the problematic list-coordination distinction is the case in which phrases or embedded clauses are presented as separate punctuation units. The question is whether these punctuation units should be considered SIUs. With respect to phrases, the segmentation guideline is that if there are three or more phrases, each presented as a separate punctuation unit, these are considered to form a list and annotated as such. If, on the other hand, there are only two phrases that are juxtaposed and presented as separate punctuation units, these will be seen as constituting two separate SIUs. The following example illustrates this situation.

---

<sup>6</sup> The annotation manual included in Appendix I details all guidelines regarding the identification of Sentence Information Units in texts.

(64) In fact, [not one investigation], [but two]. <s47, newspaper articles>

Because the comma used to separate these phrases forms an example of an optional comma (cf. Quirk et al. 1985: 1618; Huddleston & Pullum 2002: 1740), its application is here seen as being placed deliberately and consciously by the writer in order to present these two punctuation units as two separate discourse units. This relates to the rhetorical function this use of the optional comma can have (cf. Quirk et al. 1985: 1618) and can also be linked to Foster et al.'s interpretation of deliberate intonational pauses between coordinated phrases as indicating unit boundaries (2000: 367, and see Section 2.3.2 above). Note, again, that the addition of a label that indicates the grammatical realisation of these SIUs will specify their grammatical status, which helps in distinguishing them from other types of SIUs. Consider also example (65):

(65) [Mr Hoon consistently denied there were problems with supplying our forces], [despite a huge amount of evidence to the contrary]. <s305, newspaper articles>

Similar to the motivation presented above, because the PP at the end of the sentence is presented as a separate punctuation unit, this is seen as representing a conscious decision on the part of the writer to present the information in two steps (cf. Hannay & Kroon 2005, see also Section 2.3.1 above). Section 2.5 will give more information about how these SIUs are related to each other hierarchically.

This same decision with respect to determining the SIU status of phrases has been applied to embedded clauses, which also concerns a situation in which punctuation marks can be used optionally. In this study, embedded clauses are only considered to have SIU status if they are presented by the writer as separate punctuation units. If they are integrated into the sentence in which they occur, and thus not presented as separate messages, they are also not considered to constitute separate SIUs. This means that example (66) is taken to consist of one SIU:

(66) [He has now attempted to smear the former weapons inspector by claiming he killed himself because he feared being exposed as a liar]. <s308, newspaper articles>

Example (67), on the other hand, presents a case in which the embedded clauses have been presented as separate punctuation units and are thus seen as constituting separate SIUs:

- (67) [Secretary of State Paul Murphy last night said the review would focus on the operation of the Good Friday Agreement], [and it would last two or three months]. <s846, newspaper articles>

Note furthermore that similar to embedded clauses, subordinate clauses are also only granted unit status if they are presented as separate punctuation units (with the exception of sentence-initial subordinate clauses, see section below on *punctuation and sentence-initial elements*). This means that if two or more subordinate clauses are coordinated, but are not presented as separate punctuation units, they are annotated as constituting one SIU. In these cases, to be able to capture the idea that such SIUs are syntactically somewhat more complex, the coordinator that links the subordinate clauses receives the label <coordinator\_sub>.

### Commas in complex embedded clauses

A particular use of punctuation that is also not considered to mark SIU status is the comma that is applied in complex embedded clauses to separate a modification that is placed at the beginning of the embedded clause from the rest of the clause. (68) illustrates this situation:

- (68) Experiment 1 showed that during presentation of lipread targets, articulatory suppression and irrelevant speech combined are no more disruptive than suppression above. <s5673, academic prose>

The comma placed between *targets* and *articulatory* separates the PP that is placed at the start of the embedded clause, introduced by *that*, and the main part of this clause, introduced by *suppression*. Even though this use of the comma is not considered to indicate SIU boundaries, it is acknowledged that it contributes to the readability of this sentence or to preventing a misreading that could, for example, result in seeing the two compound nouns as constituting one big compound NP (see Huddleston & Pullum 2002: 1730 for a description on this function of punctuation marks). However, Quirk et al. identify this use of the comma as being ‘in violation of an important punctuation rule’, also because it ‘obscures the

grammatical hierarchical relationship', as the PP should have been both preceded and followed by a comma. Yet they do classify it as 'acceptable' because it does not lead to misinterpretation or ambiguity (1985: 1632). In order to easily identify such cases, these commas have received a separate annotation label. This use of the comma is not particularly frequent in either language, but it occurs more frequently in English than in Dutch (39 cases vs. 9 cases), particularly in the academic genre.

### **Punctuation and sentence-initial elements**

An area that presents problematic cases in which certain practical segmentation decisions have to be made is with elements occurring before the subject of independent clauses. The reason why special guidelines have to be developed for the annotation of such units is mainly influenced by a difference in punctuation practice between English and Dutch with respect to these elements. Although in both languages the subject can be preceded by one or more phrases or clauses, in Dutch these elements are typically not presented as separate punctuation units, but instead integrated into to the clause to which they belong, depending on their length. In English, on the other hand, they are often presented as separate punctuation units, although this too depends on their function and length. Because there is a certain amount of variation between and within the languages in this respect, a number of guidelines have to be formulated in order to annotate sentence-initial elements consistently throughout the corpus.

The sentence-initial elements can be grouped into three main types. The first type concerns those cases in which the element or elements that precede the subject of the main clause are presented as separate punctuation units. All these cases are considered to have SIU status, in both languages. Consider the following examples:

- (69) [More importantly], Goodnow and Collins (1990) provide evidence that parents' ideas and expectations about development influence objective child outcomes. <s5237, academic prose>
- (70) [Integendeel], de PvdA ging zich in de loop der jaren steeds kritischer uitlaten over het Amerikaanse optreden in Vietnam. <s3790, academic prose>

(On the contrary, the Dutch liberal party became more and more critical over the years with respect to America's role/involvement in Vietnam.)

The second type of sentence-initial elements concerns those that take the form of a subordinate clause. All sentence-initial subordinate clauses are considered SIUs, irrespective of whether they are presented as separate punctuation units. This decision is thus based on syntactic grounds and not on punctuational grounds (cf. 2.4.3 above and Chapter 3.3.1 for a more detailed definition of a clause). The reason for not basing this decision solely on punctuational grounds is twofold. First of all, the use of commas to mark off sentence-initial subordinate clauses constitutes an area in which variation can be found within the languages, as this may be considered an optional comma. For instance, as Quirk et al. explain, the use of commas with sentence-initial elements is motivated not solely by length, but also by function, for instance whether the element constitutes an adjunct or a conjunct (1985: 1626-1628). The second reason for considering all sentence-initial subordinate clauses as SIUs is because the start of the sentence is an area that will receive special emphasis in this study (see Chapter 6 on the start of sentences). As all punctuation marks also receive an annotation label, it is always possible to identify and thus distinguish the subordinate clauses that are presented as separate punctuation units and those that are not. In the annotation system, both sentence-initial adverbial clauses in the (71) and (72) are thus considered SIUs:

(71) [As time went on], it became increasingly clear that nothing would be found. <s54, newspaper articles>

(72) [Om dat te bereiken] moet voldaan worden aan bepaalde voorwaarden. <s6049, academic prose>

([In order to achieve that] certain requirements have to be met.)

The third type of sentence-initial element concerns those cases in which the subject of the main clause is preceded by an element that is not presented as a separate punctuation unit. In both English and Dutch the length and function of the initial elements determine whether they are presented as separate punctuation units, but as these principles operate differently in both languages due to a difference in punctuation practice, integrated sentence-initial elements have

received a special annotation label, i.e. a <zz> label<sup>7</sup> (cf. Hannay 1997 for a similar observation, and Quirk et al. 1985: 1626 on guidelines for punctuation with sentence-initial adverbials). This zz-label distinguishes them from sentence-initial elements that are presented as separate punctuation units, which are considered SIUs, and makes them easily retrievable for analysis. Moreover, these integrated sentence-initial elements are only annotated if they precede an independent main clause and not if they precede a clause fragment. Consider the following examples:

(73) <zz>In just a few years<zz> it has become an essential part of life, providing access to a universe of information. <s80, newspaper articles>

(74) <zz>In de afgelopen jaren<zz> heeft de overheid fors geïnvesteerd in jeugdgevangenissen. <s2081, newspaper articles>

(<zz>Over the past few years<zz> the government has invested a considerable amount in juvenile detention centres.)

This section has presented a description and exemplification of how and when punctuation has been used as criterion in discourse segmentation. Moreover, it has presented cases in which segmentation decisions were not solely based in punctuational grounds, but also on syntactic grounds and it has shown how differences in punctuational practice between English and Dutch are dealt with.

## 2.5 Sentence Information Units (SIUs): determining hierarchical status

As described in Section 2.3.3 above, the identification of SIUs occurs in two steps. The first step involves identifying SIU boundaries and the second step involves determining the hierarchical status of the SIU. This section will describe how this hierarchical status can be determined, where the focus in the formulation of the guidelines is again on being able to apply these consistently in order to guarantee reliable annotation.

---

<sup>7</sup> The annotation manual included in Appendix I lists some minor exceptions to this guideline. Fronted objects have, for example, not been annotated, because these have not been included in the study of sentence-initial elements.

The notion that there are hierarchical differences between different units is in line with the view that discourse is structured hierarchically (cf. Mann & Thompson 1988). Various approaches described in sections 2.3.1 and 2.3.2 above have incorporated this notion in their view on discourse structure and discourse segmentation. For instance, Hannay and Kroon (2005), who have provided the main foundations for the SIU, distinguish between units that have nuclear status in the discourse and units that have a satellite status. These terms are introduced by Mann and Thompson in their description of Rhetorical Structure Theory (see 2.3.1 above), and have also been adopted by Carlson and Marcu (2001), who define the nucleus as representing ‘the more salient or essential piece of information’ and a satellite as providing ‘supporting or background information’ (p. 31). In addition to the distinction between nuclei and satellites, Hannay and Kroon also categorise the Punctuation Units that have a satellite status on the basis of the position they take with respect to the Punctuation Unit that has nuclear status, in that they can either precede it (preposed satellite), interrupt it (interpolated satellite) or follow it (appended satellite) (2005: 31).

In this study, each orthographic sentence is seen as containing a maximum of one nucleus, which can be realised as one nuclear SIU or two or more coordinated nuclear SIUs. Moreover, the nuclear SIU can, but does not have to be, accompanied by one or more satellite SIUs, which makes the latter’s presence optional. Similar to Hannay and Kroon’s (2005) classification of the position of satellites, these can precede, interrupt or follow the nucleus. This section will focus on how the hierarchical status of a SIU is determined and discuss the problems involved in this process.

### **2.5.1 Determining the nuclear status of a Sentence Information Unit**

The first and main step in identifying the hierarchical status of a punctuation unit is to determine which SIU functions as the nucleus in a sentence. If a sentence consists of only one SIU, this one always functions as the nucleus of that particular sentence. If a sentence consists of more than one SIU, the identification of the nuclear unit is based on the following criteria: the type of punctuation mark used (see Section 2.4.1 on the hierarchy of punctuation marks, and Dale 1991 on the function of punctuation in marking relations between units); the syntactic status of a SIU, and its semantic content. These three main criteria are not always applied simultaneously in the assignment of nuclearity: in certain cases and contexts one



may override the other. In each of the following three sentences a different criterion determines the nuclear status, with nuclear units being printed in bold:

- (75) **[Mr Blair's case is this]**: [Universities need more money]. <s215, newspaper articles>
- (76) [If taxes go down] **[there will be less money for doctors, nurses and teachers]**. <s290, newspaper articles>
- (77) (Can you get infected your first time? <s7851, leaflets>  
**[Yes]**, [if your partner has a STD and you have unsafe sex], [then you can become infected]. <s7852, leaflets>

Example (75) consists of two SIUs that are both realised as independent clauses. In determining which of these SIUs functions as the nucleus, the punctuation mark, the colon, is used as a criterion. The function of the colon in this example is to indicate that the SIU following the colon is an explication, explanation or demonstration of the SIU that precedes it, which is the typical function of the colon (cf. Quirk et al. 1985: 1620; Onrust et al. 1993: 194; Huddleston & Pullum 2002: 1743). Because the function of the SIU following the colon is to serve as an explanation of the SIU that precedes it, the first SIU is considered to be the nucleus in this sentence and the second SIU a satellite. Example (76) also consists of two SIUs, a segmentation decision that is made on syntactic grounds and not punctuational grounds (see 2.4.3 above). In this example the main criterion that has been applied is considering the syntactic status of the units. The first SIU is realised as an adverbial clause, which is syntactically subordinate to the SIU realised as a main clause that follows it. The SIU realised as main clause is here considered the nucleus and the SIU realised as adverbial clause the satellite. Finally, (77) presents yet a different situation, as the assignment of nuclear status is based on semantic criteria and contextual factors. This sentence starts with an affirmation, *yes*, which is an answer to a question that was posed in the preceding sentence. In this study, answers to questions can constitute nuclear information, depending on the context (see Ford and Thompson 1996 for a similar analysis, Section 2.3.2 above). Specifically, in some cases, responses have the function of a discourse marker and are not considered to constitute nuclear information and in other cases they do constitute nuclear information. In order to make this decision, previous discourse has to be taken into account. With respect to the analysis of example (77), this consists of three SIUs, the first of which constitutes the nucleus, followed

by two satellites. It should be noted that in most assignments of nuclearity these different criteria interact, as is the case in (75). In this sentence it is the combination of the use of the colon together with the semantic content of the clause that follows the colon that determines the hierarchical status of both units (cf. Chapter 7 for a more elaborate account of the relation between different types of punctuation mark, such as colons, semi-colons and dashes, and the hierarchical status of the units that they link).

The examples presented above showed cases of sentences with one nuclear SIU. Nuclei can, however, also be realised as two or more coordinated SIUs. The following sentences present examples of this situation:

- (78) [Ze was tegen en is tegen], [en heeft haar verstand op nul gezet]. <s1893, newspaper articles>

([She was and is against it], [and she just stopped thinking].)

- (79) [Haiti has few natural resources]; [its economy is mainly agricultural]. <s1229, newspaper articles>

Both (78) and (79) consist of two punctuation units that are coordinated with each other. In (78) the SIUs are coordinated syndetically, i.e. with an overt coordinator, and in (79) the SIUs are coordinated asyndetically, i.e. without an overt coordinator.

All nuclear SIUs have received the annotation label <C>, which stands for *core*, in this corpus. All coordinated nuclei have received the label <Ca>, <Cb>, <Cc>, and so on.

### 2.5.2 Problems in determining nuclearity

There are a number of cases that present some problems in determining the hierarchical status of a SIU. In dealing with these problems, the main challenge can be found in setting criteria that can be systematically applied to texts from four different genres. This section will present some of these problematic cases in determining hierarchical status.

### Answers to questions, discourse markers and vocatives

Even though nuclear units are prototypically realised as independent clauses, in cases where the written text contains simulated dialogue it depends very much on the context what constitutes the nuclear message (see Ford & Thompson 1996 and 2.3.2 above). For instance, whereas the academic prose genre contains hardly any instances of nuclei that are realised as non-independent clauses, in the short stories genre this applies to almost a quarter of all the nuclei (cf. Chapter 5 for exact numbers and percentages). As the syntactic status of the nucleus is thus at least to a certain degree genre-dependent, the extent to which syntactic criteria can be used to determine nuclear status is also genre-dependent. This means that some genres rely more on syntactic criteria to determine hierarchical status, whereas other genres also take semantic criteria or context into account. Consider the following examples, in which the sentences between brackets are included to provide some context:

- (80) (With the Reverend Ian Paisley's DUP now the largest party in unionism, he said Sinn Fein was exploring their position. <s515, newspaper articles>  
Yes, **the current stalemate is a crisis**, a dangerous crisis, he said. <s516, newspaper articles>
- (81) (Can you get infected your first time? <s7851, leaflets>  
**Yes**, if your partner has a STD and you have unsafe sex, then you can become infected. <s7852, leaflets>
- (82) **John**, he said. <s10590, short stories>
- (83) Annalee, **wait!** <s11400, short stories>

In the second sentence in (80), the affirmative *yes* does not serve as an answer to a question, but as an introduction to the information that follows. This is why it is considered a satellite that precedes the nuclear unit (see 2.5.3 below on satellites). In the second sentence in (81), on the other hand, the affirmative *yes* is an answer to the question posed in the preceding sentence. This is seen as constituting the nuclear information unit in this sentence. Example (82) represents a situation in which the main message is realised as a vocative, which is thus analysed as having nuclear status (see next section on reporting vs. reported clauses). Finally, (83) represents a situation in which the imperative *wait* is considered to constitute the main message in informational terms and the vocative that precedes it a satellite.

### Reporting clauses and reported clauses

Texts containing simulated dialogue or quoted speech usually contain both reported speech and reporting speech. In certain cases it is difficult to determine whether the reported clause should be considered the nuclear message or whether the reporting clause should receive this status. Consider the following example, in which the nuclear unit is printed in bold:

- (84) **U bent laat**, zei Timmer toen hij opstond en een reusachtige hand uitstak.  
<s13803, short stories>  
**(You're late**, said Timmer when he stood up and stuck out his gigantic hand.)

In order to analyse such cases consistently, an annotation guideline is formulated that states that reported speech is seen as constituting the main message in information hierarchical terms. This is in line with Biber et al.'s analysis of reported speech, who, in their explanation of the high frequency of reporting clauses in sentence-final position, state that 'most typically the quoted text is the main communicative point, and the reporting clause is tagged on at the end' (1999: 924). This means that in this corpus reported clauses typically receive nuclear status and reporting clauses typically receive satellite status, as is the case in (84) above. However, the following example forms an exception to this basic principle:

- (85) But Ms Abbott, sitting next to former Labour minister Keith Vaz, added in the minute and a half address: <s1032, newspaper articles>  
 'My involvement in the programme could hardly have been concealed given its nature'. <s1033, newspaper articles>

In cases where reporting clauses precede the reported speech, as in example (85), this is considered a less straightforward case, as it is difficult to determine whether what is being said should be seen as constituting the main message or whether the source should be seen as constituting the main message. Although it is acknowledged that this situation is not clear-cut, in order to annotate these cases consistently, both the reporting clause and the reported clause are analysed as having nuclear status, with the reporting clause receiving a special annotation label, <Cr>, to indicate that this concerns a nucleus that has the function of a reporting clause (r). With respect to (85) above, this means the first sentence, 1032, receives the label <Cr> and the second sentence, 1033, the regular label <C>.

### Coordinated phrases or embedded clauses

Section 2.4.3 above explained that coordinated phrases or embedded clauses that are presented as separate punctuation units are considered to constitute separate SIUs. With respect to hierarchical status, coordinated phrases or embedded clauses are seen as coordinated nuclei in this corpus. Consider example (86):

- (86) **[The new trafficking powers to be introduced today come on top of longer sentences for those who traffic in prostitutes], [and for purposes of domestic slavery and so-called ‘organ harvesting’].** <s1519, newspaper articles>

Example (86) consists of two punctuation units that are separated from each other by the comma in between *prostitutes* and *and*. The sentence as a whole is considered to be one complex clause that consists of two coordinated SIUs. This is why, at the level of discourse segmentation, these units receive the labels <Ca> and <Cb> respectively. However, the next step in the annotation process, which actually does not always neatly follow the discourse segmentation process, but at times also interacts with this process, concerns categorizing the SIUs grammatically (see Chapter 3). The grammatical labels would then indicate that the SIUs in (86) are realised as phrases and that it thus concerns coordination at the phrase level, instead of, for instance, at the clause level. This means that at a *discourse* level no distinction is made between the coordination of main clauses and coordination of phrases, as in both cases the discourse labels used are <Ca>-<Cb>. The *grammatical* realisation labels that will be added to the discourse labels do, however, specify what type of coordination is involved. The reason for using punctuation as a decisive criterion in the annotation of coordinated elements is twofold. On the one hand, the writer’s decision to present information in separate punctuation units is taken seriously and seen as reflecting his discourse intentions. On the other hand, the reason for treating the coordination of all types of elements similarly as long as they are presented as separate punctuation units also has a practical motivation, which is to achieve consistent annotation. Precisely because all SIUs receive both a discourse and a grammatical label, the exact status of these units in both levels of analysis will be captured in the current approach, which makes it possible to identify differences between different types of discourse unit.

### Correlatives

Another situation that is difficult with respect to determining hierarchical status is in cases where two units are introduced by correlative markers. Quirk et al. (1985: 935ff, 999ff) distinguish between correlative coordinators and correlative subordinators, with the former linking units of the same rank and the latter units of a different rank. Sentences (87) and (88) present examples of both types of correlatives, which are marked in italics (cf. Appendix I for a full overview of the different types).

(87) *The longer you take sleeping tablets, the more likely you are to become physically or psychologically dependent on them.* <s7386, leaflets>

(88) *Niet alleen bij hepatitis A komt geelzucht voor, ook bij andere leveraandoeningen kan dit optreden.* <s10383, leaflets>

(Hepatitis A does *not only* occur with jaundice, it can *also* occur with other liver diseases.)

In example (87) the two SIUs are linked by the correlative subordinators *the..the*. Quirk et al. argue that because the first clause introduced by *the* can be rephrased as a conditional clause, with *the* being replaced by *if*, it should be considered subordinate to the second clause (1985: 1000). Quirk et al.'s approach is adopted in the present study, which means that in the case of correlative subordinators, syntactic criteria, i.e. the actual correlatives, and semantic criteria, i.e. the ability to rephrase the unit as a subordinate clause, interact to determine the hierarchical status of units. With respect to (87), this means that the unit that can be rephrased as a subordinate clause is seen as having satellite status and the unit that follows as having nuclear status. Example (88), on the other hand, presents an example of correlative coordinators that are considered to link units of the same rank. This means that these clauses are taken to constitute two coordinated SIUs, <Ca>-<Cb>. It should be noted that only a very small number of punctuation units in the corpus are linked by correlatives, which makes this an exceptional phenomenon. This might mean that most correlatives occur within punctuation units.

### 2.5.3 Determining satellite status and type of satellite: prepended, appended and interpolated

In addition to identifying the nuclear SIU unit, satellite SIUs also have to be identified and categorised as prepended, interpolated or appended satellites. The identification and annotation guidelines for each type of satellite will be presented in more detail in the following sections.

#### Prepended satellites

In theory nuclei can be preceded by an infinite number of satellites. An analysis of sentence-initial elements in English and Dutch has, however, shown that nuclei in both languages are typically preceded by up to a maximum of three elements (cf. Smits 2002). If these elements are presented as separate punctuation units or if they constitute subordinate clauses, the first satellite to precede the nucleus receives the label <A>, the second <B>, and the third <-Z>. If, on the other hand, the sentence-initial elements are not presented as separate punctuation units and are realised syntactically as phrases, the first element receives the label <zz> and, in cases where there is a second one, <zz2> (see also 2.4.3 above on <zz>). A sentence can also contain both a <zz> unit and an <A> satellite, depending on whether or not it constitutes a punctuation unit and on whether it takes the form of a phrase or a clause. The following sentences present examples of all situations.

- (89) <A>As time went on<A>, it became increasingly clear that nothing would be found. <s54, newspaper articles>
- (90) <A>Immers<A>, <B>als uw keuze niet in het Donorregister staat<B>, moet uw familie na uw overlijden een beslissing nemen. <s8939, leaflets>
- (<A>After all<A>, <B>if your wishes are not registered in the organ donor register<B>, your family has to make a decision after your death.)
- (91) <A>Once exceptions are granted<A>, <zz>then<zz> everything is up for grabs, and trade and talks would be dragged down by interterminal bargaining. <s1267, newspaper articles>

In (89) the nucleus is preceded by a satellite that is presented as a separate punctuation unit and has thus received the label <A>. In example (90), the nucleus is preceded by two punctuation units, of which the first receives the label <A> and the second the label <B>. And in (91), the nucleus is both preceded by a

punctuation unit that receives the label <A> and another sentence-initial element *then* that is not presented as a separate punctuation unit and constitutes a phrase, and therefore receives the label <zz>.

### **Coordinated and subordinated prepended satellites**

If there are two or more satellites that precede the nucleus, instead of constituting two separate satellites that follow each other, they can also be in a relation in which they are coordinated with each other, or in which one is subordinate to the other. The following example presents a case in which two satellites are coordinated with each other, in which the coordination is identified on the basis of the coordinator *and* (cf. Chapter 3, Section 3.4.1 on coordination vs. subordination):

- (92) <Aa>If you snort drugs<Aa>, <Ab>and you use a note or a straw to snort through<Ab>, you shouldn't share it with anyone else, as blood can be passed from the inside of a person's nose to another. <s7848, leaflets>

If there are two satellites preceding the nucleus, these can also be of a different rank and be in a relation in which one is subordinate to the other. Consider in this respect the following example:

- (93) <A>Now<A>, <1>a year after Saddam fell<1>, Mr Straw says the lid has come off the pressure cooker. <s164, newspaper articles>

In (93), the adjunct *now* is followed by another adjunct, *a year after Saddam fell*, which provides a further specification of when exactly *now* was. Because it provides this further specification, the relation between these two units is seen as being different than, for instance, the relation between the sentence-initial units in (92). Specifically, in (93) the second adjunct is seen as being lower in hierarchical rank than the first adjunct, which is thus based on an analysis of the semantic relation between these units (cf. Smits 2002 for a similar analysis and Chapter 7 for an elaborate account of the possible relations between multiple sentence-initial units). The annotation labels used reflect this relationship between the two satellites, with the first adjunct receiving the label <A> and the second the label <1>, the latter of which indicates a more parenthetical and subordinate status (see below on *interpolated satellites*).



### Appended satellites

In theory, nuclei can also be followed by an infinite number of satellites. The first satellite to follow the nucleus receives the label <D>, the second receives the label <E>, and so on. Appended satellites can also be coordinated with each other. Consider the examples below:

- (94) There is no increase in time spent alone over the three later age groups, <D>suggesting that by age 7-8 some children are already solitary in the playground <coordinator\_sub>and<coordinator\_sub> that this may not increase in the primary school years<D>, <E>although other aspects of exclusion may do so<E>. <s5303, academic prose>
- (95) This error is duplicated in a recent paper by Henss (2000), <Da>who used photographs of women<Da>, <Db>and altered their apparent body mass<Db>. <s5627, academic prose>

In example (94) the nucleus is followed by two satellites, <D> and <E>. Note that the first satellite to follow the nucleus consists of two coordinated embedded clauses that are presented as one SIU. To be able to retrieve this type of coordination at a later stage, the coordinator receives a special annotation label, i.e. <coordinator\_sub>. Furthermore, in example (95), the two appended satellites are coordinated with each other and thus belong together, indicated by the labels <Da> and <Db> respectively.

### Interpolated satellites

In theory, both nuclei and satellites can be interrupted by an infinite number of interpolated satellites. Interpolated satellites, or interruptions, are always presented as separate punctuation units and surrounded by punctuation marks. The first interruption in a sentence receives the label <1>, the second interruption the label <2>, and so on. Moreover, additional labels indicate the position of the interruption with respect to the finite verb of the unit it interrupts, if this unit has a finite verb, as either preceding it or following it. The following sentences contain examples of interruptions:

- (96) Lord Ryder, <1\_prefinite verb>the acting chairman<1\_prefinite verb>, issued a cringing apology to the government, even though lawyers had told the governors that the Hutton report was legally flawed. <s183, newspaper articles>

- (97) Volgens LPF-Kamerlid Eerdmans zijn de door minister Donner ( <1\_postfinite verb>Justitie<1\_postfinite verb>) gedane voorstellen veel te beperkt. <s2400, newspaper articles>

(According to LPF MP Eerdmans the proposals put forward by Minister Donner (Justice) are far too limited.)

In (96) the nucleus is interrupted by a SIU that occurs in between the subject and the finite verb. This type of interruption can easily be identified by the fact that it is presented as a separate punctuation unit and as it occurs in between the subject and verb, it clearly interrupts the flow of the sentence. In (97) the interruption occurs after the finite verb *zijn* (to be), which is specified by the label *postfinite*. The interruption is clearly marked off from the rest of the sentence by the brackets that surround it.

Moreover, interruptions that occur in embedded clauses receive an additional specification that indicates that they interrupt an embedded clause. The labels prefinite or postfinite indicate where the interruptions occur with respect to the finite verb of the embedded clause. Consider the following examples:

- (98) Donner is zo tevreden dat hij binnenkort al, <1\_embcl\_prefv>terwijl de proef nog maar nauwelijks is begonnen<1\_embcl\_prefv>, de Tweede Kamer op de hoogte zal stellen van de eerste successen. <s2503, newspaper articles>

(Donner is so satisfied that he soon, even though the experiment has barely started, will notify the Dutch Lower Chamber of the initial successes.)

- (99) The unpalatable truth is that Haiti just does not matter very much, <1\_embcl\_postfv>strategically, economically or politically<1\_embcl\_postfv>, in the world as presently organised. <s1246, newspaper articles>

In both examples the interpolated SIUs interrupt an embedded clause introduced by *dat* and *that* respectively. In (98) the interruption precedes the finite verb of the embedded clause, whereas in (99) it follows the finite verb.

Similar to the other types of satellite, interpolated satellites can also be coordinated with each other, as in (100):

- (100) Paul Dacre, <1a>editor-in-chief of the Daily Mail group<1a>, <1b>and a man who guards his own privacy fiercely<1b>, made a rare appearance on a public platform last month when he crossed swords with a Commons select committee. <s1649, newspaper articles>

Although it is an instance of coordination at the phrase level when seen from a syntactic perspective, because these phrases are presented as separate punctuation units, they are annotated as coordinated interpolated satellites.

In addition, interpolated SIUs themselves can also be interrupted by SIUs, with the latter than being lower in hierarchy than the interpolated satellites they interrupt. These are only annotated if they are presented as separate punctuation units and receive the roman number <i>, <ii>, and so on, depending on whether it is the first, second or third interruption within an interruption. Similar to the ‘regular’ interpolated satellites, these interruptions of interruptions also receive the additional labels that indicate whether they precede or follow the finite verb of the interruption they interrupt. Consider the following examples:

- (101) Full digitisation - <1\_prefv>bringing better picture definition, interactivity and, <i>for many users<i>, internet access<1\_prefv> - is inevitable: the question is how fast we move towards it. <s1128, newspaper articles>
- (102) Vijftig procent minder politici, <1\_prefv>de ‘norm’ waarmee de nieuwe locoburgemeester van Amsterdam, <i\_prefv>Mark van der Horst<i\_prefv>, de bureaucratie te lijf gaat<1\_prefv>, kan rekenen op brede politieke steun. <s3445, newspaper articles>

(Fifty percent fewer politicians, <1\_prefv>the ‘norm’ with which the new acting mayor of Amsterdam, <i\_prefv>Mark van der Horst<i\_prefv>, fights bureaucracy<1\_prefv>, can count on full political support.)

In (101) the interruption that occurs before the finite verb of the main clause is itself interrupted by a prepositional phrase (PP), *for many users*. In (102) the interruption that occurs before the finite verb of the main clause is interrupted by an apposition, *Mark van der Horst*. The indication of position with respect to the finite verb of the SIU it interrupts is of course only provided if that particular SIU also contains a finite verb.

### Problems and borderline cases in identifying interruptions

There are cases in which it is difficult to distinguish between prepended and appended satellites on the one hand and interpolated satellites on the other hand. Consider the following example:

- (103) <zz>Yet<zz> [these notes reveal a great deal about Foxe's sources for his account of Elizabeth and his motives in writing it] [(which are by no means as straightforward as has usually been assumed)]. <s4202, academic prose>

Sentence (103) consists of two SIUs, the nucleus, and the appended satellite that is realised as a non-restrictive relative clause. Note that *yet* is here labelled <zz>, because it constitutes a phrasal element that precedes the subject of the nucleus, but has not been presented as a separate punctuation unit (see 2.4.3 above). The non-restrictive relative clause is analysed as an appended satellite and not an interpolated satellite, although it is recognised that the status of this SIU is ambiguous, as the punctuation marks used, the brackets, do specify the parenthetical status of the sentence final SIU. The motivation for classifying it as an appended satellite instead of an interpolated satellite is precisely because it follows the nucleus and does not interrupt it. In the analysis of such sentences, the type of punctuation mark used will be taken into consideration, as these are seen as forming a hierarchy with different marks indicating different degrees of separation (see 2.4.1 above and, for instance, Jones 1996: 129-130 for a similar observation and classification of brackets).

A similar ambiguous situation applies to cases in which the nucleus is preceded by several satellites. As described in the section on prepended satellites, these satellites can be of the same hierarchical status or one can be subordinated to the other. In a third situation, the satellite that precedes the nucleus can be interrupted by an interpolated unit. The following sentences represent an example of each of these three types<sup>8</sup>:

- (104) <A>As those who lived through it were clearly - <1>perhaps a little too clearly<1> - aware<A>, the story of the Lancashire cotton famine was highly creditable to all concerned. <s4776, academic prose>

---

<sup>8</sup> See Chapter 7 for a full account and detailed analysis of sentences starting with more than one sentence-initial element (cf. also Smits 2002 for a typology of complex adverbials).

- (105) <A>In veel landen<A>, <1>ook in de ons direct omringende landen<1>, komt hondsdolheid voor. <s10073, leaflets>  
(<A>In many countries<A>, <1>also in the countries that immediately surround us<1>, rabies occurs.)
- (106) <A>For example<A>, <B>in the context of research on early visual perspective-taking skills<B>, Flavell (1988) proposes a similar precursor which he describes as 'understanding a cognitive connection' (p. 248). <s5379, academic prose>

Example (105) presents the most straightforward situation, as the SIU that is labelled <1> clearly interrupts the sentence-initial adverbial clause that precedes the nucleus. The exact difference between examples (105) and (106), i.e. between the <1> and the <B>, may be less straightforward in this respect, also because this analysis is mainly based on semantic criteria. In (105), the second initial element, the <1>, is considered to provide a further specification of or addition to the information that is presented in the preceding A satellite. In (106), however, the second SIU, the <B> is not considered to modify or refer back to the <A> that precedes it in any way. Note the interruptions in (104) and (105) have both received the label <1>, but perform different functions in these sentences. The discourse patterns formed by these sentences also reflect this difference, as (104) follows an A1AC pattern, whereas (105) follows an A1C pattern (see Chapter 7 for detailed explanation of these patterns).

This section has described the second step in the identification process of the SIU, determining its hierarchical status. A distinction is made between units that have a nuclear status and units that have a satellite status, where the latter can either precede, interrupt or follow the nuclear unit. The discussion has focused on presenting the guidelines for how to determine hierarchical status, focussing on those cases that present difficulties in this respect. In addition to being motivated by theoretical considerations, a number of guidelines were also based on by practical consideration, mainly to achieve consistent annotation.

## 2.6 Conclusion

The main aim of this chapter was to find a unit of analysis that can be applied consistently to segment texts into units of discourse in order to carry out a sentencing analysis of English and Dutch. With this purpose, it has presented a wide variety of approaches to discourse segmentation, categorised, on the one hand, by either taking a theoretical approach or an applied approach to the subject and, on the other hand, by either concentrating on the analysis of spoken language or the analysis of written language. The focus in the description of these approaches was on the extent to which they provided a clear definition of a basic unit of analysis and the criteria put forward for the identification of it. What the overview has made clear is that the applied approaches to discourse tend to focus more on not only providing the theoretical foundations of their view of discourse organization, but also how these can be translated into a practical tool for the consistent analysis of data. A subset of these also provides insight into the problems encountered during the segmentation process, although it should be noted that these form the exception to the rule. The fact that the theoretical approaches do not explicitly address or anticipate the segmentation difficulties and the fact that many applied approaches do not report on the difficulties encountered could in fact be seen as a serious shortcoming, as it makes them less useful and applicable for other data analysts. However, the overview did provide insight into the types of segmentation criteria used, which proved valuable to the present study. The majority of the approaches use a combination of different criteria, with a frequent combination being syntactic and intonational or syntactic and punctuational criteria, dependent on the nature of texts under analysis.

On the basis of the overview, a new unit of analysis has been put forward for the analysis of the data in the present study, the Sentence Information Unit. As none of the units presented in the overview of approaches exactly met the criteria to be able to deal with the data of this study, the SIU has been based on a combination of various approaches. It has been predominantly based on the Punctuation Unit as presented by Chafe (1988), Hannay (1997) and Hannay and Kroon (2005), by considering punctuation as at least to a certain extent reflecting the writer's discourse intentions. In those cases in which punctuation does not mark unit boundaries, a combination of syntactic and semantic criteria has been used to identify unit boundaries. In addition to representing a unit in discourse, the SIU also provides information about the hierarchical status and position of this unit

with respect to its surrounding units by indicating whether it has nuclear or satellite status and, for the satellites, also indicating whether they precede, interrupt or follow the nucleus. In addition to presenting the foundations of the SIU, its exact definition and the criteria used to identify it, the chapter has also presented the main segmentation guidelines, in which special emphasis was placed on possible segmentation difficulties and solutions to deal with these. Although the vast majority of these segmentation guidelines have a theoretical foundation, some of them are also practically motivated, as it constitutes an important aim of the present study to achieve a consistent analysis of sentencng patterns.

The next chapter will describe how the grammatical realisation of the SIUs was annotated in the present study, with special attention again being paid to the situation in which annotation is not straightforward.





## 3. Grammatical categorisation

### 3.1 Introduction

Chapter 2 focused on what discourse segmentation involves and presented the unit of analysis, the Sentence Information Unit (SIU), which has been developed in order to carry out a sentencing analysis of English and Dutch. However, as sentencing concerns the packaging of information into linguistic units, segmenting sentences into basic units concerns just the first step in the analysis. The second step involves analysing how these units are realised syntactically. It is only when both steps are carried out that insight can be gained into the interplay between discourse segmentation and linguistic realisation, and thus in the sentencing process. Similar to the problems that were involved in discourse segmentation, determining the grammatical realisation of units, which entails grammatical categorisation, is also not unproblematic. Consider the following example:

- (1) <C>The change to China's constitution, <1>adding a clause guaranteeing private property rights<1>, may be an important step in protecting individual landowners<C>, <D>if enforced<D>. <s1150, newspaper articles>

Sentence (1) is analysed as consisting of three SIUs: a nuclear unit (<C>), which is interrupted by an interpolated satellite (<1>) and followed by an appended satellite (<D>). Determining the grammatical realisation of the nuclear unit and the appended satellite is fairly straightforward, with the former taking the form of an independent clause and the latter the form of a non-finite adverbial clause of condition. However, determining the grammatical realisation of the interpolated satellite may be subject to debate. On the one hand, solely taking into account the form of this clause, it can be categorised as a non-finite clause. On the other hand, when the function of this particular clause is taken into account, it can also be analysed as appositive postmodification (cf. Quirk et al. 1085: 1271). Just as the SIU had to be defined clearly in order to be applied consistently to the segmentation of discourse, grammatical categories also have to be defined clearly in order for an analysis of the grammatical realisation of SIUs to be carried out consistently. This means that in the case of the interpolated satellite in (1) it is necessary to determine whether the syntactic categorisation is determined on the basis of the form the clause takes or on the basis of the function it fulfils. As both ways of

categorising are possible, determining how this clause should be categorised concerns mainly a matter of decision-making.

Moreover, in the labelling of the grammatical realisation of SIUs, decisions not only have to be made about whether a form approach or a function approach is taken, but also about how grammatical categories are to be distinguished from each other when the boundaries between them are fuzzy. This poses a challenge in certain situations, as, to use Quirk et al.'s words '[g]rammar is to some extent an indeterminate system', in which '[c]ategories and structures ... often do not have neat boundaries' (1985: 90). An additional challenge specific to this study is how to consistently categorise certain linguistic forms that are more typically associated with the spoken mode in the simulated dialogue parts of the short stories genre, which is an area that not all grammar books deal with as extensively and consistently. In cases where boundaries are fuzzy and categories are not clearly defined, clear identification and distinction criteria have to be set to guarantee consistent categorisation.

It is the main focus of the present chapter to provide insight into the decision-making process of classifying SIUs with respect to their grammatical realisation, where special attention will be paid to the grammatical categories that have fuzzy boundaries. This chapter will first explain what labels have been used to classify SIUs in both English and Dutch (3.2). It will then present a number of general categorisation issues, such as the definition of a clause and the form-function debate (3.3.1 and 3.3.2). After these more general issues, it will focus on a number of issues that are particular to clause combining relations or, in this case, SIU combining relations, such as the coordination-subordination gradient (3.4.1) and the complex category of apposition (3.4.2). The chapter will conclude with a brief discussion of a number of genre-specific issues (3.5), where the focus will be on categorisation challenges caused by the simulated dialogue parts of the short stories genre. It should be noted that the focus throughout the chapter will be on the decisions that have been made in the present study in order to achieve consistent categorisation and thus achieve reliable sentencing analyses.

### 3.2 Quirk et al. as a basis for grammatical classification

Precisely because of the fact that certain grammatical categories can be approached and defined in various ways, variation can be found between the ways in which different grammar reference books define grammatical categories. In order to guarantee the consistent labelling of grammatical categories, the present study has adopted the grammatical categorisation as presented by Quirk et al. (1985). Even though this is a grammar of English, the categorisation and labels as used by Quirk et al. have also been used for the labelling of Dutch grammatical categories. In addition to the practical considerations for doing so, a more profound reason is presented by the fact that the most comprehensive grammar of Dutch, *Algemene Nederlandse Spraakkunst* (Haeseryn et al. 1997), is in a number of cases less comprehensive and detailed than Quirk et al. For instance, in adopting Quirk et al.'s grammatical categorisation to classify the prepended satellites (<A>) in the following examples, it is possible to not only describe their grammatical form, i.e. adverbs, but also their grammatical function, i.e. conjuncts (1985: 631):

- (2) <A>Similarly<A>, the graduates who command the really big salaries - city bankers, lawyers and accountants - could afford to pay back more after university than those in public services. <s701, newspaper articles>
- (3) <A>Kortom<A>, redenen genoeg voor alle betrokkenen (NS, overheid en consumentenorganisaties) om gezamenlijk te bezien of het treinkaartje wel extra duur gemaakt moet worden. <s2007, newspaper articles>
- <A>In short<A>, reasons enough for all those involved (NS, government and consumer organisations) to decide together whether the train ticket should be made more expensive).

In addition to classifying the sentence-initial adverbials on the basis of their grammatical form, Quirk et al. also distinguish between four broad categories of grammatical functions that adverbials can fulfil: adjunct, subjunct, disjunct and conjunct. The adverbials in (2) and (3) have the function of conjuncts, which are described as fulfilling the function of 'conjoining independent units' (1985: 631). As the present study not only focuses on the grammatical form of SIUs, but also their grammatical function in a number of cases, Quirk et al.'s subcategorisation of, for instance, the grammatical functions of adverbials proves particularly useful (see

3.3.2 below on form vs. function). This is to be contrasted with Haeseryn et al.'s approach, in which adverbials are only categorised on the basis of their form and not on the various functions they can fulfil in the sentence.

### **3.3 General categorisation issues**

In categorising the grammatical realisation of SIUs, there are a number of issues that can be described as concerning more general and profound categorisation issues that lie at the basis of, and thus precede, all other categorisation issues. For instance, as example (1) above illustrated, in order to categorise the interpolated satellite, one of the first decisions that has to be made is whether this should be classified as a clause or phrase. Another decision would then concern whether this particular satellite is to be categorised on the basis of its form or its function. It is these two decisions that affect all the other categorisation issues, such as what particular category this satellite would belong to, the category of adverbial clauses or of appositions, for instance.

As these two main issues precede other decisions that have to be made in the categorisation process, they will be described and exemplified in more detail in the following subsections. The discussion will start by pinpointing what exactly is problematic about determining clause status or deciding whether to classify elements according to their form or function, which will be followed by an exemplification of the categorisation decisions that have been made and a description of how these have been applied to the analysis of data.

#### **3.3.1 The cline of clausiness**

In categorizing the grammatical realisation of a Sentence Information Unit, one of the first decisions involves determining whether the SIU should be analysed at the level of the clause or at a level below the clause. Consider in this respect the following example, taken from Quirk et al. (1985: 912):

(4) I caught the train – just.

In the present study, example (4) would be analysed as consisting of two SIUs, a nuclear unit that is followed by an appended satellite (cf. 2.5.1 & 2.5.3).

Determining the grammatical realisation of these units and assigning them to a grammatical category involves determining whether they should be analysed at the level of the clause or at a level below the clause. However, determining what constitutes a clause proves not to be a straightforward matter and it depends on the definition adopted whether both units in (4) should be analysed as clauses. In Quirk et al.'s view, both units constitute clauses, although they characterise the former unit as constituting a more prototypical clause than the latter unit (1985: 911). This section will first explore how clause status is typically determined and will then describe how this has been applied to the analysis of data in this study.

Clause status is typically determined on the basis of the extent to which a clause contains its necessary clause elements, which range from central elements to peripheral elements (cf. Quirk et al. 1985; Haeseryn et al. 1997; Huddleston & Pullum 2002). One way of determining which elements should be considered central and which ones peripheral is by basing this on the complementation properties of the verb, where, for example, in the case of intransitive verbs the verb itself and the subject are characterised as central elements and any accompanying adjuncts as peripheral elements (cf. Quirk et al. 1985; Huddleston & Pullum 2002). This means that the classification as either a central or peripheral element is dependent on whether the verb is classified as an intransitive, monotransitive, ditransitive or complex transitive verb (Quirk et al. 1985: 54). Quirk et al. classify the distinction between central and peripheral elements as being 'relative rather than absolute' (1985: 50). They identify seven clause types, based on different complementation patterns, and explain that these types form 'the most general classification that can be usefully applied to the whole range of English clauses whether main or subordinate' (1985: 53).

Another way of determining clause status is by identifying a finite clause prototype, where a gradient between verb phrases is established to determine the extent to which the VP can be characterised as finite (Quirk et al. 1985: 149-50; Aarts 2007: 118). A prototypical clause can then be characterised by the fact that it 'carries tense, is independent, and complete' (Aarts 2007: 118). Less prototypical clauses, on the other hand, are clauses with a non-finite verb, which are always dependent clauses. This notion of prototype creates a cline of clauses that extends from most typical instances of clauses to least typical instances of clauses. In his study of syntactic gradience, Aarts presents the following continuum of clauses, with the most prototypical instances at one end and the least prototypical instances at the other end (2007: 120):

Main clause > ('complete' > 'incomplete') hypotactic finite clause >  
 ('complete' > 'incomplete') embedded finite clause > ('complete' >  
 'incomplete') hypotactic non-finite clause > ('complete' > 'incomplete')  
 embedded non-finite clause > ('complete' > 'incomplete') verbless clause

In establishing what constitutes a prototypical clause, Aarts also introduces the notion of completeness. He explains that particular grammatical functions can be left implicit in many cases and classifies clauses as less prototypical if one or more of these functions that are required by the transitivity properties of the verb are not expressed overtly (2007: 119). Consider in this respect the following examples, taken from Aarts (2007: 120):

- (5) We saw an elephant.  
 (6) While running, the elephant squirted water at us.

Aarts classifies (5) as a prototypical clause, as it constitutes a main clause that can stand on its own. The sentence-initial subordinate clause in (6), on the other hand, is classified as a less prototypical clause, because it contains a non-finite verb whose subject is not expressed overtly (2007: 120). In presenting this 'cline of clausiness', Aarts acknowledges that the exact make-up of the continuum depends to some extent on one's grammatical framework (2007: 121).

Related to Aarts' notion of completeness is Quirk et al.'s notion of ellipsis, which is also used in relation to determining clause status. Ellipsis is described with reference to the principle of 'verbatim recoverability', which means that 'the actual word(s) whose meaning is understood or implied must be recoverable' (1985: 884). Different degrees of ellipsis are identified, where the particular degree can be determined on the basis of a number of criteria. Quirk et al. also apply the notion of ellipsis to determine the clause status of the appended satellite in example (4) above, repeated here as (7) (1985: 912):

- (7) I caught the train – just.

They classify the appended satellite as an appended clause, which they describe as a special type of elliptical clause for which the whole or part of the preceding clause constitutes the antecedent. They explain that both clauses 'presuppose that two separate assertions are being made', which would look as follows, if written in full (1985: 912):

(8) I caught the train – I just caught the train.

Quirk et al. acknowledge that ‘the boundaries of ellipsis cannot be easily defined’ and classify it as a fuzzy concept. They therefore use the term quite generally for ‘grammatical reduction through omission’ (1985: 889, note p. 890). Even though criteria are provided to determine the degree of ellipsis, it is still classified as a fuzzy concept and it is the task of the present study to set clear identification criteria for clauses, taking into account the concept of ellipsis.

This section showed that there are different approaches to defining clause status. One of these is to analyse the extent to which a clause contains its necessary clause elements, based on the complementation pattern of the verb. Another approach involves establishing a clause prototype and defining the exact properties of this prototype. This creates a cline of clausiness (Aarts 2007: 121), in which clauses that exactly correspond to the prototype are placed at one end and clauses that deviate from the prototype more towards the other end. The following section will describe how clause status has been determined in the present study and how clauses have been distinguished from phrases.

### **Application of the cline of clausiness to the analysis of data**

While the present study acknowledges a cline of clausiness similar to the one presented by Aarts (2007), for the sake of the consistent analysis of data, this cline was segmented into clear-cut categories that range from independent clauses at one end to clause fragments at the other end. The following categories were distinguished: independent clause, subordinate clause, embedded clause and clause fragment. This section will present the criteria on the basis of which these categories can be consistently distinguished from each other and will describe how these criteria can be applied to the analysis of data. It will start by discussing the more straightforward cases, i.e. clauses that show minimal deviations from the prototype, and will continue with the more problematic cases.

The definition of an independent clause adopted in this study corresponds to Aarts’ view of the most prototypical clause: a main clause that can stand on its own (2007: 120). In order for a clause to be classified as an independent clause it has to satisfy a number of criteria. Similar to the distinction that Quirk et al. (1985: 49) make between central and peripheral clause elements to determine clause status, a clause is considered independent if it contains all the complements that are required by the transitivity properties of the main verb. Even though it is

acknowledged that a distinction between central and peripheral clause elements is not always straightforward (cf. Quirk et al. 1985: 50), a clause is considered independent if it contains at least those elements that are typically identified as the most central clause elements, namely a subject and a finite verb. In order to deal with situations in which clause elements are left unexpressed, and thus to make concrete the fuzzy notion of ellipsis, the general guideline that has been set in this study is to determine clause status on the basis of the overtly expressed clause elements. However, there are a number of exceptions to this rule, the most important of which will be exemplified below.<sup>9</sup> Consider the following examples:

- (9) <C>David Blunkett said that the day Beverley Hughes resigned was the worst of his life<C>. <s30, newspaper articles>
- (10) <C>Vertrouwt de overheid haar eigen betaalmiddel niet meer<C>? <s2019, newspaper articles>  
(<C>Does the government no longer trust its own currency<C>?)
- (11) <C>Begin ermee zodra hij baby-af is<C>. <s9025, leaflets>  
(<C>Start with it as soon as he is no longer a baby<C>.)
- (12) <Ca>Aides privately admitted a referendum was a 'big risk'<Ca> <Cb>but insisted Mr Blair did not regard defeat as a resignation issue<Cb>. <s884, newspaper articles>

First, example (9) consists of one nuclear SIU, <C>. The *that*-clause is not analysed as a separate SIU because it is embedded in the nuclear unit and because both the independent clause and the embedded clause are presented as one punctuation unit (cf. 2.4.3 for determining SIU status in the case of embedded clauses). As the second step in the analysis of this sentence is to label the grammatical realisation of the SIU, (9) as a whole needs to be provided with one grammatical label. In the present study, the nuclear SIU in (9) receives the grammatical label of independent clause, despite the fact that it could syntactically be analysed as consisting of more clauses. This means that in the case of (9) the grammatical classification follows completely from the discourse segmentation: because it is analysed as consisting of one SIU, it also receives one syntactic label. It should, however, be noted that there are also cases in which discourse segmentation and syntactic structure interact and

---

<sup>9</sup> For a full account of the annotation guidelines and exceptions to the rules, see the annotation manual in Appendix I.



where discourse segmentation is thus influenced by syntactic structure, as in cases where a distinction needs to be made between lists and multiple coordination (cf. 2.4.3).

Second, example (10) also consists of one nuclear SIU that is grammatically realised as an independent clause that takes the form and syntactic word order of a question, which means that, in Dutch, the subject and finite verb are inverted. The grammatical labels used in this study do not indicate such a change in word order, but because all punctuation marks are annotated all questions can be retrieved for analysis.

Third, example (11) again consists of one nuclear SIU. This clause can be categorised as having the communicative function of a directive, which, in its typical form, contains no subject and has an imperative verb (cf. Quirk et al. 1985: 87). Even though this clause does not have an overt subject, it is still analysed as an independent clause. Clauses that have the communicative function of directives can thus still be analysed as independent clauses, even though not all the central clause elements are expressed.

Fourth and finally, example (12) is analysed as consisting of two nuclear SIUs that are coordinated with each other. It was explained in 2.4.3 that the identification of SIUs in the case of coordinated elements is partly based on syntactic criteria, for instance in cases where a distinction needs to be made between lists and multiple coordination. It was also explained that the reason that examples such as (12) are analysed as consisting of two coordinated SIUs is that each coordinate has the form of an independent clause. This means that the second coordinate is still considered an independent clause despite the fact that the subject of this clause is ellipted. This often occurs in cases of coordination where, because the clauses are parallel in structure, they can be reduced by leaving clause elements unexpressed (cf. Quirk et al. 1985: 858). However, as coordination often involves ellipsis, in certain cases the question arises how the coordinated elements should be analysed. Consider in this respect (13) below:

- (13) <Ca><coord\_a\_phrase>This is the department responsible not just for the sensitive issues of asylum and immigration<coord\_a\_phrase><Ca>, <Cb><coord\_b\_phrase>but also for law and order and security against terrorism< coord\_b\_phrase><Cb>. <s41, newspaper articles>

On the one hand, (13) can be analysed as an elliptical version of clause coordination, in which case the subject and part of the complement are ellipted in

the second coordinate. On the other hand, this sentence can be analysed as a single clause that contains two coordinated phrases. Quirk et al. also identify these two ways of analysing examples like (13) and favour the latter one in the case of simple coordination (1985: 942), which involves the coordination of 'single grammatical constituents such as clauses, predications, phrases and words' (p. 973). The present study also favours the latter analysis and analyses such instances of coordination on the basis of carrying out a surface analysis of the clause elements that are overtly expressed. This means that in order for two coordinates to be identified as coordinated clauses, all central clause elements except for the subject need to be expressed in the second coordinate, as in (12) above. If more clause elements than the subject are omitted, this is not analysed as an elliptical version of clause coordination, but as coordination of the elements that are overtly expressed, i.e. of PPs in the case of (13). Note moreover that the reason (13) is analysed as consisting of two coordinated SIUs is because the coordinated phrases are presented as separate punctuation units (cf. 2.4.3). The grammatical labels then indicate that this concerns coordination at the phrase level and not at the clause level.<sup>10</sup>

If, on the basis of the criteria presented above, a SIU cannot be classified as an independent clause, it has to be determined to which of the other syntactic categories it belongs, i.e. a subordinate clause, an embedded clause or a clause fragment, or whether it should be classified as a phrase. In order to distinguish between these categories, a number of criteria have been set that are based on a

---

<sup>10</sup> It should be noted that the choice of grammatical labels applied to those coordinated phrases that are presented as separate punctuation units was mainly influenced by the idea to apply symmetrical labels in the case of coordination, with both labels being identical in make up and indicating exactly what elements are coordinated with each other (i.e. coord\_a\_phrase, coord\_b\_phrase). An alternative way of labelling such cases, which would perhaps have provided a better insight into the precise nature of coordination, could have been to use asymmetrical labels, indicating that the first coordinate is realised as an independent clause, in the case of example (13) above, and the second coordinate as a phrase that is coordinated with only a part of the independent clause that precedes it. However, it should be noted that the latter interpretation is intended by the use of symmetrical labels. It should also be noted that cases in which neither of the coordinates constitutes an independent clause can be easily identified, as these coordinates received the additional label <fragment> to clearly indicate that the nucleus of the sentence, the C, is not realised as an independent clause, as in the example below:

<Ca><coord\_a\_fragment>Nothing squishy<coord\_a\_fragment><Ca>, <Cb><coord\_b\_fragment>but still<coord\_b\_fragment><Cb>: <D><fragment\_NP>the lip-to-lip kind of kiss<fragment\_NP><D>. <s10704, short stories>

surface analysis, i.e. on the analysis and categorisation of clause elements that are overtly expressed. As for the identification criteria for subordination in English, the indicators of subordination as identified by Quirk et al. (1985: 997) are adopted. They state that a subordinate clause generally contains one or more of the following indicators (*ibid*):

- (i) the clause is initiated by a subordinating conjunction;
- (ii) the clause is initiated by a *wh*-element;
- (iii) initial elements in the clause are inverted;
- (iv) the presence of certain verb forms in finite clauses is determined by the type of subordinate clause;
- (v) the verb element in the clause is either non-finite or absent.

More than one of these indicators may be present in a subordinate clause, as in the following example:

- (14) British researchers have had a disproportionately large role in the initiation and development of the field, <D>as evidenced, for example, by the authorship of the major textbooks in this area<D> (e.g. Evans, 1982, 1989; Evans, Newstead, & Byrne, 1993; Garnham & Oakhill, 1994; Manktelow, 1999). <s5695, academic prose>

In (14) the appended satellite <D> is realised as a subordinate clause, which is indicated by both the subordinating conjunction *as* and the non-finite verb *evidenced*.

For the classification of Dutch subordinate clauses, the main identification criterion is the position of the finite verb, if present, in the clause (cf. Haeseryn et al. 1997: 1095). In Dutch subordinate clauses the finite verb occurs in clause final position, as opposed to its sentence-initial position in main clauses. Consider the following examples, where (15a) is the original sentence that consists of a sentence-initial subordinate clause followed by a main clause. (15b), on the other hand, is a rewritten version of the sentence-initial subordinate clause in (15a), and is reformulated as a main clause to indicate the changing position of the finite verb, marked in italics:

- (15a) Toen Pronk een paar dagen later tijdens een kennismakingsgesprek met Verdonk het woord niet *wilde* terugnemen, escaleerde de situatie en wees Verdonk hem woedend de deur. <s2474, newspaper articles>

(When Pronk a few days later during an introductory talk with Verdonk not *wanted* to take the word back, the situation escalated and a livid Verdonk showed him the door.)

- (15b) Pronk *wilde* een paar dagen later tijdens een kenningsmakingsgesprek met Verdonk het woord niet terugnemen.

(Pronk *wanted* a few days later during an introductory talk with Verdonk the word not take back).

In addition to the finite verb placement criterion, several of the indicators of subordination as identified by Quirk et al. also apply to Dutch. For instance, subordinate clauses can be identified by the initiating subordinating conjunction (cf. Haeseryn 1997: 539). Similar to the English *wh*-elements, there are a wide variety of relative clauses in Dutch that can be initiated by various elements that function as relative pronouns (cf. Haeseryn 1997: 858/859). Last, the presence of a non-finite verb in a clause also indicates that it can be classified as a subordinate clause (1997: 1101).<sup>11</sup> Dutch subordinate clauses can thus be identified by one or more of these indicators of subordination.

A type of subordinate clause that, on the cline of clausiness, is considered by Aarts (2007: 119, see above) to be the least prototypical instance of a clause is the verbless clause. In order for this clause to be consistently recognised as such and distinguished from clause fragments or phrases, it needs to either be initiated by a subordinating conjunction or it needs to be able to be reformulated as a clause if a form of the verb *to be* is added (cf. Quirk et al. 1985: 996; Haeseryn 1997: 928). Consider the following examples:

- (16) <A>With the Reverend Ian Paisley's DUP now the largest party in unionism<A>, he said Sinn Fein was exploring their position. <s515, newspaper articles>

- (17) Beter lijkt het, <1>zoals tot dusver gebruikelijk<1>, besluiten over crisisbeheersingsoperaties per geval te nemen. <s2874, newspaper articles>

(It seems better, <1>as has thus far been customary<1>, to make decisions about crisis management operations for each case separately).

---

<sup>11</sup> For a full specification of the identification criteria, such as a list of all subordinating conjunctions in both languages, see the annotation manual in Appendix I.

The prepended SIU in (16) is categorised as a verbless clause and not as a prepositional phrase because the preposition *with* here has the function of a subordinator (cf. Quirk et al. 1985: 1003, 1090 on *with* as subordinator and requirement that it is followed by NP, as it is in (16)). Moreover, it is possible to insert a missing form of the verb *to be*, as in *with the Reverend Ian Paisley's DUP now **being** the largest party in unionism*. In (17) the interpolated SIU is categorised as a verbless clause because it is initiated by the subordinator *zoals* (as) and because it can be reformulated as a finite clause by adding a form of the verb *zijn* (to be).

In addition to the categories of independent clause and subordinate clause, there is the category of embedded clauses. As embedded clauses are typically not presented as separate punctuation units in English and Dutch, which serves as one of the main criteria for assigning such a clause SIU status, the label embedded clause does not occur frequently in this study. This study only analyses those embedded clauses that have the function of complement clauses that are presented as separate punctuation units. The main case in which a SIU takes the form of an embedded clause is when several embedded clauses are coordinated with each other and the clauses are presented as separate punctuation units. Consider the following example:

- (18) <Ca><coord\_a\_emb>But that does not mean that Mr Martin's supporters are outside the bounds of civilised debate<coord\_a\_emb><Ca>,  
<Cb><coord\_b\_emb>or that any MP who speaks up for Mr Martin should be treated like a parliamentary leper<coord\_b\_emb><Cb>. <s1404, newspaper articles>

In (18) the verb *mean* is followed by two coordinated *that*-clauses. In the present study, the two coordinated SIUs receive the labels <Ca> and <Cb> and the syntactic realisation labels <coordination\_embedded> to indicate that it involves coordination of embedded clauses. In fact, this label indicates that the <Cb> SIU is coordinated with only a part of the <Ca> SIU.

In addition to the clausal categories discussed so far, there is the category of clausal fragments (cf. 3.5.1 for a detailed account of fragments in this study), which are positioned in-between clauses and phrases. The fragment label used to classify these cases indicates that a particular element belongs to a container category, which means that it cannot be classified as a particular type of clause on the one hand or a particular type of phrase on the other hand. Consider the following instances of clausal fragments:

- (19) How can I tell if someone has an STD? <s7853, leaflets> <C><fragment>You can't<fragment><C>. <s7854, leaflets>
- (20) <C><fragment>Terecht dus dat een aantal kabinetten op rij de vermindering van de 'administratieve lastendruk' hoog in het vaandel had<C><fragment>. <s2024>
- (<C><fragment>Rightly so that a number of cabinets in a row considered the reduction of the 'administrative expenses' of paramount importance<C><fragment>)

In (19) the second sentence is grammatically labelled as a clause fragment. The main reason for classifying it as such is simply that it does not meet the criteria of any of the other categories identified in the present study, and the label is thus used in the container sense. Even though it could be argued that it should be classified as a clause because the ellipted clause elements are recoverable from the context, as is often the case in answers to questions (cf. Quirk et al. 1985: 82, 883), in the present study independent clause status is based on the analysis of the clause elements that are overtly expressed. A similar analysis applies to the classification of (20), as this sentence too lacks the overt expression of the clause elements of subject and finite verb. The main reason for not classifying these sentences as phrases is that they do not consist of just a phrasal element.

Although it does not officially form a category on the cline of clausiness, in order to make a distinction between clausal fragments and phrasal fragments, a description of the latter is included in the present section. The label of phrasal fragment is given to those elements that cannot be classified as clauses of any type and consist of just a phrasal element. Consider the following examples:

- (21) Killing people is wrong. <s1397> <C><fragment\_NP>Full stop<fragment\_NP><C>. <s1398, newspaper articles>
- (22) <C><coord\_a\_phr>De investeringen van het bedrijfsleven, <1><coord\_b\_phr>en ook de overheid<coord\_b\_phr><1>, waren over het vorige jaar 3,4 procent minder dan in 2002<coord\_a\_phr><C>. <s2450, newspaper articles>
- (<C><coord\_a\_phr>The investments by companies, <1><coord\_b\_phr>and also the government<coord\_b\_phr><1>, were 3.4 percent less over the last year than in 2002<coord\_a\_phr><C>.)

In (21) the second sentence is syntactically realised as a noun phrase, to which the label <fragment> is added to indicate explicitly that the nucleus is not realised as an independent clause. Example (22) consists of two SIUs, a nuclear SIU that is interrupted by an interpolated SIU (cf. 2.5.3). The interpolated SIU is syntactically realised as a noun phrase that is coordinated with the subject of the nuclear SIU, also realised as a noun phrase. In the present study, both SIUs receive the label <coordination\_phrase> to indicate that the SIUs are syntactically related to each other by coordination at the phrase level. Note that the analysis of the sentence at both the level of discourse and syntax lays bare that the SIUs in (22) are hypotactically related to each other at the level of discourse and paratactically at the level of grammar.<sup>12</sup> Quirk et al. capture this mismatch between discourse and grammar by labelling the syntactic realisation of the interpolated SIU ‘interpolated coordination’ (1985: 976, cf. 3.4.1 below).

In short, even though the present study acknowledges a cline of clausiness, this cline is presented as though it consists of clearly distinguishable categories in order to guarantee consistent analysis of the grammatical realisation of SIUs. To be able to distinguish between the categories, identification criteria have been set. The classification of SIUs on the cline of clausiness precedes the step of determining what type of clause or phrase is involved, and is therefore very important in the analysis of the syntactic realisation of SIUs.

### 3.3.2 Form vs. function

Besides determining where a SIU should be positioned on the cline of clausiness, another important step in the categorisation process involves determining whether a SIU should be classified on the basis of its syntactic form or its function. Quirk et

---

<sup>12</sup> Note that in all cases of coordinated elements, both coordinates received grammatical labels that were similar in make up, such as coord\_a\_phrase/coord\_b\_phrase, mainly for reasons of symmetry. In retrospect it could be argued that particularly with examples such as (22), asymmetrical labelling may have provided a better representation of the discourse-grammatical structure of such sentences. Specifically, the nucleus could then have received the label <independent clause> and the interpolated satellite the label <coord\_b\_phrase>. However, it should be noted that in the analysis of the grammatical realisation of nuclei (see Chapter 5) such cases have been included in the count of nuclei that take the form of an independent clause. It should also be noted that only 34 (2.5%) of all interruptions analysed in this study take the form of <coord\_b\_phrase> (see Chapter 8 for a detailed analysis of sentence patterns formed by interpolated satellites).

al. also identify these two main ways of classifying constituents (1985: 48). If constituents are classified on the basis of their form, they are classified by their internal structure, for example as noun phrases or verb phrases. If, on the other hand, constituents are classified on the basis of their function, they are classified on the basis of the syntactic function they have in the clause, such as subject, verb, object, complement or adverbial. Quirk et al. make this distinction in order to explain ‘complicated facts of constituency’, such as the extent to which a unit’s position is fixed, the extent to which a unit is optional, the extent to which a unit is mobile, and so on (1985: 48).

In the present study a distinction is also made between form and function. If SIUs are classified on the basis of their form, they are classified by their internal structure, in line with Quirk et al.’s interpretation of form. However, if SIUs are classified on the basis of their function, this label does not provide any information about a unit’s ‘privilege of occurrence’ in terms of the characteristics listed above (1985: 48); rather, it is intended to provide more information about how this SIU is semantically related to the SIUs that surround it. A functional label is here taken to be ‘the most informative label possible’. The main purpose of providing a SIU with ‘an informative grammatical label’ is because this contributes to making a detailed sentencing analysis. This sentencing analysis provides further insight not only into how different SIUs within one sentence are hierarchically related to each other, which is captured by the discourse segmentation model designed for this study (cf. Chapter 2), but also into how these are functionally or semantically related to each other by, for instance, indicating that a prepended A satellite is in a concession relation (adverbial clause of concession) with the nucleus it precedes. The notion and relevance of ‘an informative label’ will be illustrated in the following section.

### **Application of form-function distinction to the analysis of data**

As categorizing SIUs on the cline of clausiness automatically involves classifying them on the basis of their form, it means that all SIUs in this study are in any case classified on the basis of their formal characteristics. However, in a number of cases a formal classification of SIUs proves less informative than a functional or semantic classification, which is why a functional classification may follow a formal one in a number of cases. These cases will be discussed in more detail below. First consider the following example:



- (23) <A><conjunct>In addition<conjunct><A>, unlawful possession of a controlled drug is a criminal offence. <s7411, leaflets>

In (23) the prepended satellite *in addition* can both be classified on the basis of its form and its function. In order to decide what category it falls in on the cline of clausiness, the satellite is first of all classified on the basis of its formal characteristics, i.e. a prepositional phrase in this particular case. However, this is not considered to be a very informative label in this particular example, as it does not give any information about the function of this phrase in the sentence as a whole. A functional classification of *in addition* involves labelling it as an adverbial. Within the category of adverbials, a further subdivision can be made into four main grammatical functions that Quirk et al. distinguish: adjuncts, subjuncts, disjuncts and conjuncts (1985: 501). Although the distinction between these four categories is not always clear-cut (cf. Quirk et al. 1985: 52, 501), by setting clear identification criteria it is possible to consistently distinguish between these four categories in the analysis of data.<sup>13</sup> With respect to (23), *in addition* is classified as a conjunct, which can be characterised by its conjoining or connecting function (1985: 631). Classifying it on the basis of its functional characteristics thus gives information about the function of this particular PP in the current sentence. For this reason, most adverb phrases or prepositional phrases that are presented as separate SIUs and that can be described in terms of their syntactic function, are classified as adjuncts, subjuncts, disjuncts or conjuncts.

A similar approach is taken to the classification of subordinate clauses. As subordinate clauses can have various functions in the sentence in which they occur, merely classifying them as subordinate clauses is here not considered to be the most informative label. Quirk et al. explain that subordinate clauses can function as subject, object, complement or adverbial in a superordinate clause (1985: 1047).<sup>14</sup>

---

<sup>13</sup> See the annotation manual in Appendix I for the identification criteria of the four grammatical categories of adverbials.

<sup>14</sup> In addition to distinguishing between different functions of subordinate clauses, a distinction has also been made between finite and non-finite clauses. It should, however, be noted that all non-finite clauses have also received the label 'adverbial clause', despite the fact that a number of non-finite clauses, e.g. participial clauses, are in fact not adverbial clauses. The motivation for classifying them all as adverbial clauses was mainly to ease the annotation process by restricting the total number of categories, and also because the vast majority do function as adverbial clauses. As the position of the clause in the sentence as a whole has been indicated at the level of discourse, non-finite clauses that most likely do not

As the present study is only concerned with classifying those subordinate clauses that have received SIU status, it means that the analysis of these clauses is automatically mainly restricted to those that have the function of adverbials in the superordinate clause, as these are the ones that are most likely to be presented as SIUs (or separate punctuation units, cf. 2.4.3). Within this class, a further distinction can be made between the functions of adjunct, conjunct, disjunct and subjunct, where it should be noted that most adverbial clauses function as either adjuncts or disjuncts (1985: 1068). However, the four functional labels are in the present study restricted to those adverbials that are realised as phrases instead of clauses. For the classification of subordinate clauses, a semantic classification is used, based on the one presented by Quirk et al. (1985: 1077ff). These labels indicating the semantic role are considered to be the most informative labels, as they provide information about the semantic function of the adverbial clause in the sentence. However, instead of adopting the full classification as presented by Quirk et al., the classification is restricted to a limited number of categories in order to ease the identification and classification of the various roles. The following semantic roles are distinguished: clauses of time, place, condition, concession, reason, purpose, result, comparison and comment clauses. Consistently distinguishing between these various roles can still be difficult, which is why clear identification criteria have been set.<sup>15</sup>

Another type of clause that is not classified on the basis of its formal features alone is the reporting clause. Consider the following examples:

- (24) <Cr><reportingcl><fragment>Prime Minister Bertie Ahern  
said<fragment><reportingcl><Cr>: <s530>  
<C><reportedcl><indepcl>It is obviously important to comply with the  
law<indepcl><reportedcl><C>. <s531, newspaper articles>
- (25) <C><reportedcl><indepcl>Het zou me niet verbazen,  
<1><reportingcl><fragment>vervolgde hij bijna trots<fragment><reportingcl><1>, als

---

function as adverbial clauses, for instance some of those that are presented as interpolated satellites, can easily be retrieved for reanalysis. It should, however, be acknowledged that a more detailed analysis of various types of non-finite clauses would have provided more insight into their precise nature and behaviour.

<sup>15</sup> See the annotation manual in Appendix I for guidelines on how distinctions are to be made between the different semantic roles of adverbial clauses.

hij hoogstpersoonlijk naar Rotterdam zou afreizen om Ed de Goey een toontje lager te laten zingen<indepcl><reportedcl><C>. <s15568, short stories>

(<C><reportedcl><indepcl>It wouldn't suprise me, <1><reportingcl><fragment>he continued almost proudly<fragment><reportingcl><1>, if he were to personally go down to Rotterdam to bring Ed de Goey down a peg<indepcl><reportedcl><C>.)

Both examples (24) and (25) contain a reporting clause, as indicated by the labels that surround them. Although both clauses perform a similar semantic function in that they both introduce direct speech, they are analysed in a different way as they occur in different positions in the sentence. The reporting clause in (24) could be analysed as a main clause if the direct speech that follows it is analysed as the object of that clause. However, as direct speech could extend over many sentences, in the present study it is never analysed as the direct object of the reporting clause, but instead as constituting a separate SIU, which can be realised syntactically in various ways (cf. 2.5.2). Thus, in (24) *Prime Minister Bertie Ahern said* is analysed as a reporting clause, despite the fact that it cannot be classified as a complete clause. In this respect, it forms an exception to the clause criterion rule that all necessary clause elements need to be expressed overtly. This exception is taken into account by both indicating the function of this clause fragment at the level of discourse segmentation by applying the discourse label Cr(eporting) (cf. 2.5.2) and by adding the grammatical label <fragment> to the reporting clause label. The 'analytical problem[s]' posed by such constructions, as Quirk et al. refer to them (1985: 1022), are thus taken into account by using grammatical labels that provide information about both the form and function of these types of clauses. In (25), on the other hand, the reporting clause occurs clause-medially, taking the form of an interpolated satellite, and can be analysed as a clause fragment (cf. Quirk et al. 1985: 1023). The particular behaviour of these types of reporting clauses is again both taken into account at a discourse level, by indicating their hierarchical status and position, and by the grammatical labels that specify both their functional and formal characteristics.

Consider in this respect one further exception to the rule of using the clause label for situations in which the relevant SIU does not meet the clause status criteria as formulated in 3.3.1 above:

(26) <Cr><reportingcl><fragment\_NP>Kamerlid Jan Boelhouwer <1>(PvdA)<1>  
<fragment\_NP><reportingcl><du\_Cr> : <s3449, newspaper articles>

<C><reportedcl><indepcl>Ik heb hetzelfde gevoel als Van der Horst<indepcl><reportedcl><C>. <s3450, newspaper articles>

(<Cr><reportingcl><fragment\_NP>MP Jan Boelhouwer  
<1>(PvdA)<1><fragment\_NP><reportingcl><du\_Cr>  
<C><reportedcl><indepcl>I have the same feeling as Van der Horst<indepcl><reportedcl><C>.)

In (26) the reporting clause is reduced to only the name of the speaker who produces the direct speech. Instead of merely classifying this name as an NP, the more informative label of reporting clause is added as well to specify the function of this particular NP in this sentence. The combination of the labels <reporting clause> and <fragment\_NP> thus indicate both its formal and functional characteristics, which is in line with the current aim to use the most informative label possible.

A further example that deserves attention here concerns non-restrictive apposition (cf. 3.4.2 below for a detailed account of the present definition of apposition). Consider the following sentences, in which the bracketed phrases are classified as appositions:

(27) Lord Ryder, <apposition\_NP>the acting chairman<apposition\_NP>, issued a cringeing apology to the government, even though lawyers had told the governors that the Hutton report was legally flawed. <s183, newspaper articles>

(28) De voortschrijdende techniek, <apposition\_PP>bijvoorbeeld op het gebied van DNA<apposition\_PP>, maakt dan wel steeds preciezer opsporing en bewijsvoering mogelijk, maar leidt ook tot een samenleving waar de staat de burger kan volgen in al zijn doen en laten. <s2972, newspaper articles>

(The advancing technique, <apposition\_PP>for example in the area of DNA<apposition\_PP>, does make more specific tracing and furnishing proof possible, but also leads to a society in which the government can monitor the citizen in all his actions).

Merely classifying the interpolated SIUs in (27) and (28) on the basis of their form would lead to their classification as an NP and a PP respectively. However, these labels do not give any information about how these phrases relate to the nuclear SIU they interrupt. Providing them with the label of apposition indicates that in (27) the information contained in the interpolated satellite realised as an NP is related

to the NP it follows by a relation of designation (cf. Quirk et al. 1085: 1310). In (28) the PP is related to the NP it follows by a relation of exemplification (1985: 1315).

One last example of a situation in which the grammatical labels used provide insight into both a SIU's grammatical form and its function is presented by the following phrases that occur in sentence-initial position, marked in bold:

(29) **Okay**, Alex squints up, straining for X-ray vision. <s10686, short stories>

(30) **Wow**. <s10690, short stories>

(31) **Nee serieus**, kippen en tantes lijken toch op elkaar, associatief gezien bedoel ik. <14991, short stories>

(No seriously, chicken and aunts do look alike, by association I mean).

Each of these phrases in bold receives the label <discourse marker> in the present study. As the grammatical category of discourse markers will be discussed in more detail in 3.5.2 below, the present discussion will be restricted to motivating why a functional label is used in such situations instead of a formal one. As can be seen in all three example sentences above, the label discourse marker is used for a wide variety of phrases that have varying formal characteristics. The reason for using the label discourse marker is because it can be argued that these phrases all perform a similar function in spoken discourse (cf. 3.5.2).

This section served to motivate why the present study sometimes classifies the grammatical realisation of SIUs on the basis of their formal characteristics, sometimes on the basis of their functional or semantic characteristics, and often on a combination of both formal and functional characteristics. As formal classification automatically occurs in this study when a SIU is positioned on the cline of clausiness, a more functional label may be added if it is considered more informative to indicate how a SIU is related to the surrounding SIUs.

### 3.4 Categorisation issues in combining relations

In cases where sentences consist of more than one SIU, at the level of discourse segmentation it is necessary to determine how these SIUs are related to each other, which can either be by coordination of two nuclei or two satellites, or by a nucleus-satellite relationship (cf. 2.5). In the grammatical categorisation of SIUs, the

relations between the different units have to be captured in grammatical terms. As was shown in example (22) above, where the relation between the two SIUs was classified as one of coordination at the level of grammar and a nucleus-satellite relationship at the level of discourse, these two levels of analysis do not necessarily have to overlap. Although many grammars make the traditional distinction between units that are paratactically related to each other, involving a relation between units of equal status, or hypotactically related, involving a relation between units of unequal status (Downing & Locke 2002: 280-281), it is generally acknowledged that the distinction between these relations is not always clear-cut (cf. Quirk et al. 1985: 918-919: see Huddleston & Pullum 2002: 1350 for a three-way distinction into coordination, dependency construction and supplementation; Cosme 2007: 39ff on gradient relation between coordination and subordination).

The following section is divided into two subsections, the first of which focuses on the difficulties in distinguishing between coordination and subordination and the second on defining the relation of apposition in relation to this distinction. Both subsections will describe and exemplify how these relations have been analysed in the present study.

### **3.4.1. Coordination vs. subordination**

Quirk et al. describe coordination and subordination as both involving the linking of units of the same rank. However, they specify that in coordination ‘the units are constituents at the same level of constituent structure, whereas in subordination they form a hierarchy’, with the subordinate unit being a constituent of the superordinate unit (1985: 918). There are various ways of distinguishing between and identifying coordination and subordination, a common one being by means of the explicit indicators of both types of linking constructions, i.e. coordinating and subordinating conjunctions respectively. In addition to listing clear examples of either type of conjunction, Quirk et al. also identify a number of conjunctions that are in an intermediate position between the two types (1985: 927, cf. Huddleston & Pullum 2002: 1289 for a similar observation). With these intermediate cases it can be difficult to determine whether the relation that they indicate should be classified as one of coordination or subordination. In order to make this decision, Quirk et al. present six criteria that can be applied to determine where on the coordination-subordination gradient a conjunction should be positioned. The more criteria a conjunction satisfies, the more it can be classified as a coordinating

conjunction. The fewer criteria a conjunction satisfies, the more it can be classified as a subordinating conjunction. Intermediate cases can then be characterised by the fact that they satisfy only some of these criteria, and are classified as semi-coordinators (1985: 927, 928). These criteria are as follows (taken from Quirk et al. 1985: 927):

- (a) The conjunction is restricted to clause-initial position.
- (b) A clause beginning with a conjunction is sequentially fixed in relation to the previous clause, and hence cannot be moved to a position in front of that clause.
- (c) The conjunction cannot be preceded by another conjunction.
- (d) The conjunction links not only clauses, but predicates and other clause constituents.
- (e) The conjunction can link subordinate clauses.
- (f) The conjunction can link more than two clauses, and when it does so all but the final instance of the linking item can be omitted.

Consider the following example of the coordinating conjunction *and* that meets all six criteria (taken from Quirk et al. 1985: 922):

- (32) He was unhappy about it, *and* yet he did as he was told.

In (32) *and* is classified as a coordinating conjunction because it satisfies all six criteria. For example, *and* satisfies criterion (a) in that it is restricted to clause initial position. It satisfies criterion (b) in that it is sequentially fixed in relation to the clause that precedes it. It satisfies criterion (c) in that *and* cannot be preceded by another conjunction such as *but* or *or*. In (32) the conjunct *yet* is preceded by the coordinating conjunction *and*. The fact that this is possible is a distinguishing feature between conjuncts and conjunctions, as the latter cannot be preceded by another conjunction. Further, *and* can link not only clauses, but also predicates, other clause constituents, and subordinate clauses, thereby satisfying criteria (d) and (e). Last, *and* satisfies criterion (f) because it can link more than two clauses, which creates a construction of multiple coordination.

Now consider (33), which contains an example of the semi-coordinator *yet*:

- (33) He tried hard, *yet* he failed.

Even though Quirk et al. classify *yet* as a conjunct, they distinguish it from other conjuncts such as *however* and *therefore* because it shares some of the distinguishing features of coordinators (1985: 928). One such feature is that in initial position, *yet* resembles coordinators in that it commonly occurs with asyndetic coordination (1985: 923). Huddleston and Pullum identify a similar set of intermediate cases, of which *yet* is one. They classify *yet* as a connective adverb, but explain that it has uses where the resemblances are such that it may be seen as a marginal member of the coordinator category (2002: 1319, and see also Biber et al. 1999: 80, who characterise *yet* as blurring the borderline between coordinators and linking adverbials). Similar to Quirk et al., Huddleston and Pullum identify a number of features that prototypical coordinators have. The features that *yet* shares with more prototypical coordinators are that it normally occurs in clause initial position; it can be combined with a coordinator, but typically occurs without one, and it can link a wide range of coordinates, not just clauses (2002: 1320).

Quirk et al. also identify a group of linking items that they label quasi-coordinators, as these sometimes behave like coordinators and sometimes like subordinators or prepositions. Consider *as well as* in (34), which links the verb phrases *publishes* and *prints* (1985: 982):

(34) He publishes *as well as* prints his own books.

By identifying a gradient, it becomes clear that a classification of the intermediate cases is dependent on a number of factors, such as the extent to which they satisfy the criteria identified or the extent to which they show features of the categories they are on a gradient with, such as subordinators, prepositions and conjuncts. In order to guarantee consistent classification, clear distinction criteria have to be set, as will be illustrated in the application of the coordination-subordination gradient section below.<sup>16</sup>

Although the at times problematic distinction between coordination and subordination also applies to Dutch (see section below), unlike in English, there are no obvious examples of Dutch conjunctions that have an intermediate status between these relations, which probably explains why no explicit reference is made to cases like these in Haeseryn et al. 1997. It should also be noted, as was exemplified by sentences (15a) and (15b) above, that a clear identification criterion

---

<sup>16</sup> For an overview of all intermediate cases, see the annotation manual in Appendix I.



for subordinate clauses in Dutch is presented by the word order, which is different between independent clauses and subordinate clauses mainly with respect to the position of the finite verb (cf. Haeseryn et al. 1997: 1095).

In addition to identifying conjunctions that are on a gradient between coordination and subordination, Quirk et al. also identify three less common types of coordination, i.e. complex coordination, appended coordination and interpolated coordination. Complex coordination can be distinguished from more common instances of coordination in that it involves the linking of combinations of units rather than single units. It is, however, the latter two types of coordination that are of particular interest in the present context, as it can be argued that these are also on a gradient between coordination and subordination. Huddleston and Pullum also acknowledge the distinguishing characteristics of these types of constructions and classify them under a different category, namely that of supplementation (2002: 1350). Let us consider each of these constructions in turn. Sentence (35) presents an example of what Quirk categorises as appended coordination (1985: 975):

(35) I am not sure whether Jane wrote the letter, or Sally.

They describe appended coordination as a loose kind of coordination, which they regard as clause coordination with ellipsis. The second coordinate is then considered to be added to the first one as an afterthought. This type of construction is related to those classified as appended clauses, briefly introduced and exemplified in Section 3.3.1 above, where *just* in *I caught the train – just* was classified as an appended clause by Quirk et al. (1985: 912). Appended clauses and appended coordination are similar in that both constructions can be classified as afterthoughts, but can be distinguished from each other by the coordinating conjunction that is present in appended coordination (1985: 918). Huddleston and Pullum, however, do not classify similar constructions as instances of coordination, but of supplementation, as in the following example (2002: 1362):

(36) I'm convinced it was masterminded by Tom – or Ginger, as everyone calls him.

The unit introduced by *or* in (36) is classified as a supplement instead of an instance of coordination on the basis of two main characteristics. The first of these is that supplements are not syntactically dependent on a head, but instead are

semantically related to a so-called anchor (2002: 1351). The second characteristic is that supplements are not integrated into the syntactic structure of the sentence to which they are attached (2002: 1350). Huddleston and Pullum thus distinguish between three types of linking relations, namely subordination, coordination and supplementation. This three-way distinction seems to imply that they claim that certain constructions cannot be classified along the lines of the traditional distinction between coordination and subordination. Although Quirk et al. label this type of coordination as a less common instance, they still categorise it as coordination.

In addition to appended coordination, Quirk et al. identify a less common type of coordination that they label interpolated coordination. A similar type of construction is also identified by Huddleston and Pullum, but is again not categorised as coordination, but instead as supplementation. Consider the following example (Quirk et al. 1985: 976):

(37) John – and Sally, too – writes extremely well.

Quirk et al. explain that one of the conjoins or coordinates in (37) behaves as if it is inserted in the middle of the clause, as a parenthesis. The interpolated element is again seen as elliptical reduction of clause coordination. A distinguishing feature of interpolated coordination that they identify is that the second conjoin is normally separated by prosody or punctuation from the rest of the clause and ‘thereby shows itself to be syntactically dislocated’ (1985: 976). Another distinguishing feature is that the second conjoin does not have to constitute a grammatical unit that can occur on its own. Even though Huddleston and Pullum’s examples are not identical to the ones Quirk et al. present, the constructions can be compared as they show some overlap. Consider (38) (2002: 1361):

(38) He told the manager – and her secretary – that the report was defamatory.

Huddleston and Pullum classify the interpolated element in (38) as an instance of supplementation rather than coordination. They explain that it is closely related to the coordinative construction in which the interpolation is not set apart from the rest of the sentence by means of punctuation. It is precisely because the interpolation is separated from the rest of the sentence by means of punctuation that it is classified as a supplement, as this separation causes it to be presented as secondary information instead of a coordinated element (cf. 2.5.3 for a similar

approach and argumentation with respect to interpolated satellites in this study and example (22) above). Not only Quirk et al.'s classification of such structures as less common instances of coordination, but particularly Huddleston and Pullum's classification of them as constituting a different type of linking relation underlines the fact that these constructions may be difficult to classify when the traditional labels coordination and subordination are taken as the common reference points. By also taking into account the discourse status of such elements, i.e. as either interpolated or appended satellites, as is done in the present approach to sentencing analysis, more light could be shed on their intermediate status in between subordination at the level of discourse and coordination at the level of grammar, as will be explained in more detail in the section below.

This section explained that there are situations in which it is not always clear to distinguish between coordination and subordination. This may be due to the intermediate status of certain conjunctions or it may be because units are linked through an appended or interpolated construction. In order to guarantee consistent categorisation, it is important to set clear guidelines that help in distinguishing between units that are linked through coordination and units that are linked through subordination, which is what the following section will go into.

### **Application of subordination-coordination distinction to the analysis of data**

This section will present the guidelines that have been set in the present study in order to deal with the gradient relation between coordination and subordination. It will first present the guidelines with respect to the conjunctions that are identified as having an intermediate status between the two linking relations, and then give a description of the guidelines that were developed to deal with the categories that are here labelled as appended clauses, appended coordination and interpolated clauses.

As was shown above, Quirk et al. identify a number of conjunctions that have an intermediate status between coordinators, conjuncts and subordinators, categorised as semi-coordinators and quasi-coordinators (1985: 928, 982). Guidelines on how to classify these intermediate cases in English in the present

study affect both the hierarchical classification of SIUs and their grammatical categorisation.<sup>17</sup> Consider (39):

---

<sup>17</sup> Note that the reason that the discussion of such intermediate cases focuses on English and less on Dutch is because English presented the main problems in the annotation process in this respect. Note also that only those intermediate cases are taken into consideration that link two or more SIUs *within* one sentence, not between different sentences (cf. Appendix I for a more detailed account of semi/quasi coordinators).

- (39) <Ca><coord\_a>They say they are sticking to their opposition on principle<coord\_a><Ca>, <Cb><coord\_b>**yet** they will be voting to harm the chance of working-class children going to university<coord\_b><Cb>. <s428, newspaper articles>

Sentence (39) consists of two independent clauses that are linked by *yet*. The two independent clauses are both classified as constituting separate SIUs, a decision that is based on syntactic criteria and not on punctuational criteria in the case of linked independent clauses (cf. 2.4.3 and 3.3.1 above). The next step involves determining the hierarchical status of the SIUs. This step is again based on syntactic criteria in this particular case, which means that the hierarchical classification and grammatical categorisation interact. Specifically, *yet* is classified as performing the function of a coordinating conjunction in this sentence, the reasons for which are threefold. The first is that *yet* has been identified as performing this function when it occurs in clause-initial position of the second clause<sup>18</sup>; the second is that it can be replaced by a coordinator such as *but* in this sentence; and the third is that if *yet* has the function of a coordinating conjunction it is often preceded by a comma, as it is here (cf. Quirk et al. 1985: 923, 1615). By classifying *yet* as a coordinating conjunction in this sentence, the hierarchical status of the SIUs can be classified as constituting two nuclei that are coordinated with each other and the grammatical categorisation can be classified as involving the coordination of independent clauses. The coordination of these clauses is seen as constituting syndetic coordination and not asyndetic coordination, precisely because *yet* is here classified as a semi-coordinator.

In addition to *yet*, the semi-coordinators *so*, *nor*, *for*, and *so that* can perform a similar function, which also applies to the quasi-coordinators *as well as*, *as much as*, *rather than* and *more than*.<sup>19</sup> Similar to all other instances of coordination, these conjunctions are only taken into account when the relation of coordination is annotated at the level of discourse, a decision that is based on both punctuational and syntactic criteria. This means that the quasi-coordinator *as well as* in (40) is not analysed as linking two separate SIUs, but simply two coordinated verb phrases *within* one SIU (cf. 2.4.3, example taken from Quirk et al. 1985: 982). Coordinators that occur within one SIU are not annotated in this corpus.

---

<sup>18</sup> Note that the present discussion of semi-coordinators and quasi-coordinators is restricted to those that occur *between* SIUs *within* one sentence.

<sup>19</sup> See the annotation manual in Appendix I for more examples of semi-coordinators and quasi-coordinators.

- (40) He publishes *as well as* prints his own books.

In addition to conjunctions that can be on a gradient between coordination and subordination, the constructions of appended coordination and appended clauses could be seen as being on a similar gradient, as was shown above. Consider the following example, which would be classified by Quirk et al.'s guidelines as appended coordination:

- (41) <C><indepcl>Not a day goes by without an attack on US troops or civilians<indepcl><C> - <D><appendedcl><coordinator>and<coordinator> usually both<appendedcl><D>. <s352, newspaper articles>

Sentence (41) is analysed as consisting of two SIUs, the first of which constitutes the nuclear unit that is followed by an appended satellite. An important reason for classifying the clause final SIU as an appended satellite and not as a nucleus that is coordinated with the nucleus it follows is presented by the dash that separates the two units. When compared to the comma, the dash marks a stronger break from the surrounding text and is 'not used to separate coordinates' (Huddleston & Pullum 2002: 1750, see also Dale 1991; Jones 1996: 129ff on the role of punctuation in distinguishing nuclear from satellite units, and Chapter 7 for a more elaborate account of the relation between different types of punctuation mark and the hierarchical status of the units that they link). Another reason for classifying this unit as a satellite and not a coordinated nucleus is based on the grammatical realisation of this unit. It would be classified by Quirk et al. as appended coordination, which is described as a 'loose kind of coordination', in which the second conjoin is added as an afterthought (1985: 975). It is precisely because it can be classified as having the status of an afterthought that it is here taken to involve a satellite relation with the nucleus that precedes it. Note furthermore that the appended satellite in (41) is classified as an appended clause and not as appended coordination, as no distinction is made between these two categories in the present study.<sup>20</sup> Sentences such as (41) thus exemplify the added value of

---

<sup>20</sup> See the annotation manual in Appendix I for more examples of appended clauses. It should be noted that appended clauses constitute a broad category in the present study and that appended satellites are classified as such when they cannot be classified as any other type of subordinate clause that provides a more informative label. In this respect, it has the function of a container category.

performing both a discourse and a grammatical analysis of sentences, as an analysis at both these levels can provide insight into how these levels interact with each other. This is particularly interesting when the relation can be seen as one of hypotaxis from a discourse perspective and parataxis from a grammatical perspective (cf. Quirk et al. 1985: 919 for classification of appended coordination and appended clauses as constituting paratactic relations, and see Huddleston and Pullum 2002: 1350 for the notion of supplementation to classify such intermediate cases).

Another example of what is classified as an appended clause in the present study is presented by (42):

- (42) <C><indepcl>De bruto bedrijfsinvesteringen nemen komend jaar met 2 procent af<indepcl><C>, <D><appendedcl>iets positiever dan uit de vorige berekeningen naar voren kwam<appendedcl><D>. <s2553, newspaper articles>

(<C><indepcl>The gross company investments will decrease by 2 percent this coming year<indepcl><C>, <D><appendedcl>somewhat more positive than the earlier calculations showed<appendedcl><D>.)

This example clearly shows that the category of appended clause is taken to be a broad category that allows for a wide range of constructions that do not fit any of the other categories identified in this study. Characteristics of the elements that are classified as such are that they always occur in clause-final position; that they are always separated from the unit they follow by at least a comma, but usually a dash, and that they have the discourse status of satellites. It should be noted that the category of appended clauses as it is here defined shows overlap with or can be positioned in between the categories of non-restrictive relative clauses on the one hand and appositions on the other hand. The main way to distinguish between relative clauses and appended clauses is that the former needs to be introduced by a relative pronoun. The main way to distinguish between apposition and appended clauses is that apposition is restricted to constituting a relation between phrases (cf. 3.4.2 below). This means that even though the satellite in (42) cannot be classified as a full clause, it is still not considered an instance of apposition, because this satellite does not refer back to, nor is it linked with a phrase in the preceding nucleus. Instead, it refers back to the information that is contained in the full clause that precedes it. It is, however, acknowledged that the distinction between what here counts as an appended clause on the one hand or apposition on the other hand is not always clear cut. In order to achieve consistent identification and

categorisation, both categories have been exemplified in great detail in the annotation manual.<sup>21</sup>

Besides appended coordination, Quirk et al. identify a type of coordination that they label interpolated coordination (1985: 973). It was shown above that Huddleston & Pullum (2002) classify similar constructions as supplementation instead of instances of coordination. These diverging classifications already seem to point to the intermediate status of such constructions. The present approach to such cases will be explained on the basis of the following example (presented above as (22)):

(43) <C><coord\_a\_phrase>De investeringen van het bedrijfsleven,  
<1><coord\_b\_phrase>en ook de overheid<coord\_b\_phrase><1>, waren over  
het vorige jaar 3,4 procent minder dan in 2002<coord\_a\_phrase><C>. <s2450,  
newspaper articles>

(<C><coord\_a\_phr>The investments by companies, <1><coord\_b\_phr>and also  
the government<coord\_b\_phr><1>, were 3.4 percent less over the last year  
than in 2002<coord\_a\_phr><C>.)

In line with Quirk et al.'s analysis, the construction in (43) is classified as an instance of coordination at the level of grammar. Because it concerns coordination at the phrase level, i.e. of the PP postmodifications of the noun *investeringen* (*investments*), the interpolated phrase receives the label <coordination\_b\_phrase>. Even though this phrase is only coordinated with a part of the clause that it interrupts, the entire surrounding clause receives the label <coordination\_a\_phrase>, to indicate that this concerns coordination at the phrase level. It is only when the clause that the interpolation interrupts does not constitute an independent clause that an addition is made to the label <coordination\_a\_phrase> to indicate that it interrupts a non-independent clause. Furthermore, what the analysis of (43) neatly shows is how the present analysis of sentences at both the level of discourse and grammar makes clear how these interact and can thus provide insight into the relation between discourse and grammar. Specifically, the analysis of (43) shows that the interpolated element is also identified and labelled as such at the level of discourse structure, i.e. as an interpolated satellite. This means that in the present system this type of

---

<sup>21</sup> See the annotation manual in Appendix I for more guidelines and examples of both apposition and appended clauses.



construction is considered to constitute the linking of units that are hierarchically related to each other at the level of discourse, indicated by the fact that one unit is in a satellite relation with the nuclear it interrupts, while being in a coordination relation at the level of grammar. This analysis is in line with Quirk et al.'s description of the interpolated conjoin as an element that is 'inserted, as a parenthesis' in the sentence (1985: 976), but in the present system also clearly represented and visualised as such by the annotation at both levels of analysis.

This section has presented a number of cases in which no clear distinction can be made between coordination and subordination. In the case of conjunctions that have an intermediate status between the two relations, criteria have been set to determine what type of relation a conjunction introduces in a particular situation. In the case of both appended coordination and interpolated coordination, the present system of analysis is able to capture the indeterminate status of units by both describing the relation at the level of discourse and grammar, which, as these examples show, do not always have to overlap.

### 3.4.2 Apposition

In addition to describing the gradient relation between coordination and subordination, the section on combining relations should also include a discussion of apposition, as this too has been classified as a gradient relation by some and involves the combining of various units in a sentence (cf. Quirk et al. 1985: 1300; Meyer 1992; Acuna-Farina 1999; Hannay & Keizer 2005). As Quirk et al. observe, '[g]rammarians vary in the freedom with which they apply the term apposition' (1985: 1302). Some restrict the definition to a relation that primarily exists between noun phrases that are identical in reference (cf. Quirk et al. 1985: 1301; Haeseryn et al. 1997: 846; Huddleston & Pullum 2002: 1357). Others apply the term more loosely and have expanded the category to also include instances of non-nominal apposition, in which case the appositives can take the form of other types of phrases or even a clause (cf. Meyer 1992; Acuna-Farina 1996, 1999, also for an extensive overview of different perspectives). When defined more narrowly, the term apposition is applied if each of the following conditions is met (Quirk et al. 1985: 1302):

- (44) (a) Each of the appositives can be separately omitted without affecting the acceptability of the sentence.
- (b) Each fulfils the same syntactic function in the resultant sentences.

- (c) It can be assumed that there is no difference between the original sentence and either of the resultant sentences in extralinguistic reference.

Appositions that meet all three conditions are labelled 'full apposition' and appositions that do not meet all three are labelled 'partial apposition'. Quirk et al. also make a distinction between so-called strict and weak apposition, where 'strict' means that both appositives are of the same syntactic class and 'weak' that the appositives belong to a different syntactic class (1985: 1303). This means that an example such as (45) would be classified as partial weak apposition (taken from Quirk et al. 1985: 1321):

- (45) His explanation, that is to say that he couldn't see the car, is unsatisfactory.

Example (45) is classified as partial apposition because it does not fulfil condition (a): the first appositive, *his explanation*, cannot be omitted without affecting the acceptability of the sentence. It is classified as weak apposition because the first appositive, *explanation*, and the second appositive, *that is to say that he couldn't see the car*, do not belong to the same syntactic class.

In line with Quirk et al.'s distinction between full and partial apposition, Meyer (1992), in his extensive study on apposition, presents similar criteria and classifies apposition as a gradable relation with constructions that are most appositional or 'central appositions' and constructions that are less appositional or 'peripheral appositions' (p. 41). If the two units of apposition are structurally independent of each other, the construction is characterised as a central apposition. If, on the other hand, the two units are structurally dependent on each other, it is characterised as a peripheral apposition. In order to distinguish between central and peripheral apposition, Meyer posits the following syntactic criteria (1992: 41):

- (46) (a) The first unit of the apposition can be optionally deleted.  
 (b) The second unit of the apposition can be optionally deleted.  
 (c) The units of the apposition can be interchanged.

If a construction fulfils all three criteria, it is classified as central apposition and if a construction does not meet all criteria it is classified as peripheral apposition (1992: 42). Meyer also classifies constructions as peripheral apposition if they are on a gradient with other relations, such as coordination, modification or

complementation (cf. Quirk et al. 1985: 1301). These perceived correspondences will be considered more closely, as it is of particular importance to the present study that a distinction between these different relations can be made in order to guarantee the consistent classification of the grammatical realisation of SIUs. The discussion will, however, be restricted to those appositions that Quirk et al. label non-restrictive appositions, which are described as constituting separate information units, classified as such because they constitute separate tone units or separate punctuation units, and thus also constitute separate SIUs (1985: 1303-1304). Quirk et al. characterise non-restrictive apposition as parenthetical (1985: 1304), which is reflected in their discourse status as typically constituting interpolated or appended satellites.

As for the correspondence between apposition and coordination, Quirk et al. explain that this similarity is not only brought about by the fact that both involve the linking of units of the same rank, but also by the fact that the coordinators *and* and *or* may occasionally be used as markers of apposition (1985: 1301, also see Meyer 1992: 43). Consider in this respect (47), taken from Meyer (1992: 44)

(47) They have a thing called *First University Examination*, or *FUE*.

Meyer explains that when taking the syntactic criteria into account, the italicised elements in (47) are not in apposition, as neither of the NPs can be deleted because one of the NPs is then left with the coordinating conjunction *or*. From a semantic perspective, however, the NPs are appositional, as they are coreferential, which Meyer identifies as a key semantic characteristic of the majority of appositional relations (1992: 58). He concludes by saying that it is 'extremely difficult to differentiate coordinative apposition from simple coordination', as the relations might be different from each other from a semantic perspective but not always from a syntactic perspective (1992: 45, but see Acuna-Farina 1999 for a critical view of Meyer's broad definition of apposition and classification of it as a gradable relation). Quirk et al. label appositions that do not satisfy all criteria and therefore show overlap with coordination as partial apposition, thus acknowledging the gradient relation of the concept.

In addition to coordination, apposition also shows overlap with non-restrictive postmodification. In the following example Quirk et al. (1985: 1301) explain that the NP *my best friend* may be considered a reduced form of a non-restrictive relative clause:

(48) Anna, my best friend, was here last night.

They explain that some grammarians have included non-restrictive relative clauses as apposition (eg. Jespersen 1961; Burton-Roberts 1975) and theorise about what may have motivated this decision (p. 1301). They present the following three reasons: first, the fact that the second appositive can often be expanded into a relative clause; second, the loose attachment or optional status of the non-restrictive relative clause to the sentence; and third, the requirement of co-reference between the *wh*-word in the clause and an antecedent NP (*ibid*). However, they decide not to accept non-restrictive relative clauses as appositions mainly because apposition involves linking units of the same rank, and primarily NPs (1985: 1301, see also Burton-Roberts 1975, 1999; Meyer 1992: 55, on the correspondence between these constructions).

Besides classifying apposition as a gradable relation on a syntactic or structural level, Quirk et al. (1985) and Meyer (1992) also classify it as a gradable relation in terms of semantics. They identify 'a semantic scale running from equivalence (i.e. 'most appositive') to loose and unequal relationship ('least appositive')' (Quirk et al. 1985: 1308, Meyer 1992: 90). The relation of equivalence subsumes relations such as appellation, identification, designation and reformulation. The relation of attribution is positioned in between equivalence and inclusion. And the least appositive semantic relation is the one of inclusion, which includes relations such as exemplification and particularisation (Quirk et al. 1985: 1308, see also Hannay & Keizer 2005 for a different classification of the semantic relations). In his discussion of the semantic classes of apposition, Meyer presents the following example, categorised as an instance of the semantic class of identification, the most appositive type of class (1992: 76):

(49) The Living Room has *another scoop: Jane Russell will make one of her rare night club singing appearances there, opening Jan. 22.*

Meyer explains that the italicised elements in (49) are in apposition, with the second appositive identifying the scoop referred to in the first appositive. The fact that Meyer classifies this as an appositional relation is quite controversial. He defends it by stating that the relation of apposition should not be restricted to a relation between two coreferential NPs, as this would severely limit the class of apposition (1992: 3, 57). At the same time, he acknowledges that the class of appositions should not become too broad either, as a point could then be reached

where ‘virtually any construction satisfying the literal definition of apposition (i.e. “placed alongside of”) is considered apposition’ (1992: 3,4). In his view, he prevents that from occurring by clearly listing the criteria that appositions should meet, not only syntactically, but also semantically and pragmatically. In addition to that, he distinguishes between central and peripheral apposition, as described above. Meyer’s definition is, however, still considered far too loose by others, such as Acuna-Farina (1999). Quirk et al. also restrict apposition to constituting a relation between NPs, even though they acknowledge that ‘appositive-like’ relations exist between other units than NPs. They argue that ‘to talk about apposition of units other than noun phrases makes the concept of apposition too weak’ (1985: 1308).

Even though these diverging perspectives illustrate that apposition does indeed appear to be ‘a term that is often used in the literature with a remarkable lack of precision’ (Acuna-Farina 1999: 59) and about which much controversy exists, a definition has to be formulated for the present study that not only finds the right balance between these varying perspectives, but that can also be applied consistently to the grammatical categorisation of SIUs.

#### **Apposition: definition and application to the analysis of data**

In line with Meyer’s view (1992) on the relation of apposition, the present study adopts a fairly broad definition of this relation. However, in order to prevent the category from becoming too broad and therefore too weak or meaningless (cf. Quirk et al. 1985: 1308; Meyer 1992: 3-4; Acuna-Farina 1999: 62), apposition is restricted to constituting a relation between phrases and not between clauses (see 3.3.1 above for the distinction between clauses and phrases). It is, however, not restricted to constituting a relation between noun phrases. Instead, the category is extended to also include relations between other types of phrases, such as adjective phrases or prepositional phrases. The main motivation for this extension relates to the aim of providing the most informative label in classifying the grammatical realisation of SIUs (cf. 3.3.2 above on form vs. function). To illustrate how this aim relates to the category of apposition, consider (50):

- (50) Means-tested benefits more than doubled - <1><appos\_PP>from 16% of all benefits to 34%<appos\_PP><1> - under the previous 18 years of Conservative rule. <s1297, newspaper articles>

In (50) the interpolated satellite, <1>, takes the form of a prepositional phrase. If included in the category of apposition at all, even though it does not satisfy any of the three criteria listed above for full apposition, Quirk et al. would classify this as an instance of weak apposition, because the appositives, the VP *doubled* and the PP *from 16% of all benefits to 34%*, do not belong to the same syntactic class (1985: 1303). Meyer, on the other hand, would classify this as an instance of non-nominal apposition, even though he does reserve the label peripheral apposition for appositions of this type (1992: 30). The present study adopts Meyer's line of thought, as the classification of the PP in (50) as an apposition is considered to be the most *informative* label. Despite the fact that it is acknowledged that this presents a non-standard approach to classifying the relation between these units, the motivation for it lies in the aim to distinguish this PP from those PPs or adjuncts that are not as closely related semantically to the unit that they follow. Specifically, in (50) the PP specifies what is meant exactly by the verb *doubled*. If the PP were only to be classified on the basis of its form in this example, this would not provide any information on how exactly it is related to the words that precede it. If, on the other hand, this PP is classified as an apposition, the combination of both its SIU status as an interpolated satellite and the grammatical label of apposition provide detailed information on how this satellite is hierarchically and semantically related to the nucleus it interrupts. Moreover, although no distinction is made between full and partial or central and peripheral apposition in the present study, the syntactic class of the apposition has been indicated, as the label in (50) shows, which does make it possible to determine what type of apposition is involved.

As a consequence of the relatively broad definition of apposition, the criteria set by Quirk et al. (1985) and Meyer (1992) to distinguish between full and partial apposition or between central and peripheral apposition respectively have not been followed in the present study. The main identification criteria that are applied are as follows. First, in terms of discourse structure, the element that receives the label apposition has the discourse status of either an interpolated or an appended satellite. This means that the prepended satellite in (51) is not classified as an apposition in this study, but as a verbless clause (cf. 3.3.1 on the exceptional status of verbless clauses, and see Huddleston & Pullum 2002: 1358 on clause-initial NP supplements; Quirk et al. 1985: 996ff).

- (51) <A><subcl\_advcl\_verbless>A complex and sensitive organ<subcl\_advcl\_verbless><A>, the brain is carefully protected inside the bony skull and it regulates and controls everything we do. <s8111, academic prose>

Second, the apposition takes the form of a phrase and not a clause, which means that the appended satellite in (52) is not classified as an apposition.

- (52) <C><indepcl>Mr Blair's case is this<indepcl><C>: <D><indepcl>Universities need more money<indepcl><D>. <s215, newspaper articles>

Even though the semantic relation between the two SIUs in (52) could be classified as a relation of identification, where the second SIU identifies what is mentioned in the first SIU, the appended satellite is not classified as an apposition, because it constitutes a clause. Instead, both SIUs have received the label <independent\_clause>. Although these labels do not provide much information about the particular relation between these two units, as all punctuation marks have also been annotated in the present study, the fact that these clauses are linked by a colon provides more insight into the relation between them (cf. Chapter 7 on a closer analysis of punctuation marks). Third, the relation of apposition can be explicitly marked by an indicator of apposition, such as *namely*, *that is* or *in other words*, which also expresses the semantic relationship between the appositives (cf. Quirk et al. 1985: 1307).<sup>22</sup> Even though the appositions have not been further categorised into semantic classes, these indicators have been used in the identification of apposition.

As the distinction between apposition on the one hand and coordination and non-restrictive relative clauses on the other hand is not always clear-cut (see above), clear criteria need to be set to distinguish between these categories and consistently classify the grammatical realisation of SIUs. First, as Meyer also showed in (47) above, repeated here as (53) for convenience, in certain cases it is 'extremely difficult' to distinguish between apposition and coordination, especially when the elements are linked by the coordinative conjunction *or* (cf. Quirk et al. 1985: 1302)

---

<sup>22</sup> See the annotation manual in Appendix I for an overview of explicit indicators of apposition.

- (53) They have a similar thing called *First University Examination*, or *FUE*.

On the basis of Meyer's criteria, the italicised elements do qualify as appositives at a semantic level, but not at a syntactic level. With respect to the present study, although the grammatical classification of SIUs is presented as though it follows discourse segmentation, this example presents a situation in which the two steps in the annotation process interact. The sentence-final punctuation unit would here be classified as a SIU simply because it constitutes a separate punctuation unit (cf. 2.4.3). The next step would then involve determining its hierarchical status, which is either that of a coordinated SIU or an appended satellite. This distinction is mainly made on semantic grounds, as the coordinative conjunction *or* is not classified as such in this example, but as an explicit indicator of apposition that indicates the semantic relation of reformulation (cf. Quirk et al. 1985: 1311). Because of the classification of *or* as such, it means that this SIU is identified as an apposition and therefore also receives the discourse status of satellite and not of a coordinated SIU. Now consider (54), in which the Dutch coordinative conjunction *en* (*and*) is classified as a marker of apposition:

- (54) Of anders gezegd, de buitenlandse politiek - <1><appos\_NP>en de Vietnamoorlog in het bijzonder<appos\_NP><1> - bleek uitermate geschikt om van ' een progressieve grondhouding ' te getuigen. <s3786, academic prose>

(Or put differently, the foreign politics – <1><appos\_NP>and the Vietnam War in particular<1><appos\_NP> – proved to be particularly suitable for demonstrating 'progressive principles'.)

The interpolated satellite, indicated by <1>, is classified as an apposition in the present study both because of its discourse status, which is reinforced by the dashes that surround it, and because of the marker of apposition *in het bijzonder* (*in particular*), which expresses the semantic relation of particularisation (cf. Quirk et al. 1985: 1316 and Meyer 1992: 76).

In addition to being on a gradient with coordination, apposition has also been described to show overlap with non-restrictive relative clauses (see above). However, as apposition is here restricted to constituting a relation between phrases, this distinction is fairly clear-cut in the present study (see 3.3.1 above for the distinction between clauses and phrases). This means that the interpolated



satellite in (55) is classified as an apposition, whereas the satellite in (56) is classified as a non-restrictive relative clause:

(55) Thirty-four years later the family of Professor Jellinek - <1><appos\_NP\_list>an assimilated Jew and son of a Rabbi<appos\_NP\_list><1> - had not reason to celebrate as their fate lay now in the hands of Jellinek's former student.  
<s4291, academic prose>

(56) Het ziet er naar uit dat het kabinet voornemens is de missie van de Nederlandse militairen in Irak, <1><nonrestr\_relcl>die in juli afloopt<nonrestr\_relcl><1>, te verlengen. <s2489, newspaper articles>

(It looks as though the cabinet is planning the mission of the Dutch militaries in Irak, <1><nonrestr\_relcl>which ends in July<nonrestr\_relcl><1>, to extend.)

Note that because the apposition in (55) consists of two coordinated noun phrases, this apposition is classified as apposition list, a label that is applied to situations in which the second appositive consist of more than one phrase that function together as one appositive (cf. Quirk et al. 1985: 1306). These situations in which more than two units are in appositions can be of two main types. The first type is presented by (55), in which the second appositive consists of more than one phrase that together function as one appositive. The second type is presented by (57), where there is a hierarchy of appositional relationships:

(57) Second, the new man in charge of the Home Office's police standards unit, <1><appos\_NP>Paul Evans, <i><appos\_NP>the former Boston police chief<appos\_NP><i> <appos\_NP><1>, will be there to stop British commentators over-romanticising the FBI. <s1196, newspaper articles>

In (57) the interpolated satellite, <1>, is realised as an apposition that is followed by another satellite that is again lower in hierarchy, <i>, which also takes the form of an apposition. These appositions are thus hierarchically related to each other, something that is clearly indicated in the present study at the level of discourse structure (cf. 2.5.3).

In short, even though apposition is acknowledged to constitute a relation that has been and can be defined in various ways, the particular approach that has been adopted here is motivated by the aim to provide SIUs with the most informative grammatical label and to guarantee consistent classification of the grammatical realisation of SIUs.

### **3.5 Genre-specific categorisation issues**

In addition to the more general classification issues, there are a number of classification issues that were brought about by the particular genres that were included in this corpus, such as the public information leaflets genre and the short stories genre. The short stories genre posed particular problems as this contains large parts of simulated dialogue. Even though these pieces of simulated dialogue cannot be classified as real conversation, they do show overlap with real conversation and therefore give rise to certain discourse segmentation and grammatical classification issues that are particular to the spoken language (cf. 2.3.1 & 2.3.2 on discourse segmentation of spoken language). The way in which these issues have been dealt with in the present study is by broadening a number of categories in order to allow for a wider range of structures. The main motivation for not extending the total number of categories is driven by practical purposes, such as easing the categorisation procedure and thereby achieving consistent annotation. An additional reason is that the present study focuses on the structure of written discourse and not spoken discourse (but see 2.2 on the continuum of spoken - written language). The following sections will exemplify the problematic categories in more detail.

#### **3.5.1 Fragments**

In the present study the category of fragments is designed as a fairly broad category that subsumes fragmentary phrases or clauses of various types. It is mainly used to classify the grammatical realisation of SIUs in the short stories genre, especially in those parts of text that simulate dialogue. As Biber, Johansson, Leech, Conrad & Finegan (1999) devote a considerable part of their comprehensive grammar of the English language to the analysis of spoken data, their classification and categorisation of spoken data is seen as being more extensive than Quirk et al.'s (1985) in this respect. For instance, whereas Quirk et al. classify fragmentary sentences as irregular sentences (1985: 838), which implies a biased categorisation towards the written mode, Biber et al. provide a fairly detailed classification of what they call non-clausal units (1999: 1069). They subdivide the class of non-clausal units into two main types, namely inserts and syntactic non-clausal units. The present definition of fragments (see below) shows overlap with Biber et al.'s category of syntactic non-clausal units, although it should be noted that their

classification is much more detailed and extensive. Even though Biber et al. explain that syntactic non-clausal units are often classifiable according to standard phrase categories or certain types of dependent clause, in their description they also focus on the particular function that these units perform and classify them along those lines, by using category names such as elliptic replies, condensed questions, echo questions, and so on (1999: 1099-1104). In the present study, these syntactic non-clausal units have mainly been categorised on the basis of their form and not their function. The main motivation for this is that the different functions of the non-clausal units may not always be easy to distinguish from each other, which could jeopardise consistent categorisation. In addition to that, as these particular types of units predominantly occur in just one genre, and only in the simulated dialogue parts of that genre, it would mean that a fair number of categories would be especially designed for this one genre. This does not correspond to the present aim of restricting the total number of categories in order to ease the categorisation procedure and to increase consistent annotation.

In the present study, the label fragment is used in two situations. The main use concerns the situation in which nuclear SIUs cannot be classified as independent clauses, but either take the form of subordinate clauses or of phrases. For these cases the fragment label serves to indicate that the grammatical categories that are typically associated with a grammatical status of dependency are presented as independent discourse units with nuclear status. In addition to the basic fragment label, these labels always contain additional information about the precise grammatical nature of the SIU, such as a specification of its syntactic class. The other situation in which the fragment label is used is when it functions as a label that indicates that a particular element belongs to a container category, which means that the particular element cannot be classified as belonging to any of the other categories that also occur in the other genres. Instead of subcategorizing all these elements into categories that only occur in one particular genre, the practice has been to group them into one category that bears the label fragment. The main characteristic of these fragments is that they typically have a status that is in between clauses and phrases, such as clause fragments. Both uses of the fragment label will be exemplified in the paragraphs below.

As an illustration of the fragment label in the present study, consider the following piece of simulated dialogue, taken from an English short story:

- (58) (1) <C><reportedcl><indepcl>Shall we go for a walk<indepcl><reportedcl><C>?  
<D><reportingcl><fragment>Andy asked<fragment><reportingcl><D>.  
<s12724, short stories>
- (2) <C><reportedcl><fragment\_no>No<fragment\_no><reportedcl><C>. <s12725,  
short stories>
- (3) <C><reportedcl><fragment\_PP>Just down to the beach  
<fragment\_PP><reportedcl><C>? <s12726, short stories>

First consider the use of the fragment label in (58.1), in the appended satellite <D>. In this example it is used in combination with the label <reporting\_clause> to indicate the particular function of this SIU, which is to be contrasted with the SIU that precedes it that contains the direct speech, indicated by the label <reported clause>. As reporting clauses have this particular function in combination with the reported clauses that they often accompany, they receive a label that indicates its function and its syntactic form (see 3.3.2 above). Another example of the use of the fragment label is presented in (58.2). This nuclear SIU consists of the word *no*, which is an answer to a question posed in (58.1). Because of its function as an answer to a question and because it is presented as a separate unit, it is here analysed as a nuclear SIU (cf. 2.5.1 on determining nuclearity). The fragment label is used to indicate that this SIU is not realised as a particular type of clause and the addition <\_no> to the label specifies its exact function. Another example of the main use of the fragment label is presented by (58.3), a nuclear SIU that is realised as a PP. As the default grammatical realisation of nuclear SIUs is a clause of some type, usually an independent clause, this label serves to underline that this nuclear SIU deviates from this default realisation; that it occurs in a dialogue (reported clause), and that it is realised as a PP. It thus underlines that the nuclear unit is syntactically realised as a phrase.

In addition to the high frequency of the fragment label in the short stories genre, this label is also used to categorise a particular syntactic situation that is relatively frequent in the genre of public information leaflets. Consider in this respect (59):

- (59) <C><fragment\_comp\_list>Geschikte organen zijn<fragment\_comp\_list><C> :  
<D><np\_list>lever, longen, hart, nieren, alvleesklier  
(<1><appos\_NP>pancreas<appos\_NP><1>) en de dunne darm<np\_list><D>. <s8898,  
leaflets>

(<C><fragment\_comp\_list>Suitable organs are<fragment\_comp\_list><C> :  
 <D><np\_list>liver, lungs, heart, kidneys, pancreas  
 (<1><appos\_NP>pancreas<appos\_NP><1>) and the small intestine<np\_list><D>.)

Public information leaflets in both languages can be characterised by the fact that much information is presented in the form of lists (cf. van den Boomen & van der Lans 1991: 103, 114-115; Woerkum & Kuiper 1995: 128ff). One such example is presented by (59). The discourse unit that introduces the list, which usually takes the form of a nuclear SIU <C>, is often realised as an incomplete clause, in the sense that the obligatory complement of *of*, for example, the verb is realised as a list of items that is presented in a separate SIU. To indicate this particular function of the incomplete clause in this genre, the fragment label has received the addition <complement\_list>, which indicates that the fragment takes the form of an incomplete clause that is completed by the appended satellite which takes the form of a list.

A final example serves to illustrate the use of the fragment label as a container category. Consider (60) and (61):

- (60) (1) <C><indepcl>Slechts één op de drie werknemers die in schadelijk  
 geluid werkt, beschermt zijn oren<indepcl><C>. <s9329, leaflets>  
 (2) <C><fragment>Waarom zo weinig<fragment><C>? <s9330, leaflets>
- (<C><indepcl>Only one in three employees who works in harmful  
 noise, protects his ears.<indepcl><C>.  
 <C><fragment>Why so few<fragment><C>?)
- (61) (1) <C><reportedcl><indepcl>Are you preparing for an exhibition <indepcl>  
 <reportedcl><C>? <D><reportingcl><fragment>asked John as he stared  
 down at one of the unfinished works<fragment><reportingcl><D>.  
 <s10596, short stories>
- (2) <C><reportedcl><fragment\_no>No<fragment\_no><reportedcl><C>,  
 <D><reportedcl><fragment>nothing like that at the moment  
 <fragment><reportedcl><D>, <E><reportingcl><fragment>said  
 Robin<fragment><reportingcl><E>. <s10597, short stories>

Sentence (60.2) presents an example of an elliptical clause, in which the verb and part of the NP are left out, both of which can be recovered from the sentence that precedes it, (60.1) (cf. Biber et al. 1999: 1100 on condensed questions in conversation). Example (60.2) received the label *fragment* because it literally

constitutes a clause fragment. A similar example is presented by (61.2), in which the nuclear SIU, realised as an answer to a question posed in (61.1), is followed by an elliptical clause (cf. Quirk et al. 1985: pp. 884-888 on ellipsis, and Biber et al. 1999: 1099 on elliptic replies in conversation). In this clause the subject and the verb are left implicit, which is why this SIU is classified as a fragment.

### 3.5.2 Discourse markers, vocatives, tags

In line with the category of fragments, the categories of discourse markers, vocatives and tags have all been kept fairly broad in order to classify a wide number of elements that mainly occur in the simulated dialogue parts of the short stories genre. Biber et al. use the umbrella term *inserts* for many of these non-clausal units (1999: 1082-1098). Again, as they aim to provide a fairly detailed account of the grammar of spoken language, they subdivide inserts into further subcategories, such as interjections, greetings and farewells, discourse markers, and so on (but see p. 1083 on gradual boundaries between subcategories of inserts). In the present study, the number of categories has been restricted because it is not a study of spoken language, but rather of written language. Moreover, the number of categories has been restricted as these types of units only occur in one part of one genre.

The first broad category concerns the category of discourse markers, which includes a wide range of non-clausal units that typically consist of one word. Consider the following examples:

- (62) <C><reportedcl><indepcl>He's six years old<indepcl><C>, <D><dm>for Christ's sake<dm><reportedcl><D>. <s10561, short stories>
- (63) <C><dm>Wow<dm><C>. <s10690, short stories>
- (64) <C><reportedcl><dm>Mmmm<dm><reportedcl><C>, <D><reportingcl><fragment>zei Timmer die bleef staan en in de verte tuurde<fragment><reportingcl><D>. <s13807, short stories>
- ( <C><reportedcl><dm>Mmmm<dm><reportedcl><C>, <D><reportingcl><fragment>said Timmer who stood still and looked into the distance<fragment><reportingcl><D>.)
- (65) (1) <C><reportedcl><dm>Gefeliciteerd<dm><reportedcl><C>. <s13845, short stories>

- (2) <C><reportedcl><dm>Dank u<dm><reportedcl><C>. <s13846, short stories>  
 (<C><reportedcl><dm>Congratulations<dm><fragment><reportedcl><C>.  
 <C><reportedcl><dm>Thank you<dm><reportedcl><C>.)

*For Christ's sake* in (62), *wow* in (63), *Mmmm* in (64) and *Gefeliciteerd* (*congratulations*), and *dank u* (*thank you*) in (65) have all been classified as discourse markers in the present study. If the classification of Biber et al. had, for instance, been adopted, *for Christ's sake* would be classified as an expletive (1999: 1094); *wow* as an interjection (p. 1083); *Mmmm* as a response form (p. 1091); and *gefeliciteerd* and *dank u* as polite speech-act formulae (p. 1093). And of course it should be noted that their classification also represents one of various possible ones (1999: 1086). The motivation for grouping this wide range of phrases under the category of discourse marker has been motivated above and is mainly driven by practical purposes. Note, however, that certain units that Biber et al. classify as discourse markers, such as *you know* and *I mean*, are here classified as a particular type of fragment, because they are considered to contain more clausal than phrasal features. Consider (66), in which *I mean* is categorised as a fragment with the addition of comment clause (cf. Quirk et al. 1985: 1114-1115 for a similar categorisation):

- (66) <A><reportedcl><commcl\_fragment>I mean<commcl\_fragment><reportedcl><A>,  
 <C><indepcl>it depends on the woman<indepcl><C>. <s11199, short stories>

The second category concerns the category of vocatives. Compared to the classification of discourse markers, the categorisation of vocatives is much more straightforward, as these simply concern those cases in which someone is addressed explicitly, either by his or her name or some other label. Consider the following examples of vocatives:

- (67) <A><reportedcl><vocative>Frank<vocative><reportedcl><A>,  
 <C><reportedcl><indepcl>don't go<indepcl><reportedcl><C>,  
 <D><reportingcl><fragment>Caro implored<fragment><reportingcl><D>. <s11264, short stories>
- (68) <C><reportedcl><dm>Come on<dm><reportedcl><C>, <D><reportedcl><vocative>little girl<vocative><reportedcl><D> ... <s11389, short stories>

Last, in line with Biber et al. (1999: 1080), who distinguish between seven kinds of tag, the category of tags has been made relatively broad in the present study. Consider the following examples:

- (69) <C><reportedcl><indepcl>I think you'd better go<indepcl><reportedcl><C>, <D><reportedcl><tag>don't you<tag><reportedcl><D>? <s11845, short stories>
- (70) <C><subcl\_advcl\_reas>Because you can't tell how it's done from a distance<subcl\_advcl\_reas><C>, <D><tag>see<tag><D>. <s12068, short stories>
- (71) <C><reportedcl><indepcl>Het is niet meer helemaal pap<indepcl><reportedcl><C>, <D><reportedcl><tag>hè<tag><reportedcl><D>? <s14294, short stories>
- <C><reportedcl><indepcl>It's not all mushy<indepcl><reportedcl><C>, <D><reportedcl><tag>right<tag><reportedcl><D>?)

*Don't you* in (69), *see* in (70) and *hè (right)* in (71) have all been classified as tags in the present study, which in line with Biber et al.'s classification of tags (1999: 1080).

This section focused on a number of linguistic forms that mainly occurred in the simulated dialogue parts of the short stories genre. Both because this study focuses on the analysis of the written language and in order to guarantee consistent annotation, the number of categories into which these various forms fall has been restricted. The additions made to, for instance, the fragment label, the labelling of the markers of direct speech and reporting clauses and the discourse annotation all provide information to be able to sufficiently analyse these parts of the data.

### 3.6 Conclusion

The main aim of this chapter was to discuss and exemplify the main difficulties involved in determining the grammatical classification of Sentence Information Units. As a sentencing analysis involves both segmenting discourse into basic units of analysis and categorising these units grammatically, both steps in the annotation process have been described in great detail, focusing on the decisions that were made and distinction criteria set in order to guarantee consistent annotation. For the sake of clarity, these two levels of analysis have been described in two consecutive chapters, thereby partly reflecting the order of the respective analyses,



as discourse segmentation typically precedes grammatical categorisation. However, it has also been noted on various occasions that the two levels of analysis do not always follow each other, but at times also interact, for instance when syntactic criteria are used to determine unit boundaries and unit status, and vice versa, when discourse status is used to make decisions about grammatical classification. For this reason, these two chapters together form the basis of the model of analysis developed for the sentencing analysis and should be considered in tandem.

The discussion of issues in grammatical classification was divided into three main parts. The first part dealt with issues that can be categorised as constituting more elementary or general problems, in that they precede all further steps in the categorisation process. These concerned determining the status of a unit on the cline of clausiness and deciding whether to classify a unit on the basis of its grammatical form or on its function or semantic role. The second part dealt with categorisation problems that arise when classifying certain grammatical relations consistently, such as the distinction between subordination and coordination, and the gradable status of apposition. The third part focused on categorisation issues that were brought about by particular genres, especially the simulated dialogue parts of the short stories genre. In the description of all these categorisation issues, the purpose has been to pinpoint the exact difficulties and describe how these were dealt with consistently in the analysis of data.

## 4. Methodology

### 4.1 Introduction

Chapter 1 explained what sentencing involves and described the main aim of this study as identifying and analysing sentencing patterns in English and Dutch. Chapters 2 and 3 showed that the identification of such patterns involves an analysis of sentences at both the level of discourse and grammar. The present chapter will describe how exactly the sentencing analysis was carried out, i.e. by using the method of corpus linguistics.

In order to identify and analyse sentencing patterns in English and Dutch, a corpus was compiled, designed and annotated at the two levels of analysis. The reason for using the method of corpus linguistics to carry out this type of analysis is because this makes it possible to identify patterns in large sets of data that can easily be processed and manipulated (cf. McEnery, Xiao, Tono 2006: 6). In addition, this method makes it possible to test hunches based on intuition or introspection by performing a quantitative analysis of large sets of empirical data. Despite the fact that there are a wide range of corpora available, there was no comparable corpus of English and Dutch texts that met the requirements set by the research questions, i.e. that included the same range of genres as the corpus designed for this study and that contained discourse and grammatical annotation at the sentence level.

This chapter will first describe what type of corpus was developed for the present study in order to answer the main research question (4.2). It will then describe how it was designed and composed (4.3), where the focus will first be on general aspects that are of primary concern in designing a corpus, such as achieving representativeness and balance (4.3.1). This is followed by a detailed description of the decisions made at the levels of the four subcorpora of which the corpus consists (4.3.2). As the type of annotation added to the corpus is essential for answering the main research question, one section will be devoted to describing how the corpus was annotated and focus on how reliable and accurate annotation was achieved (4.4). It should be noted that the complete list of discourse and grammatical labels applied to the corpus is provided in the annotation manual in Appendix I, which also contains the exact annotation criteria and decisions,

supported by a wide range of examples taken from the corpus. The chapter will end with a description of the statistical tests that were performed to test the results of the corpus-based sentencing analysis (4.5).

## 4.2 A corpus-based study

The present study can be characterised as a corpus-based study, where corpus-based is used as an umbrella term and no distinction is made between corpus-based and corpus-driven research (cf. McEnery et al. 2006: 8). This means that the method of corpus research has been used in order to identify, analyse and compare sentencing patterns in English and Dutch. The corpus that has been especially designed (see 4.3) and annotated (see 4.4) for this study can be classified as a bilingual comparable corpus. Even though definitions vary as to what constitutes a comparable corpus and how this is to be distinguished from a parallel corpus (or translation corpus), the term comparable is here used to characterise a corpus that consists of two or more monolingual subcorpora of original texts that are not translations of each other and that are designed using the same sampling techniques (cf. McEnery & Wilson 1996: 57; Johansson & Hasselgård 1999: 145; Hunston 2002: 15; McEnery et al. 2006: 47; Altenberg & Granger 2002: 8, and see Granger (2003) for an overview of types of corpora available in cross-linguistic research). Specifically, it means that the subcorpora are collected using ‘the *same proportions* of the texts of the *same genres* in the *same domains* in a range of *different languages* in the *same sampling period*’ (McEnery et al. 2006: 48, italics in original).

This collection of compilation criteria are mainly set to achieve comparability of data, which is one of the most difficult and also one of the most debated issues with respect to comparable corpora, and may even be considered, to use Cosme’s words, a ‘fundamental weakness’ (2007: 152, see Johansson & Hasselgård 1999: 146, Altenberg & Granger 2002: 8 on comparability). How these compilation criteria were taken into consideration in the design of the corpus will be described in the following sections.

## 4.3 Corpus design and composition

In designing a corpus and in distinguishing it from a random selection of texts, a primary concern is that the corpus can be considered representative of the language or language variety under consideration. Determining how representativeness should be defined or how it can be achieved in corpus linguistics is not a straightforward matter and has been described and approached from various angles (cf. Biber 1993; Kennedy 1998: 68; Sinclair 2004b: 7-9; McEnery et al. 2006: 21). One way of qualifying a corpus as representative is when the findings of the analyses carried out on the corpus can be generalised to the language or language variety it is supposed to represent (cf. Leech 1991: 27). Another way of approaching the concept is by defining it as 'the extent to which a sample includes the full range of variability in a population', where the sample constitutes the corpus and the population the language or language variety the corpus is supposed to represent (Biber 1993: 243). In designing a representative corpus, what needs to be taken into account is the range of genres included in a corpus, i.e. its balance, and the way in which the text chunks taken from these genres are selected, i.e. the sampling criteria used (cf. McEnery et al. 2006: 13). The present section will first describe and motivate the range of genres that have been included in the corpus and will then outline the sampling criteria used for each of these genres, such as determining the sampling method, the sampling size and the size of the corpus.

### 4.3.1 General concerns in corpus design

As representativeness depends to a large extent on how balanced a corpus is, i.e. the range of genres included, the choice of what genres to select is of primary concern in corpus design (cf. Biber 1993: 243; McEnery et al. 2006: 16). This decision is motivated by the intended use of the corpus. As the main aim of the present study is to identify and analyse the sentencing patterns in the English and Dutch written language, the corpus designed in order to achieve this aim has to be considered representative of written English and written Dutch. It should be noted that the written language of English has been restricted to published texts that can be classified or categorised as British English and that written Dutch has been restricted to published texts that can be classified or categorised as representing Netherlandic Dutch, and not, for instance, Belgian Dutch. The reason for restricting it to published texts is motivated by one of the main criteria of using punctuation in

discourse segmentation, which is that punctuation has been applied deliberately and consistently (cf. 2.4.2). In addition, the reason for restricting it to one variety within each language is that there may be differences between the various varieties that cannot be taken into account in the present study for practical purposes (cf. Biber et al. 1999 for differences between American and British English).

In deciding what genres can be considered representative of the written language, where the term genre is used in Biber et al.'s sense as 'situationally defined varieties' (1999:4), see also Chapter 1, 1.3.4), inspiration was taken from Biber et al.'s choice of genres in designing a corpus to describe the grammar of spoken and written English (1999: 15-17). That corpus consists of four genres, namely conversation, fiction, news and academic prose. Since the present study has been restricted to describing sentencing patterns in the written language, it includes fiction, news and academic prose, with the addition of the genre of public information leaflets. The decision to include this last genre was motivated by the aim to well represent the continuum of various diverging genres within the written language, by including genres that occupy different positions on the spoken-written continuum (cf. 2.2). In line with Biber et al.'s choice of genres, the present selection is seen as having 'the virtue of being (a) important, highly productive varieties of the language, and (b) different enough from one another to represent a wide range of variation' (Biber et al. 1999: 15-16). Biber et al. provide a detailed characterisation of their genres in terms of diverging 'situational characteristics', such as interactiveness, communicative purpose, audience and dialect domain (1999: 16). Moreover, in selecting the genres for analysis, it should of course be mentioned that the decision to limit this to four is also influenced by practical considerations. As the corpus has been entirely manually annotated (cf. Chapters 2 and 3, and Section 4.4 below), operational feasibility also influenced the design and size of the corpus (cf. Biber 1993: 245).

By using non-linguistic parameters in the selection of texts of the corpus, i.e. in selecting them on the basis of the genre they belong to and not on the basis of their linguistic features, we applied so-called external criteria for text selection as opposed to internal criteria (cf. McEnery et al. 2006: 14). The reason for using external criteria is motivated by the fact that it can be considered circular to use internal criteria, as a corpus is typically designed to study linguistic features. Specifically, '[i]f the distribution of linguistic features is predetermined when the corpus is designed, there is no point in analysing such a corpus to discover naturally occurring linguistic feature distributions' (McEnery et al. 2006: 14, see also Biber 1993: 253 and Sinclair 2004b: 10 for a similar view).

In addition, in seeking representativeness and thus creating a balanced corpus, sampling techniques also play a crucial role. These will first be described in more general terms and outlined in more detail for each of the four genres in the sections below. The sampling method adopted for the present study is that of stratified random sampling (Biber 1993: 244; Oakes 1998: 10). This means that the population is first divided into groups or strata, i.e. the four genres included in this corpus, and that these groups are sampled using random techniques. This is different from simple random sampling, where all texts in the population have an equal chance of being included (Biber 1993: 244; McEnery et al. 2006: 20). According to Biber, the former approach has the advantage of 'guaranteeing that all strata are adequately represented while at the same time selecting a non-biased sample within each stratum' and that stratified samples are 'almost always more representative than non-stratified samples' (1993: 244).

Another important decision in sampling concerns determining sample size. This involves deciding whether to include full texts in the corpus or text chunks. In order to achieve a balanced corpus, we chose to sample text segments instead of full texts for three of the four genres, i.e. academic prose, short stories and public information leaflets. As full texts within these genres typically consist of at least a few thousand words, the combination of limited time available and full manual annotation at both the level of discourse and grammar would lead to a very small selection of different texts if full texts were included. A small corpus that consists of only a few full texts can be considered less balanced precisely because the range of authors is reduced and 'the peculiarity of an individual style or topic may occasionally show through into the generalities' (Sinclair 1991: 19, as quoted in McEnery et al. 2006: 20; see also Aston & Burnard 1998: 22; Kennedy 1998: 68). The main reason for making an exception for the newspaper genre in this respect is that this genre can be characterised by the fact that text lengths can vary considerably and that many articles contain considerably fewer words when compared to the full texts in the three other genres (cf. Kennedy 1998: 75). In order to balance the part of the corpus that contains newspaper articles, the articles included have all been taken from the same section in the newspaper in which the articles roughly have the same length (see 4.3.2 below).

In deciding to use text samples, the size of the sample also has to be determined. This relates to another methodological issue in corpus linguistics (cf. Biber 1993; McEnery et al. 2006: 20). In his discussion of the length of text samples, Biber argues that text segments should be 'long enough to reliably represent the distributions of linguistic features,' which he specifies by explaining that more

common features are reliably represented in relatively short segments of texts and longer segments of texts are needed for a reliable representation of less common features (1993: 252). He does, however, state that 'broader linguistic representation can be achieved by focusing on diversity across text types rather than by focusing on longer samples from within texts' (1993: 252). Following this thought, the present study focused on including a diverse range of text types, not only by including four different genres, but also a range of subgenres within those genres (see 4.3.2 below). As a practical consequence of this decision, text segments had to be kept relatively short and set at 1,000 words.

In addition to determining the length of the text samples, the proportion and number of samples for each genre also have to be decided on, a decision which determines the size of the corpus as a whole. Although these proportions are considered difficult to define objectively (cf. Hunston 2002: 28-30), Biber (1993) makes an attempt and offers a number of recommendations on how corpus size can be determined by taking into account linguistic variation within and across different genres. He argues that ten 2,000- word samples are typically sufficient for each genre, when focusing on the occurrence of common linguistic features (1993: 252-253). It should be noted, however, that corpus size depends primarily on the research question and on the frequency and distribution of the linguistic features or 'objects' that are being studied (Leech 1991: 8-29; McEnery & Wilson 2001: 80; Meyer 2002: 33; Sinclair 2004b: 10; McEnery et al. 2006: 72). For instance, corpora used to study lexical features are typically much larger than corpora used to study grammatical features, because word features are typically described with respect to their frequency and distribution per, for example, 1 million words (cf. McEnery et al. 2006: 72). In addition to the research question or type of linguistic feature under consideration, the type of annotation added to the corpus also influences corpus size. Specifically, 'corpora that need extensive manual annotation are necessarily small' (McEnery et al. 2006: 72). With respect to the corpus designed for the present study, although it can be considered relatively small when expressed in number of words (nearly 300,000, see Table 1 below) and compared to large reference corpora such as the British National Corpus (BNC) (100 million words) or the Bank of English (524 million words), at the same time it could be argued that it contains sufficient instances of the linguistic feature under consideration, namely the sentence. The present corpus contains nearly 16,800 sentences that have all been analysed in great detail at two levels of analysis in order to carry out a sentencing analysis. It may be suggested that, for the current purposes, the size of the corpus should not be

expressed in the total number of words it consists of, as is standard practice in corpus linguistics, but in total number of sentences.

Now consider the basic composition of the corpus, as represented in Table 1 below. The detailed composition per genre, i.e. including the subgenres, will be presented in 4.3.2 below.

**Table 1**                    **Composition of the corpus in terms of number of words and sentences**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Total number of words	41.595	34.182	41.340	29.328	146.445
Total number of sentences	1438	1844	3169	1589	8040
<b>Dutch</b>					
Total number of words	40.973	30.158	41.092	29.171	141.394
Total number of sentences	1740	1754	3154	2060	8708

Table 1 shows that in terms of number of words most genres have been balanced not only with respect to each other within one language, but especially with respect to each other across the two languages.

This section described how representativeness and balance have been aimed at in general terms, by including a range of genres and adopting specific sampling techniques. The following sections will describe what design and compilation decisions were taken at the level of each of the four different genres.

### **4.3.2 Genre-specific design and compilation decisions**

In addition to establishing sampling criteria at a general level when designing a corpus, these also have to be set at the level of each subcorpus to be included in the corpus. Decisions at the level of genre include, for example, what disciplines to focus on within the academic prose subcorpus or what newspapers to include in the newspaper article subcorpus. The sampling criteria that were established at this level and the decisions made will be described and motivated in the sections below.

#### **Academic prose**

The academic subcorpus includes only samples taken from research articles and not from book sections. This decision is mainly motivated by practical purposes, as over



half of the research articles included in the corpus were available in electronic format. The articles that were not digitally available were typed in.

As there are many disciplines to choose from within the academic prose genre, we restricted the selection of texts to two disciplines, namely history and psychology. The choice of these particular disciplines is not random. As this is a contrastive study of English and Dutch and as English can to a large extent be considered the language of academic publication, we wanted to include two disciplines and two journals within each of these disciplines for which a publication in either English or Dutch is still considered equally valuable in terms of the ranking status of the publication.<sup>23</sup> For the English history academic prose section of the corpus, *English Historical Review* (EHR) and *Social History* (JSH) were selected and matched with the Dutch journals *Bijdragen en Mededelingen betreffende de Geschiedenis der Nederlanden* (BMGN) and *Economische en Sociale Geschiedenis* (ESG). These journals are comparable in terms of academic ranking, content and readership. For the English psychology section of the corpus, the *British Journal of Developmental Psychology* (BJDP) and *British Journal of Psychology* (BJP) were included and matched with the Dutch journals *Kind en Adolescent* (KA) and *Nederlands Tijdschrift voor Psychologie* (NLTP).

From each of these four journals, we selected ten 1,000-word samples (see Table 2 below). The decision to include text samples instead of full texts leads to a further decision that involves determining whether to sample text initial, middle or end parts. We decided to sample both text-initial and text-final parts, by taking a 500-word sample from the introductory sections of the journal articles and a 500-word sample from their discussion/conclusion sections. This choice was motivated by the fact that these sections can typically be characterised as containing mostly text and few graphs or tables when compared to, for example, the methods and results sections. However, it should be noted that one of the differences between the two disciplines in the structuring of research articles is that articles written within the discipline of psychology tend to be structured more typically along the lines of the rather standard structure of introduction, methods, results and discussion/conclusion, which applied less to the articles written within the discipline of history. This latter group of articles did, however, include an introduction and discussion/conclusion section, which formed another reason for choosing these sections, as it contributed

---

<sup>23</sup> Leading professors within both the discipline of psychology and history were consulted for advice on what journals to include for the academic prose subcorpus.

to the comparability of samples taken from the two disciplines. Furthermore, all samples were randomly selected from the journals in the sense that they were not matched by subject. All articles were matched for sampling period. Specifically, all history articles were published between 2000 and 2003 and all psychology articles were exactly lined up and were all published between 2000 and 2002. However, it should be noted that the publication date does not necessarily match the time at which the articles were written, as the reviewing and publication process can be stretched over a period of time, depending on the procedure followed by each individual journal. One further aspect that was taken into account was establishing the nationality and location of all the authors of the articles in an attempt to only include those authors who were either native speakers of English or Dutch respectively. Whenever possible, this was checked by a quick scan of a website or webpage of the respective authors. However, in some cases these searches resulted in finding only information about their affiliation, and not necessarily about their language background. It should be noted that in designing their corpus and in trying to take register variation into account, Biber et al. characterise the academic prose genre as a 'global' genre, in that the language is influenced 'by authors, editors, and publishing houses often located on different continents, with an eye to an international readership' (1999: 26). Bearing this in mind, it could be argued that restricting the academic subcorpus to texts written by native speakers of one variety within English and Dutch respectively may therefore not constitute a necessary or attainable aim.

The following table provides an overview of the exact make-up of the academic prose subcorpus.

**Table 2** Composition of the academic prose subcorpus

English academic prose	History		Psychology		Total
	EHR	JSH	BJDP	BJP	
Total number of text samples	10	10	10	10	40
Total number of words	10,300	10,655	10,221	10,419	41,595
Total number of sentences	334	401	339	364	1438
Dutch academic prose	History		Psychology		Total
	BMGN	ESG	KA	NLTP	
Total number of text samples	10	10	10	10	40
Total number of words	10,242	10,332	10,207	10,192	40,973
Total number of sentences	455	441	412	432	1740

### **Newspaper articles**

The newspaper articles included in the relevant subcorpus were all retrieved using the program Lexis Nexis, which provides current information via online applications and contains articles from all sections from a wide range of newspapers in digital format. This made it possible to copy all newspaper texts into the relevant subcorpus, instead of typing or scanning them in.

As there are many different types of newspapers available in both languages, a selection was made of two national newspapers per language. For English, these were *The Guardian* and the *Daily Mirror*. These represent different readership levels, as *The Guardian* can be classified as a broadsheet, or 'highbrow', newspaper and the *Daily Mirror* as a popular, tabloid newspaper (cf. Biber et al. 1999: 31). In an attempt to match these with Dutch newspapers, this selection consisted of *de Volkskrant*, which may be classified as a broadsheet, and *de Telegraaf*, which is typically seen as a popular newspaper. However, it should be noted that the difference in readership and style between the Dutch newspapers is not quite as pronounced as between the English newspapers. Furthermore, the newspaper subcorpus consists of two types of newspaper writing, namely news reportage and editorials. The news reportage articles were all taken from the domestic section of the newspapers and thus typically cover national or regional news. The reason for restricting it to national news is that articles covering international news are sometimes based on an international source and may have been translated into the respective language of the newspaper. By restricting it to national news reportage, an attempt was made at avoiding the risk of including translated texts, as this may influence sentencing patterns.

As was explained above, the newspaper articles subcorpus is the only part of the corpus that consists of full texts instead of text samples. The decision to include full texts was motivated by the fact that newspaper articles can vary considerably in length and often do not run up to the 1,000-word sample size set for the present study (cf. Kennedy 1998: 75). To keep the design and composition of this section of the corpus comparable to the other subcorpora, the total number of words of this subcorpus was matched to that of the others as far as possible. However, although the newspaper subcorpus contains fewer words than the academic prose and short stories subcorpora (see Table 1 above), it roughly equals the number of words of the public information leaflets corpus. As there were some differences between the different newspapers in terms of the length of articles, the total number of words per section was equalled by including more texts in cases where the articles were relatively short, such as with news reportage articles

covering national news in the *Daily Mirror* (see Table 3 below). This means that the subsections within this subcorpus were matched on the basis of the total number of words and not on the total number of texts included. Moreover, besides matching the news reportage articles between the languages on the basis of the sections they belonged to, they were also roughly matched by subject area, namely politics. This was achieved by selecting the articles on the basis of a keyword search, which was possible within the Lexis Nexis program. The keywords used for the news reportage articles were *minister* (MP), *Den Haag* (The Hague, the political capital of the Netherlands) for the Dutch newspapers and *MP* and *politics* for the English newspapers. The articles were quickly scanned before inclusion to check that they could indeed be classified as covering domestic politics. The editorials included in the subcorpus were not selected on the basis of their subject, but on the basis of their average length. Finally, the newspaper articles included in this subcorpus were all published in the first six months of 2003.

Table 3 below presents an overview of the exact design and composition of the newspaper articles subcorpus.

**Table 3**                      **Composition of newspaper articles subcorpus**

English newspaper articles	News reportage		Editorial		Total
	Guardian	Daily Mirror	Guardian	Daily Mirror	
Total number of texts	22	30	22	30	104
Total number of words	7766	9172	9425	7819	34,182
Total number of sentences	345	536	460	503	1844
Dutch newspaper articles	News reportage		Editorial		Total
	Volkskrant	Telegraaf	Volkskrant	Telegraaf	
Total number of texts	22	20	22	35	99
Total number of words	7279	6412	8937	7530	30,158
Total number of sentences	460	348	484	462	1754

### Short stories

All short stories selected for the short stories subcorpus were taken from published books containing collections of short stories. This means that none of the short stories included were available in digital format and all text samples were therefore typed in.

As short stories are typically included in a wide range of collections, the source of the texts did not form a sampling criterion. With the aim of designing and composing a comparable subcorpus, a number of sampling criteria were established.

First, all authors had to be native speakers of English or Dutch respectively and had to be born after 1940. Specifically, all Dutch authors had to be born and raised in the Netherlands and all British authors had to be born and raised in Britain. This information was controlled for by performing a check of the authors' backgrounds. Second, all short stories had to be published after 1990. This particular criterion was established to not only achieve comparability of texts within this genre, but also with respect to the time at which the texts were written in the other genres. Third, the short stories had to be between 12 and 20 pages in length, which was to exclude all particularly short or long short stories from the analyses, thereby trying to achieve a comparable text sample to entire text ratio. Fourth, all short stories had to be written for an adult audience. Fifth and finally, all short stories included were written by different authors, 20 of whom are male and 20 female. An attempt was made to include both relatively well-known or established authors and ones that are more new to the field. Examples of Dutch authors include Joost Zwagerman, Arnold Grunberg, Mensje van Keulen and Linda Polman, and examples of English authors include Julian Barnes, Stephan Baxter, Helen Dunmore and Julie Burchill.<sup>24</sup>

From each short story, a 1,000-word sample was taken. Specifically, the subcorpus consists of 20 x 1,000-word samples written by male authors and 20 x 1,000 word samples written by female authors. All samples were taken from the middle parts of the short stories, leaving out the introductory and concluding sections.

Table 4 gives an overview of the composition of the short stories subcorpus.

**Table 4**                      **Composition of the short stories subcorpus**

<b>English short stories</b>	<b>Male authors</b>	<b>Female authors</b>	<b>Total</b>
Total number of text samples	20	20	40
Total number of words	20,745	20,595	41,340
Total number of sentences	1418	1751	3169
<b>Dutch short stories</b>			
Total number of text samples	20	20	40
Total number of words	19,491	21,601	41,092
Total number of sentences	1431	1723	3154

---

<sup>24</sup> See Appendix II for an overview of the short stories included in the short stories subcorpus and their respective authors.

### Public information leaflets

The public information leaflets selected for the public information leaflets subcorpus were all available as printed leaflets and as digital texts, i.e. as PDF files of the printed editions, on the websites of the organisations that publish them (see below). Because of the fact that they were also available in digital format, it was possible to copy the text samples taken from these leaflets into the corpus, instead of typing or scanning them in.

As there are a wide range of public information leaflets available on various topics, a selection was made for those leaflets that could be characterised as dealing with various health issues. All leaflets were taken from the websites of either British or Dutch national charities or governmental departments and organisations that deal with a wide range of health issues. Examples of British charities or departments are the *Health and Safety Executive*, the *Department of Health*, the *National Health Service* and the *British Liver Trust*. Examples of Dutch charities and organisations are *Postbus 51* (Dutch government information service), *Voedingscentrum* (nutrition centre), *Nationaal Instituut voor Gezondheidsbevordering en Ziektepreventies* (national institute for stimulation of health and illness prevention) and *Nationaal Epilepsie Fonds* (national epilepsy foundation).<sup>25</sup> A criterion for the inclusion of a leaflet in the corpus was that it had to exist, be available and circulated in a printed version. The motivation for this is that the corpus was restricted to including published texts only, which means that texts that only exist in digital format were not included. Moreover, the leaflets selected were all aimed at the general public and at an adult readership

From each leaflet, we took a 1,000-word sample and selected 30 x 1,000-word samples per language. All samples comprise the first 1,000 words of the leaflets, as this part usually describes the problem, states the characteristics of the problem and provides possible solutions to the problem, which may be seen as constituting the core information. Moreover, 27 of the 30 text samples were exactly matched by subject, such as leaflets dealing with epilepsy, skin cancer or kidney donors in both languages. Three text samples were more loosely matched, in the sense that they all deal with health issues, but do not necessarily cover the

---

<sup>25</sup> See Appendix II for an overview of the public information leaflets included in the public information leaflets subcorpus and their respective sources.

exact same topic.<sup>26</sup> It should be noted that the leaflets were not selected or matched by publication date. The reason for this is that this was not practically feasible. All information leaflets were selected in the first six months of 2004, and were thus online available at that time, and we always selected the most updated and recent publication of the leaflet.

Table 5 presents the make-up of the public information leaflets subcorpus.

**Table 5**                      **Composition of the public information leaflets subcorpus**

<b>English public information leaflets</b>	<b>Total</b>
Total number of text samples	30
Total number of words	29.328
Total number of sentences	1589
<b>Dutch public information leaflets</b>	
Total number of text samples	30
Total number of words	29.171
Total number of sentences	2060

## 4.4 Corpus annotation

Leech defines corpus annotation as ‘the practice of adding interpretative linguistic information to a corpus’ (2004: 17). Whether or not the addition of this type of information to a corpus should be seen as added value remains a point of discussion (cf. Leech 1997, 2004; Hunston 2002; Sinclair 2004a; McEnery et al. 2006). Whereas some consider annotation an enrichment of the raw text (e.g. Leech 1997, 2004; McEnery 2006), making it possible to easily extract linguistic information, others approach annotated corpora with more caution and may prefer unannotated or raw texts, because these, for instance, lack the interpretation or view of a particular annotator (eg. Sinclair 2004a). Without going into this discussion in any more detail, which is in fact quite balanced and not simply black and white, it should be noted that the annotation that was added to the corpus compiled for this study is seen as giving it added value, making it possible to answer the particular research questions posed in this study. Specifically, to be able to identify the sentencing patterns in English and Dutch, tags

---

<sup>26</sup> Appendix II also provides information about the topics of the information leaflets and how these were matched.

were added at the level of discourse, indicating of what type and how many Sentence Information Units (SIUs) each sentence consists, and at the level of grammar, indicating the grammatical form of the SIUs. The exact practice of identifying SIUs and determining and classifying their grammatical realisation has been described extensively in Chapters 2 and 3. The present section will be restricted to describing the annotation procedure in more detail and to explaining how accuracy and reliability of annotation practice was achieved.

#### 4.4.1 Annotation practice

In order to annotate the discourse and grammatical structure of sentences, four sets of tags were developed, two of which constitute the main ones, i.e. the discourse tags and grammatical tags. In addition to these, there is a small set of lexical tags and a set of tags to label all punctuation marks. With the exception of the punctuation tags, all labels were manually added to the corpus. The entire tagset was developed on the basis of a small pilot study that involved analysing texts of different genres at the level of discourse and grammar. The annotation was carried out by three different annotators using Note Tab light, a free text and HTML editor. This section will provide more information about the make-up and type of tags used and will describe how the annotation procedure was carried out.

With respect to the annotation at the level of discourse, all Sentence Information Unit or Units (SIUs) within each sentence received tags that provide information about their hierarchical status and position (cf. see 2.4 & 2.5 and the annotation manual in Appendix I for overview and description of types of SIU). Specifically, the units with nuclear status received either the label <C> or <Ca,b,c, etc>, the latter of which indicates that the nuclear unit consists of two or more units that are coordinated with each other. As for the units with satellite status, a distinction was made between units that precede the nucleus (<A>, <B>, <-Z>)<sup>27</sup> and the units that follow the nucleus (<D>, <E>, <F>, etc). All satellites that interrupt other discourse units, i.e. both nuclei and prepended or appended satellites, received the label <1> to mark the first interruption, <2> to mark the second interruption, and so on. In addition to indicating their status, interpolated satellites also received labels that indicate their position with respect to the finite

---

<sup>27</sup> See 2.5.3 and annotation manual in Appendix I for a special type of prepended satellite, <ZZ>.



verb of the SIU they interrupt. All SIUs were surrounded by an opening tag and a closing tag, the latter of which contains a forward slash (</>). In order to provide the SIU labels with a unique label and distinguish these from the grammatical labels, all SIU tags start with the letters <du\_>, which stands for *discourse unit*. Consider sentence (1):

- (1) <du\_A>If they do</du\_A>, <du\_C>it may be to highlight Sir Philip's reported conclusion that in the managerial melee two staffers in his private office - <du\_1\_embcl\_prefv>Christine Watson and Annabelle Eyre</du\_1\_embcl\_prefv> - were paid out of his parliamentary allowance as an MP</du\_C>, <du\_D>when they should have been paid from Conservative party funds</du\_D>. <s1603, newspaper articles>

Sentence (1) consists of an appended SIU (<A>), a core SIU (<C>) that is interrupted by an interpolated satellite (<1>) and followed by an appended satellite (<D>). In this particular example, the tag that surrounds the interpolated satellite indicates that this occurs in an embedded clause (<\_embcl>), i.e. the *that*-clause, and that it precedes the finite verb of this embedded clause, *were* (<\_prefv>). When taking out the words in this sentence, the sequence of remaining tags provides insight into its discourse structure:

- (2) <du\_A> ...</du\_A> <du\_C> <du\_1\_embcl\_prefv> </du\_C> <du\_D> ... </du\_D>. <s1603, newspaper articles>

This sequence shows that the core unit is interrupted by an interpolated satellite and both preceded and followed by a prepended and appended satellite respectively.

With respect to the annotation labels used at the level of grammar, a full list is provided in the annotation manual in Appendix I (and see Chapter 3 for a detailed description of the grammatical categorisation of SIUs). To exemplify a few grammatical labels and their basic structure, consider sentence (1) again, repeated as (3) below, to which the grammatical labels are now added:

- (3) <du\_A><g\_subcl\_advcl\_cond>If they do</g\_subcl\_advcl\_cond></du\_A>, <du\_C><g\_indepcl>it may be to highlight Sir Philip's reported conclusion that in the managerial melee two staffers in his private office - <du\_1\_embcl\_prefv><g\_appos\_NP>Christine Watson and Annabelle Eyre</g\_appos\_NP></du\_1\_embcl\_prefv> - were paid out of his parliamentary allowance as an MP</g\_indepcl></du\_C>, <du\_D><g\_subcl\_advcl\_conces>when

they should have been paid from Conservative party  
funds</g\_subcl\_advcl\_conces></du\_D>. <s1603, newspaper articles>

Example (3) shows that the grammatical labels follow the discourse labels, as they provide information about the grammatical realisation of the SIUs. All grammatical labels start with <g\_>, which stands for *grammatical category*. In order to keep the tags as short as possible, but still intelligible (cf. Leech 2004: 22), the names of the grammatical categories were abbreviated, such as <\_subcl>, which stands for *subordinate clause*. The structure of the information contained in the tags goes from general to specific, by first providing grammatical information at the most general level of categorisation, followed by more specific information about types, subtypes and classes. This can be illustrated by the grammatical label <g\_subcl\_advcl\_cond>. The abbreviation <\_subcl> categorises this SIU as a *subordinate clause*; the abbreviation <\_advcl>, which stands for *adverbial clause*, provides information about the type of subordinate clause, and the abbreviation <\_cond> provides further information about the semantic class of the adverbial clause, namely *condition*. When leaving out the words again, the remaining tags now provide insight into the discourse-grammatical structure of the sentence, as exemplified by (4) below:

(4) <du\_A><g\_subcl\_advcl\_cond>... </g\_subcl\_advcl\_cond></du\_A> <du\_C><g\_indepcl>  
<du\_1\_embcl\_prefv><g\_appos\_NP></g\_appos\_NP></du\_1\_embcl\_prefv> -  
</g\_indepcl></du\_C> <du\_D><g\_subcl\_advcl\_conces>...</g\_subcl\_advcl\_conces></du\_D>.  
<s1603, newspaper articles>

Specifically, these tags indicate that the prepended satellite <A> is realised as an adverbial clause of condition; that the nucleus <C> is realised as an independent clause, which is interrupted by an interpolated satellite that takes the form of an appositional noun phrase, and that the nucleus is followed by an appended clause that takes the form of an adverbial clause of concession.

Although the main annotation task involved applying tags at the level of discourse and grammar, a small selection of words were annotated at a lexical level. These concern words that can be categorised as coordinators, subordinators or correlatives that occur at the start of a new SIU and were all annotated manually.<sup>28</sup>

---

<sup>28</sup> See Appendix I for the annotation manual, which lists the coordinators, subordinators and correlatives that have been annotated at the lexical level in this study.

All lexical labels start with the letter 'l', as in <l\_>, for *lexical*. Consider sentence (1) once more, now repeated as (5):

- (5) <du\_A><g\_subcl\_advcl\_cond><l\_subordinator>If</l\_subordinator> they do</g\_subcl\_advcl\_cond></du\_A>, <du\_C><g\_indepcl>it may be to highlight Sir Philip's reported conclusion that in the managerial melee two staffers in his private office - <du\_1\_embcl\_prevf><g\_appos\_NP>Christine Watson and Annabelle Eyre</g\_appos\_NP></du\_1\_embcl\_prevf> - were paid out of his parliamentary allowance as an MP</g\_indepcl></du\_C>, <du\_D><g\_subcl\_advcl\_conces><l\_subordinator>when</l\_subordinator> they should have been paid from Conservative party funds</g\_subcl\_advcl\_conces></du\_D>. <s1603, newspaper articles>

A fourth and final level of annotation concerned the application of tags to label all punctuation marks used in each sentence. In principle, all punctuation marks were annotated automatically with the help of a programmer who used the programming language Python to carry out this task. However, a small number of punctuation marks that were not considered to function as indicators of SIU boundaries received a manually added annotation label (cf. 2.4 on role of punctuation in marking SIU boundaries and annotation manual in Appendix I for overview of these manually applied punctuation labels). This subset of punctuation labels was applied simultaneously with the discourse labels. As the automatic annotation of punctuation marks occurred after the manual annotation process was completed, it was simple to skip all the punctuation marks that were already provided with a manually added tag. The labels applied to indicate the particular use and function of those punctuation marks that do not mark SIU boundaries are exemplified by the following sentences:

- (5) Having sat down with people who have lost children <pu\_comma\_serial>,</pu\_comma\_serial> mothers and fathers to suicide bombers and to military action in the West Bank I am really appalled by what Jenny Tonge has said. <s1692, newspaper articles>
- (6) This leaflet is about: - the health effects of exposure to loud noise <pu\_semicolon\_serial>;</pu\_semicolon\_serial> - your legal duties to protect the hearing of your workers <pu\_semicolon\_serial>;</pu\_semicolon\_serial> - how to assess and control noise <pu\_semicolon\_serial>;</pu\_semicolon\_serial> - how to choose quieter equipment and machinery <pu\_semicolon\_serial>;</pu\_semicolon\_serial> - different methods of hearing

protection <pu\_semicolon\_serial>;</pu\_semicolon\_serial> - health surveillance.  
<s6783, leaflets>

- (7) Wie Erkel hebben ontvoerd <pu\_comma\_serial>,</pu\_comma\_serial> wat hun oogmerken waren en waar ze hem vasthielden  
<pu\_comma\_NL>,</pu\_comma\_NL> is altijd onduidelijk gebleven. <s2844, newspaper articles>

(Who kidnapped Erkel <pu\_comma\_serial>,</pu\_comma\_serial> what were their intentions and where they held him <pu\_comma\_NL>,</pu\_comma\_NL>, has always remained unclear.)

- (8) Wel benadrukt het ministerie dat hoe de zaak ook zal lopen  
<pu\_comma\_A\_embedded>,</pu\_comma\_A\_embedded> deze de belastingbetaler geen geld gaat kosten. <s2610, newspaper articles>

(The ministry does emphasise that however the case will proceed  
<pu\_comma\_A\_embedded>,</pu\_comma\_A\_embedded>, it will not cost the taxpayer any money.)

- (9) <du\_Ca><g\_coord\_a>This new vaccine is given to babies when they are two  
<pu\_comma\_serial>,</pu\_comma\_serial> three and four months old</g\_coord\_a></du\_Ca>,<du\_A><g\_pp><l\_coordinator>but</l\_coordinator> unlike the previous vaccine</g\_pp></du\_A> <pu\_comma\_A\_satC>,</pu\_comma\_A\_satC> <du\_Cb><g\_coord\_b>the polio part is given in the same injection rather than by mouth</g\_coord\_b></du\_Cb>. <s7707, leaflets>

Sentences (5) and (6) contain cases in which the comma and semi-colon are used serially to separate the items on a list. As these particular uses of these punctuation marks are not considered to mark SIU boundaries, they received tags that were applied manually. In addition to a serial use of the comma, sentence (7) also contains an example of a use of the comma that is particularly frequent in Dutch, hence the name *comma\_NL*. This comma is used to separate an often complex or long subject from the finite verb, as in (7), or to separate long and complex objects from the finite verb. Although this is much more frequent in Dutch, in those cases where a comma has the same function in an English sentence, it also received the label <comma\_NL>. Sentence (8) contains an example of a situation in which a comma is inserted in a complex embedded clause to separate a modification that is placed at the beginning of the embedded clause from the rest of the clause. This use of the comma received the label <comma\_A\_embedded> to indicate that the embedded clause is preceded by an prepended satellite-like element that is not

quite presented as a separate punctuation unit, as it is followed by only one comma and not surrounded by two. By providing these commas with a special annotation label, insight can be gained into the structure of such sentences, making it possible to retrieve them for analysis at a later stage. Finally, sentence (9) contains an example of a use of the comma that has the function of separating a prepended satellite from the second coordinate of the coordinated nucleus Cb. This use of the comma received the label <comma\_A\_satellite\_C>. Besides these particular uses of these punctuation marks, especially the comma, all other uses and occurrences of punctuation marks were annotated automatically.

The order in which the tags were applied at the different levels of annotation was determined by the order in which the sentencing analysis was carried out. This means that sentences were first analysed and annotated at the level of discourse to identify SIUs and this was followed by the application of the grammatical tags to classify the grammatical realisation of SIUs. Specifically, all texts and text samples were first fully annotated at the level of discourse with the inclusion of those punctuation marks that needed to be annotated manually. After the annotation at the level of discourse was completed, the full text was annotated at the level of grammar. These different steps in the annotation procedure were deliberately separated after a small pilot study in which the annotation system was put to practice had shown that separating the steps not only kept the annotation procedure more transparent, by making the annotated text more legible, but also speeded up the annotation process considerably. Furthermore, to physically separate these annotation steps, each original text was copied three times, with one copy receiving only discourse tags, one receiving only grammatical tags and a third receiving all punctuation tags. After all texts in the corpus had been fully annotated at the different levels, the different copies of the same text were automatically merged into one document. As already became clear from the sequence of tags produced in example (4) above, in merging these different documents into one, all grammatical tags were programmed to directly follow the discourse tags. Sentence (10) below gives an example of a sentence in which the three different levels of annotation have been merged and example (11) contains the sequence of tags that is produced after the text is taken out of example (10).

(10) <du\_C><g\_indepcl>The tradition prospers because of our innate  
curiosity</g\_indepcl></du\_C> <pu\_comma>,</pu\_comma> <du\_D><g\_nonrestr\_relcl>to  
which formal education and professional teachers are not always  
kind</g\_nonrestr\_relcl></du\_D> <pu\_period>.</pu\_period> </s1078, newspaper articles>

- (11) `<du_C><g_indepcl> ... </g_indepcl></du_C> <pu_comma>,</pu_comma>  
<du_D><g_nonrestr_relcl> ... </g_nonrestr_relcl></du_D> <pu_period>.</pu_period>  
</s1078, newspaper articles>`

After all texts were merged, they were placed in one document containing the entire annotated corpus. To this document a number of standard codes were added automatically, providing additional information about, for example, the source, date and author of a text. This type of information is typically referred to as corpus mark-up (cf. McEnery et al. 2006) and although no official mark-up scheme was used for this corpus, such as the Text Encoding Initiative (TEI) or Corpus Encoding Standard (CES), the information was added automatically, consistently and in a detailed fashion (cf. McEnery et al. 2006: 22). The following example provides information about the corpus mark-up added to the corpus:

- (12) `<text no="1" tn="001" lang="en" genre="na" subgenre="ed" krant="dm" volgnr="01"  
name="na_dm_ed_en_01">  
<p>  
<header>VOICE OF THE DAILY MIRROR <pu_colon>:</pu_colon> HOON MUST  
PROBE RED CAP KILLINGS</header>  
</p>  
<p>  
<s n="1"><du_C><g_indepcl>REFUSING to hold a proper inquiry into the killing  
of six Red Caps in Iraq is an insult to the memory of those brave  
men</g_indepcl></du_C> <pu_period> . </pu_period> </s>  
</p>`

This type of standard coding contained in the annotation labels in example (12) was added to every text or text sample included in the corpus. The coding provides information about the text number (<text no>), the language (<en>, English), the genre (<na>, newspaper articles), the subgenre (<editorial>) if present, the source (<dm>, Daily Mirror), and the full name given to each merged text: <na\_dm\_ed\_en\_01>. In addition to this, all headers, titles, subtitles and paragraph breaks were also labelled automatically, and every sentence in the corpus received a label containing its unique number, such as <s n = '1'>.

Besides existing as one large text file containing all merged annotated texts, the full corpus also exists in table format that only contains the information contained in the tags. Having all this information contained in a table format makes it possible to read it into a statistical program such as SPSS, which has been used for the present study, and perform frequency counts and statistical analyses with the data (cf. 4.5 below on the statistical analyses performed in the present study).

This section described the make-up of the different tags added to the corpus and the order in which these were applied. The following section will describe how and by whom these were applied and how we have attempted to achieve consistent and reliable annotation.

#### **4.4.2 Accuracy and reliability of annotation**

One very important criterion to take into account when adding linguistic annotation to a text is that the annotation can be considered of good quality, which means that it has to be applied accurately and consistently (cf. Leech 2004: 27-28). This section will first give some information about the annotators of the corpus and then list the steps and measures that have been taken to achieve consistent and accurate annotation.

The manual annotation of the corpus was carried out by three annotators, one of whom was the author of the present study, who also developed and designed the annotation system, and the two other annotators were two university students who were hired on a temporary basis and paid an hourly rate to carry out the annotation. These were a bachelor student of English Language and Culture and a master student of Dutch Language and Culture respectively.

Each of the three annotators was responsible for a separate part of the corpus and annotated this individually. Specifically, the student of English annotated a considerable part of the English part of the corpus, the student of Dutch annotated a considerable part of the Dutch part of the corpus and the third annotator annotated the remaining English and Dutch texts that were not annotated by the students. The reason for deciding to hire a student of English for the annotation of the English corpus and a student of Dutch for the annotation of the Dutch corpus is that these students were trained in their studies to identify and label grammatical categories in the language they studied. However, as all grammatical categories received English tags (cf. Chapter 3, 3.2), the student of Dutch received additional guidance at the start of the annotation process to become familiar with the English labels.

One of the main measures taken to achieve consistent and accurate annotation was the development of an annotation manual (cf. Appendix 1, and Leech 2004: 24 on the need for annotation manual), which the annotators could consult during the annotation process. This manual explains the annotation scheme in great detail, listing the full tagset and giving the annotation rules and guidelines on how and when what label should be applied. It also contains a large number of

examples of how the labels were applied to annotate the discourse and grammatical structure of both English and Dutch texts in various genres. The annotation manual was constantly updated during the annotation process by formulating the guidelines more accurately and by providing more wide-ranging examples taken from different genres.

In addition to the development and constant consultation of the annotation manual, another measure taken to achieve consistent and accurate annotation involved the organisation of weekly annotator meetings in which all annotation problems and questions were discussed extensively and dealt with one by one. In preparation for these meetings, the annotators sent in their questions, provided with a number of examples. Questions and problems were thus dealt with on a weekly basis, and on the basis of decisions made and issues revolved during these meetings, the annotation manual was also updated on a weekly basis. As problematic cases in the texts that had led to questions were marked with a dollar sign (\$), these could be retrieved easily after the meetings and changed accordingly.

When the official annotation process was completed after two and a half years, an additional number of checks were carried out on the corpus. The most extensive and thorough check involved a full check of all the tags added to one of the four subcorpora, namely the newspaper articles subcorpus, at all levels of annotation. Both on the basis of this check and on the basis of the questions collected during the weekly annotator meetings, a list was drawn up of the tags that had posed most problems and uncertainty during the annotation process. These problematic tags were all individually checked and corrected if necessary in each of the four subcorpora with the program Wingrep, which makes it possible to search files using regular expressions and perform global replacements. By carrying out this post-editing step, certain possibly inconsistent decisions and developments in the annotation system caused by the fact that the annotation procedure was spread out over a number of years were thus controlled for.

This section described the measures taken in order to achieve reliable and consistent annotation. Even though it may not be possible to achieve 100% accuracy (cf. Leech 2004: 27), by developing an annotation manual, by organising weekly annotator meetings and by performing thorough checks on the annotated texts an attempt was made to strive for a high accuracy and consistency rate.



## 4.5 Statistical analysis

The core data reported in the results chapters (Chapters 5, 6, 7 and 8) consist of frequency counts of sentencing phenomena in the two languages and in the four genres. These frequency counts are presented in tables and summarised in the form of absolute frequencies and percentages. In those cases where the column totals within the tables are below 30, the tables list only the absolute frequencies and not the corresponding percentages. An exception to this is formed by the totals column, which always presents the percentages of all absolute frequencies.

To compare the frequencies of sentencing patterns and study language x genre interactions, log-linear analyses were performed (cf. Oakes 1998; Field 2005; Tabachnick & Fidell 2007 for more information on log-linear analysis). In those cases where the loglinear analyses detected significant three-way interactions, subsequent chi-square tests were performed to further examine the interaction effects. Standardised residuals were then studied to identify the cells contributing to significant chi squares. For all reported chi squares in the results chapters effect size estimates were determined with Cramer's V.

As corpus data often contain large frequencies and as a chi square analysis of such large frequencies may too easily lead to significant results, the alpha level was set at .01 instead of the more conventional .05. This was done to reduce the possibility of a Type I error.

## 4.6 Conclusion

The main aim of this chapter was to provide insight into how the method of corpus linguistics has been used to design and annotate the corpus compiled for this study. It has presented a detailed description of the composition of the corpus and of the annotation procedure. With respect to its composition, the discussion focused on what the design of a balanced corpus involves and on explaining how representativeness has been achieved. With respect to the annotation procedure, the relevant section explained that accuracy and reliability constitute key notions in corpus annotation and described how these have been taken into account in the annotation of the corpus designed for this study. The chapter has also referred to the annotation manual provided in Appendix I, which contains more detailed information about tags used and annotation decisions made.

The following chapters will present the main findings of the sentencing analysis. Specifically, Chapter 5 constitutes the general results chapter and presents an overview of the most frequent sentencing patterns in English and Dutch in the four different genres. Chapters 6, 7 and 8 then take a closer look at the beginnings of sentences, the use of punctuation marks in signalling discourse relations, and the subpatterns formed by interpolated satellites respectively.



# 5. Sentencing patterns in English and Dutch

## 5.1 Introduction

As it is the main aim of the present study to uncover what the most frequent sentence patterns are in English and Dutch, to compare and contrast them and to establish to what extent genre influences sentence structure, a corpus that consists of four different genres was designed (cf. Chapter 4) and annotated at both the level of discourse (cf. Chapter 2) and grammar (cf. Chapter 3). This chapter will present the main results of an analysis of the main sentencing patterns in English and Dutch in four different genres.

The chapter starts by providing some basic information about the corpus and about average sentence length in both languages (5.2). It will then go on to present the main sentence patterns in English and Dutch (5.3). This will be followed by sections that present the most frequently occurring subpatterns of the main sentence patterns, which will be described and analysed in detail (5.4 to 5.7). In each of these sections, an analysis will be made of both the discourse structure of each sentence pattern and the grammatical realisation of each of the Sentence Information Units (SIUs) in these patterns. Section 5.8 will present a detailed summary of the main findings of the analysis, first presenting the main differences between the languages at the level of genre, then presenting the main differences between the languages irrespective of genre, and ending with an overview of structural similarities between the languages.

## 5.2 General information on the corpus and sentences

The corpus developed for this study consists of English and Dutch texts that represent four different genres: academic prose, newspaper articles, short stories and leaflets (cf. 4.3 on design and composition of the corpus). This section will provide information about the make-up of the corpus in terms of number of words and sentences per genre per language, and about differences in sentence length.

Table 1 provides the total number of words per genre per language and the total number of sentences per genre per language.

**Table 1** Text statistics corpus

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Total number of words	41.595	34.182	41.340	29.328	146.445
Total number of sentences	1438	1844	3169	1589	8040
<b>Dutch</b>					
Total number of words	40.963	30.158	41.092	29.171	141.384
Total number of sentences	1740	1754	3154	2060	8708

Of the various methods available, the one adopted here for measuring sentence length was banded sentence length (cf. Hannay 1997: 243; Cosme 2007: 259), in which the length band was arbitrarily set to 10 words. The following four categories were created: sentences of 1-10 words, 11-20 words, 21-30 words and 31 words and above. The frequencies of the sentences that fall into each of these categories are presented in Table 2 below. In interpreting sentence length, it should be noted that there is a difference in writing practice of compounds in English and Dutch that has not been taken into account in the present study. Specifically, in Dutch compounds are typically written as one word, whereas in English these are typically written as two separate words. A consequence of this difference in writing practice could be that English sentences are longer than Dutch sentences. However, it should also be noted that if the writing practice of compounds in English causes English sentences to be longer than Dutch sentences, it means that English sentences should be longer across all four genres, assuming that the occurrence and frequency of compounds themselves are not dependent on genre. Table 2 below shows that English sentences are indeed longer than Dutch sentences, but that this does not apply to all four genres. For this reason it may be possible that differences in sentence length between the languages are due to something other than the writing practice of compounds.

**Table 2** Sentence length in four categories

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
1- 10 words	74 (5.1%)	400 (21.7%)	1670 (52.7%)	355 (22.3%)	2499 (31.0%)
11-20 words	361 (25.1%)	749 (40.6%)	902 (28.5%)	709 (44.6%)	2721 (33.8%)
21-30 words	430 (30.0%)	480 (26.0%)	371 (11.7%)	362 (22.8%)	1643 (20.4%)
31 – above	573 (39.8%)	215 (11.7%)	226 (7.1%)	163 (10.3%)	1177 (14.6%)
Total	1438 (100%)	1844 (100%)	3169 (100%)	1589 (100%)	8040 (100%)
<b>Dutch</b>					
1- 10 words	184 (10.6%)	374 (21.3%)	1647 (52.2%)	710 (34.5%)	2915 (33.5%)
11-20 words	614 (35.3%)	875 (50.0%)	946 (30.0%)	1043 (50.6%)	3478 (39.9%)
21-30 words	543 (31.2%)	391 (22.3%)	362 (11.5%)	259 (12.6%)	1555 (17.9%)
31 – above	399 (22.9%)	114 (6.5%)	199 (6.3%)	48 (2.3%)	760 (8.7%)
Total	1740 (100%)	1754 (100%)	3154 (100%)	2060 (100%)	8708 (100%)

The loglinear analysis showed a significant three-way interaction between language, genre and sentence pattern ( $\chi^2(9) = 142.69$ ,  $p < .001$ ). Subsequent chi-square tests showed a difference in sentence length between English and Dutch for academic prose ( $\chi^2(3) = 129.29$ ,  $p < .001$ , Cramer's  $V = .20$ ), newspaper articles ( $\chi^2(3) = 48.52$ ,  $p < .001$ , Cramer's  $V = .11$ ) and leaflets ( $\chi^2(3) = 204.37$ ,  $p < .001$ , Cramer's  $V = .20$ ). No difference in sentence length was found length for the short stories genre ( $\chi^2(3) = 2.99$ ,  $p = .39$ , Cramer's  $V = .02$ ).

In the academic prose genre English sentences were found to be longer than Dutch sentences. In English 573 sentences (39.8%) belong to the 31 words and above category, whereas in Dutch 399 sentences (22.9%) belong to this category. In contrast, Dutch contains more sentences that belong to the 1-10 and 11-20 words categories than English (Dutch: 1-10: 184 (10.6%); 11-20: 614 (35.3%); English: 1-10: 74 (5.1%); 11-20: 361 (25.1%)).

In the newspaper genre sentences are also longer in English than in Dutch, with English containing more sentences in the 31-above category (English: 215 (11.7%); Dutch: 114 (6.5%)) and Dutch containing more sentences in the 11-20 words category (Dutch: 875 (50.0%); English: 749 (40.6%)).

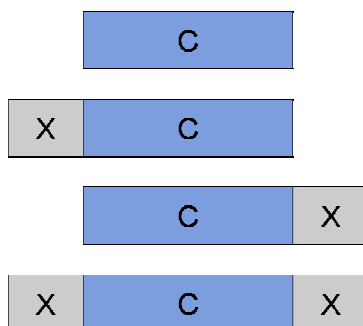
In the short stories genre the majority of sentences belong to the 1-10 words category in both languages (English: 1670 (52.7%); Dutch: 1647 (52.2%)), followed by the 11-20 category (English: 902 (28.5%); Dutch: 946 (30.0%)).

Last, in the leaflets genre English sentences are again longer than Dutch sentences. The differences are most pronounced in the 31-above category (English: 163 (10.3%); Dutch: 48 (2.3%)), but are also substantial between the 1-10 category

(English: 355 (22.3%); Dutch: 710 (34.5%)) and the 21-30 category (English: 362 (22.8%); Dutch: 259 (12.6%)).

### 5.3 Main sentence patterns

On the basis of the discourse segmentation and annotation system described in Chapter 2, all sentences can be categorised as belonging to one of four main patterns. These are sentences that consist of only a nuclear unit, (C); sentences that consist of a nuclear unit and one or more prepended satellites (XC); sentences that consist of a nuclear unit and one or more appended satellites (CX); or sentences that consist of a nuclear unit, one or more prepended satellites and one or more appended satellites (XCX) (see Figure 1 below). Each of these four main patterns can be further categorised into a range of subpatterns. The main patterns and their respective subpatterns will be described in detail in the sections below. The patterns formed by the interpolated satellites (cf. 2.5.3) will be described in Chapter 8, as these have been studied extensively in a separate case study.



**Figure 1 Main sentence patterns**

The following sentences exemplify each of these four main patterns:

#### C-pattern

- (1) <C>PvdA-leider Bos heeft gisteren nieuwe regels voorgesteld om het toezicht op bedrijven te verscherpen<C>. <s1859, newspaper articles>

<C>Dutch Labour Party leader Bos proposed new rules yesterday to increase inspection of companies<C>.

**XC pattern**

- (2) <A>Once they start<A>, <C>there is no knowing where they will end<C>. <s33, newspaper articles>

**CX pattern**

- (3) <C>De eetstoornis kan een reactie zijn op die veranderingen<C>, <D>die als angstig en bedreigend worden ervaren<D>. <s9005, leaflets>

(<C>The eating disorder could be a reaction to these changes<C>, <D>which are perceived as frightening and threatening<D>.)

**XCX pattern**

- (4) <A>Yesterday<A>, <C>Tony Blair warned what the ultimate effect of their rebellion would be<C> – <D>a return to Tory government<D>. <s447, newspaper articles>

**5.3.1 Frequencies of main sentence patterns**

Table 3 contains the frequencies of the four patterns in each of the four genres in both languages. A visual representation of these frequencies can be found in Figure 2 below.

**Table 3** Frequencies of main sentence patterns: C, XC, CX and XCX

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
C-pattern	666 (46.3%)	1264 (68.5%)	1911 (60.3%)	1044 (65.7%)	4885 (60.8%)
XC pattern	449 (31.2%)	312 (16.9%)	353 (11.1%)	265 (16.7%)	1379 (17.2%)
CX pattern	207 (14.4%)	221 (12.0%)	751 (23.7%)	231 (14.5%)	1410 (17.5%)
XCX pattern	116 (8.1%)	47 (2.5%)	154 (4.9%)	49 (3.1%)	366 (4.6%)
Total	1438 (100%)	1844 (100%)	3169 (100%)	1598 (100%)	8040 (100%)
<b>Dutch</b>					
C-pattern	836 (48.0%)	1010 (57.6%)	1740 (55.2%)	1173 (56.9%)	4759 (54.7%)
XC pattern	521 (29.9%)	466 (26.6%)	506 (16.0%)	584 (28.3%)	2077 (23.9%)
CX pattern	257 (14.8%)	206 (11.7%)	727 (23.1%)	229 (11.1%)	1419 (16.3%)
XCX pattern	126 (7.2%)	72 (4.1%)	181 (5.7%)	74 (3.6%)	453 (5.2%)
Total	1740 (100%)	1754 (100%)	3154 (100%)	2060 (100%)	8708 (100%)



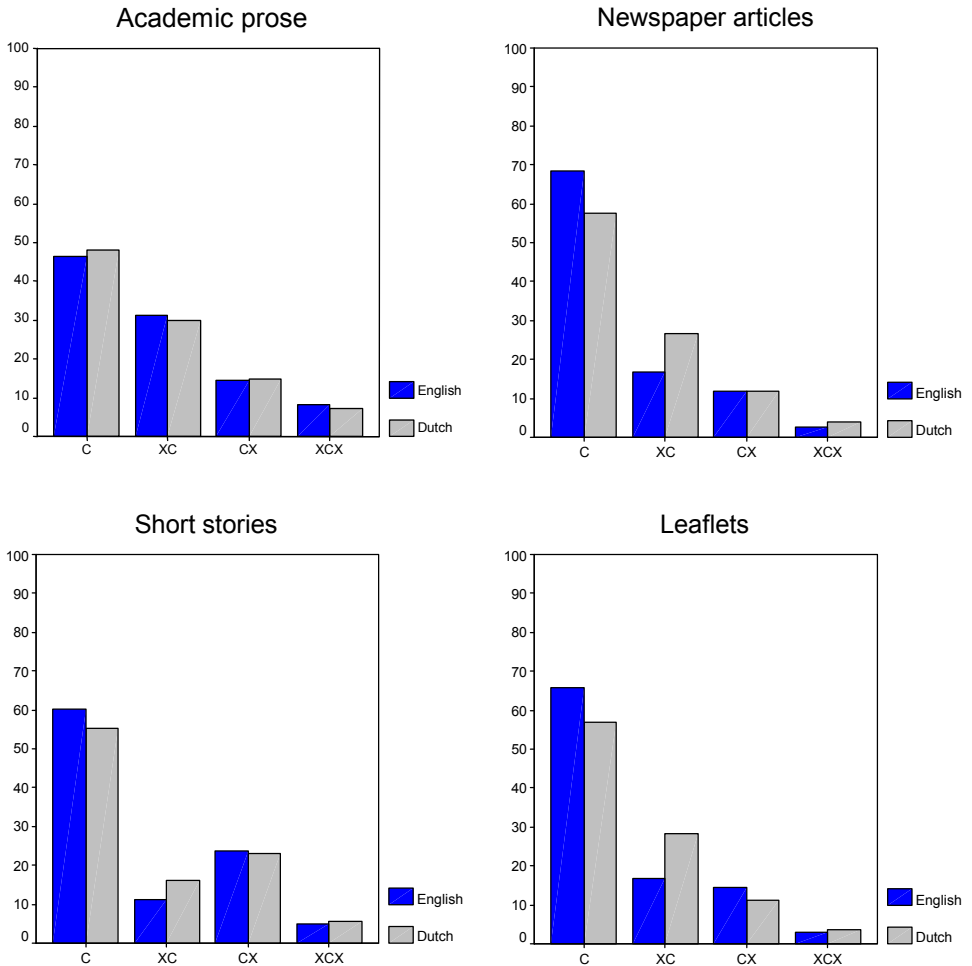
The loglinear analysis showed a significant three-way interaction between language, genre and sentence pattern ( $\chi^2(9) = 65.12$ ,  $p < .001$ ). Subsequent chi-square tests showed differences in the distribution of sentence patterns between English and Dutch for newspaper articles ( $\chi^2(3) = 62.42$ ,  $p < .001$ , Cramer's  $V = .13$ ), short stories ( $\chi^2(3) = 37.79$ ,  $p < .001$ , Cramer's  $V = .07$ ) and leaflets ( $\chi^2(3) = 72.87$ ,  $p < .001$ , Cramer's  $V = .14$ ). No differences in frequencies of the different sentence patterns were found for the academic prose genre ( $\chi^2(3) = 1.70$ ,  $p = .63$ , Cramer's  $V = .02$ ).

In the academic prose genre, the languages show rather similar frequencies, with the C-pattern forming the largest pattern (English: 666 (46.3%); Dutch: 836 (48.0%)), followed by the XC pattern (English: 449 (31.2%); Dutch: 521 (29.9%)), the CX pattern (English: 207 (14.4%); Dutch: 257 (14.8%)), and finally by the XCX pattern (English: 116 (8.1%); Dutch: 126 (7.2%)).

In the newspaper genre, differences between English and Dutch are particularly evident in the C-pattern (English: 1264 (68.5%); Dutch: 1010 (57.6%)) and the XC pattern (English: 312 (16.9%); Dutch: 466 (26.6%)). The frequencies of the CX pattern, on the other hand, are similar (English: 221 (12.0%); Dutch: 206 (11.7%)).

In the short stories genre, the main distribution difference was found in the frequencies of the XC pattern, which is again more frequent in Dutch (506 (16.0%)) than in English (353 (11.1%)). The frequencies of both the CX (English: 751 (23.7%); Dutch: 727 (23.1%)) and the XCX pattern (English: 154 (4.9%); Dutch: 181 (5.7%)) are again similar.

Last, in the leaflets genre, the main difference was again found in the XC pattern, which is more frequent in Dutch (584 (28.3%)) than in English (265 (16.7%)). The C and the CX pattern, on the other hand, are more frequent in English than in Dutch (English C: 1044 (65.7%); Dutch C: 1173 (56.9%); English CX: 231 (14.5%); Dutch CX: 229 (11.1%)).



**Figure 2 Sentence patterns in four genres in English and Dutch in percentages**

### 5.3.2 Conclusion

A comparison of the four main sentence patterns in English and Dutch in four different genres shows that with the exception of the academic prose genre, the XC pattern is more frequent in Dutch than in English across the different genres. The C-pattern is more frequent in English in the newspaper, short stories and leaflets genre. The CX pattern has a similar distribution in English and Dutch, except for the leaflets genre. Finally, no significant differences between the languages were found for the distribution of sentence patterns in the academic prose genre.

## 5.4 The C-pattern

Both Table 3 and Figure 2 above clearly show that the C-pattern is the most frequently occurring sentence pattern in both languages across the four genres. This C can either take the form of a single, uncoordinated nuclear unit or of two or more coordinated nuclear units, creating two main subpatterns of the C-pattern, i.e. the C and Ca/b/x patterns respectively (cf. Chapter 8 on subpatterns formed by interpolated satellites that can occur in the nucleus). Table 4 contains examples of these two main subpatterns of the main C-pattern and Table 5 provides their respective frequencies.

**Table 4 Subpatterns of C-pattern**

Subpattern of C	Example
C (single, uncoordinated nucleus)	<C>Blundering Bush and his right-wing advisers openly denigrate the international body<C>. <s255, newspaper articles>
Ca/b (coordinated nuclei)	<Ca>Families must take responsibility for their health<Ca> <Cb>but the Government and food manufacturers also have big role to play<Cb>. <s69, newspaper articles>

**Table 5 Frequencies of subpatterns of the C-pattern**

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
C	489 (73.4%)	1079 (85.4%)	1430 (74.8%)	846 (81.0%)	3844 (78.7%)
Ca/b	177 (26.6%)	185 (14.6%)	481 (25.2%)	198 (19.0%)	1041 (21.3%)
Total	666 (100%)	1264 (100%)	1911 (100%)	1044 (100%)	4885 (100%)
<b>Dutch</b>					
C	708 (84.7%)	872 (86.3%)	1361 (78.2%)	1001 (85.3%)	3942 (82.8%)
Ca/b	128 (15.3%)	138 (13.7%)	379 (21.8%)	172 (14.7%)	817 (17.2%)
Total	836 (100%)	1010 (100%)	1740 (100%)	1173 (100%)	4759 (100%)

The loglinear analysis showed a significant three-way interaction between language, genre and subpatterns of the C-pattern ( $\chi^2(3) = 14.34, p < .01$ ). Subsequent chi-square tests showed a difference in distribution of the subpatterns of the C-pattern between English and Dutch for the academic prose genre ( $\chi^2(1) = 29.07, p < .001$ , Cramer's  $V = .13$ ) and the leaflets genre ( $\chi^2(1) = 7.35, p < .01$ , Cramer's  $V = .05$ ). No difference between the languages in distribution of the subpatterns of the C-pattern was found

for the newspaper genre ( $\chi^2(1) = .43, p = .50$ , Cramer's  $V = .01$ ) and the short stories genre ( $\chi^2(1) = 5.80, p = .016^{29}$ , Cramer's  $V = .04$ ).

Although the subpattern formed by uncoordinated nuclei is more frequent in both languages, a difference in frequency of the two subpatterns is particularly evident for the academic prose genre, with English containing more coordinated nuclei than Dutch (177 (26.6%) vs. 128 (15.3%).

In the newspaper genre, the vast majority of sentences consist of uncoordinated nuclei in both languages (English: 1079 (85.4%); Dutch: 872 (86.3%)).

The same applies to sentences in the short stories genre, although the percentages for coordinated nuclei are relatively high when compared to the other genres (English: 481 (25.2%); Dutch: 379 (21.8%)).

Finally, although most sentences in both the English and Dutch leaflets genre consist of uncoordinated nuclei, English contains more coordinated nuclei in comparison to Dutch (English: 198 (19.0%); Dutch: 172 (14.7%)).

#### 5.4.1 Grammatical realisation of subpattern C – uncoordinated, single unit

The wide range of potential grammatical realisations of the uncoordinated C have been clustered into two main groups: those Cs that are realised as independent clauses and those that are not realised as independent clauses. Table 6 presents the frequencies of both types of realisation.

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Independent clause	484 (99.0%)	952 (88.2%)	1113 (77.8%)	816 (96.5%)	3365 (87.5%)
Non-independent clause	5 (1.0%)	127 (11.8%)	317 (22.2%)	30 (3.5%)	479 (12.5%)
<b>Total</b>	<b>489 (100%)</b>	<b>1079 (100%)</b>	<b>1430 (100%)</b>	<b>846 (100%)</b>	<b>3844 (100%)</b>
<b>Dutch</b>					
Independent clause	705 (99.6%)	838 (96.1%)	1043 (76.6%)	904 (90.3%)	3490 (88.5%)
Non-independent clause	3 (0.4%)	34 (3.9%)	318 (23.4%)	97 (9.7%)	452 (11.5%)
<b>Total</b>	<b>708 (100%)</b>	<b>872 (100%)</b>	<b>1361 (100%)</b>	<b>1001 (100%)</b>	<b>3942 (100%)</b>

<sup>29</sup> This is not considered a significant result in this study, because of the predetermined alpha level of .01 (cf. 4.5 on statistical analyses).

The loglinear analysis showed a significant three-way interaction between language, genre and grammatical realisation of the subpattern of C, in which C is a single, uncoordinated unit ( $\chi^2(3) = 73.18, p < .001$ ). Subsequent chi-square tests showed a difference in frequency of the realisation of C between English and Dutch for the newspaper genre ( $\chi^2(1) = 39.46, p < .001$ , Cramer's  $V = .14$ ) and the leaflets genre ( $\chi^2(1) = 27.03, p < .001$ , Cramer's  $V = .12$ ). No difference in frequency of the realisation of C was found for the academic prose genre ( $\chi^2(1) = 1.56, p = .21$ , Cramer's  $V = .03$ ) and the short stories genre ( $\chi^2(1) = .59, p = .45$ , Cramer's  $V = .01$ ).

For the academic prose genre, both languages follow a similar pattern in the sense that they contain hardly any instances of non-independent clauses and that the vast majority of sentences are realised as independent clauses (English: 484 (99.0%); Dutch: 705 (99.6%)).

In the newspaper genre, the difference between English and Dutch can be found in the higher frequency of non-independent clauses in English (127 (11.8%) vs. 34 (3.9%)). It should be noted that a large number of these non-independent clauses take the form of a reporting clause, introducing reported speech (cf. 2.5.2).

The short stories genre, contains the highest percentage of Cs realised as non-independent clauses, with both languages showing similar frequencies (English: 317 (22.2%); Dutch: 318 (23.4%)).

In the leaflets genre, the frequencies of C when realised as a non-independent clause are higher in Dutch than in English (97 (9.7%) vs. 30 (3.5%)).

### **Make-up of the non-independent clause category**

Even though the Cs that are realised as non-independent clauses only form a small percentage of the total number of Cs, it is interesting to have a closer look at the make-up of this category in the various genres, as there are some subtle differences between the languages. The non-independent clause-category can be further broken down into four subcategories: (1) phrasal fragments<sup>30</sup> (i.e. noun phrases, adjective phrases, conjuncts and disjuncts); (2) clausal fragments (i.e. mainly fragments and reporting clauses, see example (5) below); (3) adverbial clauses, and (4) discourse markers,

---

<sup>30</sup> The reason why this category received the label 'phrase fragment' and not just 'phrase' is to indicate explicitly that the nucleus takes the form of a phrase and not a clause (cf. 3.3.1 & 3.5.1 for more detailed explanation of fragment category).

question words, vocatives and tag questions, which predominantly occur in the short stories genre. Table 7 presents the frequencies of these four main realisation groups.

**Table 7** Make-up of the non-independent clause in subpattern C – single unit in 4 realisation groups

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Fragment phrase	2 (40.0%)	26 (20.5%)	172 (54.3%)	24 (80.0%)	224 (46.8%)
Fragment clause	3 (60.0%)	95 (74.8%)	92 (29.0%)	6 (20.0%)	196 (40.9%)
Adverbial clause	0 (0.0%)	6 (4.7%)	17 (5.4%)	0 (0.0%)	23 (4.8%)
Discourse marker/vocative/question word/tag	0 (0.0%)	0 (0.0%)	36 (11.4%)	0 (0.0%)	36 (7.5%)
<b>Total</b>	<b>5 (100%)</b>	<b>127 (100%)</b>	<b>317 (100%)</b>	<b>30 (100%)</b>	<b>479 (100%)</b>
<b>Dutch</b>					
Fragment phrase	2 (66.7%)	8 (23.4%)	198 (62.3%)	57 (58.8%)	265 (58.6%)
Fragment clause	1 (33.3%)	24 (70.6%)	69 (21.7%)	26 (26.8%)	120 (26.5%)
Adverbial clause	0 (0.0%)	2 (5.9%)	16 (5.0%)	14 (14.4%)	32 (7.1%)
Discourse marker/vocative/question word/tag	0 (0.0%)	0 (0.0%)	35 (11.0%)	0 (0.0%)	35 (7.7%)
<b>Total</b>	<b>3 (100%)</b>	<b>34 (100%)</b>	<b>318 (100%)</b>	<b>97 (100%)</b>	<b>452 (100%)</b>

In line with Table 6 above, Table 7 shows that the academic prose genre contains hardly any instances of Cs that are realised as non-independent clauses in both languages. The newspaper genre, on the other hand, does contain quite a few instances Cs realised as non-independent clauses, especially in English, most of which take the form of clause fragments, and more specifically, reporting clauses. Sentence (5) provides an example of a (reporting) clause fragment in the English newspaper genre.

- (5) <C><reportingcl><fragment>The Mid-Ulster MP told the conference in Dublin<fragment><reportingcl><C>: <s506, newspaper articles>

Compared to the other genres, the short stories genre contains most instances of nuclear units that are realised as non-independent clauses, with the languages showing similar frequencies in all four categories. In both languages, these typically take the form of phrasal fragments (English: 172 (54.3%); Dutch: 198 (62.3%)). They also take the form of clausal fragments (English: 92 (29.0%); Dutch: 69 (21.7%)), discourse markers, vocatives, question words or tags (English: 36 (11.4%); Dutch: 35 (11.2%)), as well as adverbial clauses (English: 17 (5.4%); Dutch: 16 (5.1%)).

Sentence (6) provides an example of a phrasal fragment and sentence (7) one of a clausal fragment in the short stories genre.

- (6) (Maybe he was confused. <s10779, short stories>) <C><fragment\_VP>Or  
brainwashed<fragment\_VP><C>. <s10780, short stories>
- (7) (Wat is jammer? <s14072, short stories>) <C><complement\_cl>Dat je niet inziet  
dat hij gelijk had<complement\_cl><C>. <s14073, short stories>
- ((What is a pity?) <C><complement\_cl>That you don't see that he was  
right<complement\_cl><C>.)

With respect to the leaflets genre, Table 6 above already showed that Dutch contains more Cs that are realised as non-independent clauses than English. Table 7 shows that in both languages most of these nuclear units take the form of phrasal fragments (English: 24 (80.0%); Dutch: 57 (58.8%)). However, in Dutch this C is realised as a clausal fragment or adverbial clause in quite a number of cases (clause fragment: 26 (26.8%); adverbial clause: 14 (14.4%)). Sentences (8) and (9) provide examples of a phrasal fragment and adverbial clause respectively.

- (8) (Will I have to change my lifestyle after donating? <s7107, leaflets>)  
<C><fragment\_no>No<fragment\_no><C>. <s7108, leaflets>
- (9) (Soms kan cannabisgebruik leiden tot afhankelijkheid. <s8818, leaflets>)  
<C><advcl\_condition>Met name als iemand al eerder een verslaving heeft  
gehad<advcl\_condition><C>. <s8819, leaflets>
- (Sometimes cannabis use can lead to dependency.)  
<C><advcl\_condition>Especially if someone has had an addiction  
before<advcl\_condition><C>.)

#### 5.4.2 Grammatical realisation of subpattern C – coordinated nuclei

The second subpattern of the C-pattern consists of a nucleus that is coordinated with one or more other nuclei. Table 5 above showed that the English academic prose and leaflets genres contain more instances of coordinated nuclei when compared to these genres in Dutch. Table 8 below shows that these coordinated nuclei can be realised syntactically either as coordinated independent clauses or as non-independent clauses, i.e. as coordinated subordinate clauses, phrases or fragments. Note that coordinated subordinate clauses or phrases have only been annotated when they are presented as separate discourse units (cf. 2.4.3 and 3.3.1). In these

cases the second coordinate, and possible third, fourth, and so on, consists of a subordinate clause, a phrase or a fragment that is coordinated with a subordinate clause, a phrase or a fragment of the first coordinate (see examples (11) and (12) below).

**Table 8 Grammatical realisation of subpattern C – coordinated nuclei**

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Coordinated indep. clauses	116 (65.5%)	162 (87.6%)	411 (85.4%)	167 (84.3%)	856 (82.2%)
Coordinated subcl/phrases/frag	61 (34.5%)	23 (12.4%)	70 (14.6%)	31 (15.7%)	185 (17.8%)
Total	177 (100%)	185 (100%)	481 (100%)	198 (100%)	1041 (100%)
<b>Dutch</b>					
Coordinated indep. clauses	103 (80.5%)	120 (87.0%)	325 (85.8%)	149 (86.6%)	697 (85.3%)
Coordinated subcl/phrases/frag	25 (19.5%)	18 (13.0%)	54 (14.2%)	23 (13.4%)	120 (14.7%)
Total	128 (100%)	138 (100%)	379 (100%)	172 (100%)	817 (100%)

The loglinear analysis showed no significant three-way interaction between language, genre and grammatical realisation of the subpattern of C, in which C consists of coordinated nuclei ( $\chi^2(3) = 5.82$ ,  $p = .12$ ). The analysis did show a significant two-way interaction between genre and grammatical realisation of Ca/b ( $\chi^2(3) = 33.22$ ,  $p < .001$ , Cramer's  $V = .14$ ). Even though genre differences irrespective of language are not the main concern of this study, the main genre differences can be summarised as follows. Table 8 shows that across the genres the vast majority of coordinated Cs take the form of coordinated independent clauses in both languages. The English academic prose genre is the only genre in which the percentage of coordinated Cs realised as non-independent clauses is higher when compared to the other genres and when compared to the Dutch academic prose genre (English: 61 (34.5%); Dutch: 25 (19.5%)). Sentences (10) and (11) provide examples of both types of grammatical realisation of coordinated nuclei.

- (10) <Ca><coord\_indepcl\_a>Families must take responsibility for their health<coord\_indepcl\_a><Ca> <Cb><coord\_indepcl\_b>but the Government and food manufacturers also have a big role to play<coord\_indepcl\_b><Cb>. <s69, newspaper articles>
- (11) <Ca><coord\_a\_embed\_asyn>U wordt geadviseerd om het bad dat langere tijd leeg heeft gestaan eerst te vullen met water van 60 C of



heter<coord\_a\_emb\_asyn><Ca>, <Cb><coord\_b\_emb\_asyn>dit water enige tijd (5 à 10 minuten) te laten bubbelen<coord\_b\_emb\_asyn><Cb> <Cc><coord\_c\_emb>en het bad vervolgens leeg te laten lopen<coord\_c\_emb><Cc>. <s9413, leaflets>

(<Ca><coord\_a\_embed\_asyn>You are advised to first fill the bathtub that has been empty for a longer period of time with water that is 60 C or warmer<coord\_a\_emb\_asyn><Ca>, <Cb><coord\_b\_emb\_asyn>to let this water bubble for some time (5 to 10 minutes) <coord\_b\_emb\_asyn><Cb> <Cc><coord\_c\_emb>and then to empty the bathtub<coord\_c\_emb><Cc>.)

### 5.4.3 Summary

A close analysis of the most frequent main sentence pattern, the C-pattern, shows it can be further subdivided into two main patterns: it can either consist of one uncoordinated nuclear unit or of a number of coordinated nuclear units. When it consists of one nuclear message, this can be realised as either an independent clause or a non-independent clause. An analysis of the grammatical realisation of sentences that consist of just one nuclear unit showed that this is typically realised as an independent clause in both languages. Compared to the other genres, the short stories genre contains the highest frequencies of sentences that are realised as non-independent clauses in both English and Dutch. Significant differences between the languages in the frequencies of non-independent clauses (mainly reporting clauses) were found for the newspaper genre and the leaflets genre. The English newspaper genre contains more instances of Cs realised as non-independent clauses, whereas the reverse pattern holds for the Dutch leaflets genre.

With respect to the coordinated nuclei subpattern, the analysis showed significant differences between the English and Dutch academic prose genre and leaflets genre, with coordinated nuclei occurring more frequently in English than in Dutch in these genres. In both languages these coordinated nuclei are typically realised as independent clauses in all genres, with the exception of the academic genre, which shows a higher frequency of coordinated non-independent clauses, especially in English.

## 5.5 The XC pattern

Table 3 above already showed that the XC pattern is significantly more frequent in the Dutch newspaper, short stories and the leaflets genre when compared to the

frequency of this pattern in these English genres. In the XC pattern, a sentence consists of a nuclear unit that is to be preceded by one or two sentence-initial elements, and in a very small number of cases by three elements. The sentence-initial elements can be of different types and can form various combinations with each other to form five main subpatterns. They can take the form of a prepended satellite <A>, <B> and <Z> (cf. 2.5.3), an interpolated satellite <1> (cf. 2.5.3) or a so-called <zz> element. The <zz> elements can be characterised by the fact that they are typically not presented as separate punctuation units (cf. 2.4.3). However, as the difference between prepended satellites and zz elements mainly concerns one in punctuation practice between the languages, these categories have been conflated for the sake of analysis and are both subsumed under the category 'A-satellite'.

The following section will describe the make-up of the XC pattern and compare the frequencies and distribution of the various subpatterns. The subpattern in which the X consists of one element, the AC pattern, will receive the main focus, as this is the most frequently occurring subpattern. The subpatterns in which the X consists of two or three elements will be described briefly here and in more detail in Chapter 6, which focuses on the complex beginnings of sentences.

Table 9 presents an overview of all the possible subpatterns, including examples, and Table 10 presents the frequencies of the five main subpatterns.

**Table 9 Subpatterns of the XC pattern**

Make-up of X	Subpattern of XC	Example
X = 1 element	A – C	<A>Instead<A>, <C>more Brits went to live in Spain<C>. <s120, newspaper articles>
X = 2 elements	A1A – C	<A>Although you may still hear the term AIDS (<1>Acquired Immune Deficiency Syndrome<1>)<A> <C>it is no longer used by doctors<C>. <s7029, leaflets>
	A1 – C	<A>Now<A>, <1>a year after Saddam fell<1>, <C>Mr Straw says the lid has come off the pressure cooker<C>. <s164, newspaper articles>
	AB – C	<A>Therefore<A>, <B>whenever he seems hungry<B>, <C>just put him to the breast<C>. <s7981, leaflets>
X = 3 elements	--	<A>Unlike Australia<A>, <1>where the social tensions which surfaced during Queen Victoria's 1887 jubilee were closely linked to the growth of republicanism<1>, <B>in South Africa<B> <C>anti-capitalist sentiment did not lead to the rejection of imperialism or the values of the 'old world'<C>. <s4391, academic prose>

**Table 10**      **Frequencies of the XC pattern: X = 1, 2 or 3 elements**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
XC – X= 1 element	369 (82.2%)	281 (90.1%)	302 (85.6%)	233 (87.9%)	1185 (85.9%)
XC – X = 2 elements	67 (14.9%)	29 (9.3%)	47 (13.3%)	27 (10.2%)	170 (12.3%)
XC – X = 3 elements	13 (2.9%)	2 (0.6%)	4 (1.1%)	5 (1.9%)	24 (1.7%)
<b>Total</b>	<b>449 (100%)</b>	<b>312 (100%)</b>	<b>353 (100%)</b>	<b>265 (100%)</b>	<b>1379 (100%)</b>
<b>Dutch</b>					
XC – X= 1 element	482 (92.5%)	438 (94.0%)	466 (92.1%)	549 (94.0%)	1935 (93.2%)
XC – X = 2 elements	37 (7.1%)	28 (6.0%)	38 (7.5%)	32 (5.5%)	135 (6.5%)
XC – X = 3 elements	2 (0.4%)	0 (0.0%)	2 (0.4%)	3 (0.5%)	7 (0.3%)
<b>Total</b>	<b>521 (100%)</b>	<b>466 (100%)</b>	<b>506 (100%)</b>	<b>584 (100%)</b>	<b>2077 (100%)</b>

The loglinear analysis showed no significant three-way interaction between language, genre and distribution of the three main subpatterns of the XC pattern ( $\chi^2(6) = 2.78$ ,  $p = .83$ ). The analysis did show a significant two-way interaction between language and frequencies of the three subpatterns of the XC pattern ( $\chi^2(2) = 49.97$ ,  $p < .001$ , Cramer's  $V = .12$ ). Differences between the languages can be found in the frequencies of the XC subpattern in which the X consists of two elements, which is more frequent in English than in Dutch (English: 170 (12.3%); Dutch: 135 (6.5%)). The same applies to the subpattern in which the X consists of three elements: this is also more frequent in English (24 (1.7%)) than in Dutch (7 (0.3%)), although frequencies of this pattern are very low in both languages. The subpattern in which the X consists of one element is the most frequent subpattern in both languages (English: 1185 (85.9%); Dutch: 1935 (93.2%)).

### **5.5.1 The AC subpattern**

As the AC subpattern is by far the most frequent subpattern of the three main subpatterns of the XC pattern (see Table 10 above), this will be described in more detail in the following section. In fact, 14.7% (1185) of all English sentences in the entire corpus belong to this subpattern, compared to 22.2% (1935) of all Dutch sentences.

### Grammatical realisation of A in the AC subpattern

Table 11 below presents the frequencies of the A-element when it takes the form of either a phrase or a clause.<sup>31</sup>

**Table 11** Grammatical realisation of A in the AC subpattern as phrase or clause

	Academic prose	Newspaper articles	Short stories	Leaflets	Total
A = phrase	290 (78.6%)	197 (70.1%)	229 (75.8%)	142 (60.9%)	858 (72.4%)
A = clause	79 (21.4%)	84 (29.9%)	73 (24.2%)	91 (39.1%)	327 (27.6%)
Total	369 (100%)	281 (100%)	302 (100%)	233 (100%)	1185 (100%)
<b>Dutch</b>					
A = phrase	422 (87.6%)	376 (85.8%)	372 (79.8%)	447 (81.4%)	1617 (83.6%)
A = clause	60 (12.4%)	62 (14.2%)	94 (20.2%)	102 (18.6%)	318 (16.4%)
Total	482 (100%)	438 (100%)	466 (100%)	549 (100%)	1935 (100%)

The loglinear analysis showed a significant three-way interaction between language, genre and grammatical realisation of the AC subpattern ( $\chi^2(3) = 12.52$ ,  $p < .01$ ). Subsequent chi-square tests showed a difference in frequency of the realisation of A as a phrase or clause between English and Dutch for the academic prose genre ( $\chi^2(1) = 12.28$ ,  $p < .001$ , Cramer's  $V = .12$ ), the newspaper genre ( $\chi^2(1) = 26.20$ ,  $p < .001$ , Cramer's  $V = .19$ ) and the leaflets genre ( $\chi^2(1) = 36.89$ ,  $p < .001$ , Cramer's  $V = .21$ ). No difference in frequency of the realisation of A between the languages was found for the short stories genre ( $\chi^2(1) = 1.72$ ,  $p = .18$ , Cramer's  $V = .04$ ).

In all genres in both English and Dutch the A-element typically takes the form of a phrase. Differences between the languages can be found in a higher frequency of clauses in English when compared to Dutch. In the English academic prose genre, 79 A-elements (21.4%) take the form of a clause, compared to 60 in Dutch (12.4%).

A similar pattern can be found in the newspaper genre, in which the frequency of As realised as clauses is again higher for English than Dutch (84 (29.9%) vs. (62 (14.2%)).

<sup>31</sup> Note that similar to all the other discourse units, the A-element can also take the form of either a single uncoordinated element or of two or more coordinated elements. However, as only 16 sentences in English (1.3%) and 11 sentences in Dutch (0.5%) contain a coordinated A-element (Aa-Ab), these have not been further investigated.

In the short stories genre both languages follow a similar pattern, in that A typically takes the form of a phrase (English: 229 (75.8%); Dutch: 372 (79.8%)) and frequencies for A realised as clause are also similar for both languages (English: 73 (24.2%); Dutch: 94 (20.2%)).

Finally, the leaflets genre is the genre in which differences between the languages are most pronounced, with the frequencies of As taking the form of clauses again being much higher in English than in Dutch (91 (39.1%) vs. (102 (18.6%)).

The phrasal element can be further subcategorised into four main realisation groups. It can take the form of (1) an adjunct or prepositional phrase, (2) a conjunct, (3) a disjunct or subjunct, or (4) a discourse marker, question word, vocative or the words *yes/no*. To this last realisation group a very small number of grammatical forms have been added that could not be added to the other categories. These include a few instances of fronted adjective phrases and fragments. Sentences (12) to (15) provide examples of each of the grammatical forms the phrasal element in the AC pattern can take and Table 12 presents the frequencies of each of the four types.

(12) <zz><pp>In deze diepe slaap<pp><zz> zijn mensen moeilijk te wekken. <s8389, leaflets>

(<zz><g\_pp>In this deep sleep<pp><zz> people are difficult to wake.)

(13) <A><conj>Instead<conj><A>, this article has focused on normalities. <s4643, academic prose>

(14) <zz><disj>Natuurlijk<disj><zz> heeft Bos gelijk als hij zich zorgen maakt over het betrouwbaarheidsimago van de bedrijven. <s1863, newspaper articles>

(<zz><disj>Of course<disj><zz> Bos is right if he is concerned about the credibility of companies.)

(15) <A><dm>All right then<dm><A>, Jordan Pennance is up to something with Rose Sancreed. <s11740, short stories>

**Table 12**      **Frequencies of grammatical realisations of A as a phrase in the AC subpattern in four realisation groups**

English	Academic prose	Newspaper article	Short stories	Leaflets	Total
A = adjunct/PP	159 (54.8%)	138 (70.1%)	134 (58.5%)	85 (59.9%)	516 (60.1%)
A = conjunct	110 (37.9%)	34 (17.3%)	38 (16.6%)	52 (36.6%)	234 (27.3%)
A = disjunct/subjunct	19 (6.6%)	16 (8.1%)	26 (11.4%)	5 (3.5%)	66 (7.7%)
A = discourse marker, question word, vocative, yes/no	2 (0.7%)	9 (4.6%)	31 (13.5%)	0 (0.0%)	42 (4.9%)
Total	290 (100%)	197 (100%)	229 (100%)	142 (100%)	858 (100%)
<b>Dutch</b>					
A = adjunct/PP	295 (69.9%)	245 (65.2%)	298 (80.1%)	317 (70.9%)	1155 (71.4%)
A = conjunct	110 (26.6%)	100 (26.6%)	23 (6.2%)	104 (23.3%)	337 (20.8%)
A = disjunct/subjunct	15 (3.6%)	27 (7.2%)	9 (2.4%)	22 (4.9%)	73 (4.5%)
A = discourse marker, question word, vocative, yes/no	2 (0.5%)	4 (1.1%)	42 (11.3%)	4 (0.9%)	52 (3.2%)
Total	422 (100%)	376 (100%)	372 (100%)	477 (100%)	1617 (100%)

The loglinear analysis showed a significant three-way interaction between language, genre and grammatical realisation of the phrase in the AC subpattern ( $\chi^2(9) = 48.63$ ,  $p < .001$ ). Subsequent chi-square tests showed a difference in frequency of the realisation of the phrase between English and Dutch for the academic prose genre ( $\chi^2(3) = 17.33$ ,  $p < .001$ , Cramer's  $V = .15$ ), the newspaper genre ( $\chi^2(3) = 12.43$ ,  $p < .01$ , Cramer's  $V = .14$ ), the short stories genre ( $\chi^2(3) = 44.34$ ,  $p < .001$ , Cramer's  $V = .27$ ) and the leaflets genre ( $\chi^2(3) = 10.91$ ,  $p < .01$ , Cramer's  $V = .13$ ).

In the academic prose genre the group of adjuncts/Ps is larger in Dutch than in English (English: 159 (54.8%); Dutch: 295 (69.9%)). In English, on the other hand, the phrasal A-element more often takes the form of a conjunct when compared to Dutch (English: 110 (37.9%); Dutch: 110 (26.6%)).

For the newspaper genre the main differences between the languages can be found in the frequencies of conjuncts, which are more frequent in Dutch (100 (26.6%)) than in English (34 (17.3%)). In both languages the adjunct/PP group is again the largest realisation group (English: 138 (70.1%); Dutch: 245 (65.2%)).

The short stories genre shows the largest differences between the languages with respect to the grammatical realisation of A. Although the adjunct/PP group is the largest group in both languages, this is much larger in Dutch (298 (80.1%)) than in English (134 (58.5%)). In English more As take the form of either a conjunct (38 (16.6%) vs. 23 (6.2%)) or of a disjunct/subjunct (26 (11.4%))

vs. 9 (2.4%). The frequencies of the discourse marker group are similar for both languages (English: 31 (13.5%); Dutch: 42 (11.3%).

Finally, in the leaflets genre the main differences between the languages can be found in the conjunct realisation group, which occurs more frequently in English (52 (36.6%)) than in Dutch (104 (23.3%)). And although the adjunct/PP category is also the most frequently occurring realisation group in this genre in both languages, it is more frequent in Dutch (317 (70.9%)) than in English (85 (59.9%)).

The clausal element can also be further specified and subcategorised into seven realisation groups. These groups contain the following types of clauses: 1) adverbial clauses of time, 2) adverbial clauses of condition, 3) adverbial clauses of concession, 4) adverbial clauses of purpose, 5) adverbial clauses of reason, 6) other adverbial clauses, and 7) other types of clauses. Realisation group 6 contains many clauses that are introduced by the subordinator *as* in English or *zoals* in Dutch and a large number of non-finite clauses (see Tables 14 and 15 below) that have not been further specified into a semantic class. In English 59 (69.4%) of the 85 sentences in this realisation group are non-finite clauses and in Dutch this applies to 27 (54.0%) of the 50 sentences in this group. Realisation group 7 contains a few instances of complement clauses and comment clauses. Sentences (16) to (22) provide examples of each of the grammatical forms that the clausal element in the AC subpattern can take. The frequencies of these realisation groups are presented in Table 13 below.

(16) <A><advcl\_time>Zolang ik me herinner<advcl\_time><A>, heeft mijn moeder dat gezegd. <s13598, short stories>

(<A><advcl\_time>As long as I remember<advcl\_time><A>, my mother said that.)

(17) <A><advcl\_cond>But if we finally discover what really happened in Paris that fateful night<advcl\_cond><A>, the wait will have been worthwhile. <s144, newspaper articles>

(18) <A><advcl\_conces>Hoewel de gegevens uit deze database natuurlijk niet representatief genoeg zijn om tot algemene kwalitatieve uitspraken over de landelijke daling te komen<advcl\_conces><A>, leverde ze wel inzicht op in de richting van deze verandering. <s5138, academic prose>

(<A><advcl\_conces>Although the data of this database are of course not representative enough to draw general qualitative conclusions about the

national decrease<advcl\_conces><A>, they did provide insight into the direction of this change.)

- (19) <A><advcl\_purp\_nonfin>Om ernstige gevolgen te voorkomen<advcl\_purp\_nonfin><A> moet de infectie echter wel op tijd worden vastgesteld. <s9866, leaflets>

(<A><advcl\_purp\_nonfin>To prevent serious consequences from occurring<advcl\_purp\_nonfin><A> the infection however has to be identified in time.)

- (20) <A><advcl\_reas>Because they are overactive and impulsive<advcl\_reas><A>, children with ADHD often find it difficult to fit in at school. <s7198, leaflets>

- (21) <A><advcl\_nonfin>Having won the war<advcl\_nonfin><A>, Mr Bush is perilously close to losing the peace. <s250, newspaper articles>

- (22) <A><complementcl>Dat zij telkens in de krant moeten lezen dat het allemaal niet deugt<complementcl><A>, dat kan ik niet hebben. <s3481, newspaper articles>

(<A><complementcl>That they keep having to read in the newspaper that nothing is any good<complementcl><A>, that I can't take.)

**Table 13** Frequencies of grammatical realisations of A as a clause in the AC subpattern in seven realisation groups

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Adverbial cl_time	10 (12.7%)	19 (22.6%)	41 (56.2%)	10 (11.0%)	80 (24.5%)
Adverbial cl_condition	8 (10.1%)	29 (34.5%)	10 (13.7%)	46 (50.5%)	93 (28.4%)
Adverbial cl_concession	26 (32.9%)	12 (14.3%)	2 (2.7%)	7 (7.7%)	47 (14.4%)
Adverbial cl_purpose	4 (5.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	4 (1.2%)
Adverbial cl_reason	1 (1.3%)	0 (0.0%)	1 (1.4%)	6 (6.6%)	8 (2.4%)
Adverbial cl_not specified	28 (35.4%)	22 (26.2%)	16 (21.9%)	19 (20.9%)	85 (26.0%)
Different type of clauses	2 (2.5%)	2 (2.4%)	3 (4.1%)	3 (3.3%)	10 (3.1%)
Total	79 (100%)	84 (100%)	73 (100%)	91 (100%)	327 (100%)
<b>Dutch</b>					
Adverbial cl_time	8 (13.3%)	12 (19.4%)	51 (54.3%)	6 (5.9%)	77 (24.2%)
Adverbial cl_condition	14 (23.3%)	27 (43.5%)	23 (24.5%)	68 (66.7%)	132 (41.5%)
Adverbial cl_concession	7 (11.7%)	5 (8.1%)	3 (3.2%)	4 (3.9%)	19 (6.0%)
Adverbial cl_purpose	4 (6.7%)	5 (8.1%)	1 (1.1%)	7 (6.9%)	17 (5.3%)
Adverbial cl_reason	5 (8.3%)	2 (3.2%)	1 (1.1%)	6 (5.95)	14 (4.4%)
Adverbial cl_not specified	22 (36.7%)	8 (12.9%)	12 (12.8%)	8 (7.8%)	50 (15.7%)
Different type of clauses	0 (0.0%)	3 (4.8%)	3 (3.2%)	3 (2.9%)	9 (2.8%)
Total	60 (100%)	62 (100%)	94 (100%)	102 (100%)	318 (100%)



Due to the detailed level of categorisation, the expected frequencies of certain cells are too low to be tested statistically. A comparison of the frequencies between English and Dutch across the different genres shows that differences between the English and Dutch academic prose genre can be found in the frequencies of the adverbial clauses of condition and those of concession. The former is more frequent in Dutch (14 (23.3%) vs. 8 (10.1%)), whereas the latter is more frequent in English (26 (32.9%) vs. 7 (11.7%)). The distribution of the other types of clauses is similar in both languages. In both languages the adverbial clause group that has not been specified according to semantic class forms the largest group and, as explained above, these mainly contain instances of non-finite clauses (English: 28 (35.4%); Dutch: 22 (36.7%)).

In the newspaper genre, the group of adverbial clauses of condition forms the largest group in both languages, showing higher frequencies in Dutch than in English (27 (43.5%) vs. 29 (34.5%)). The second most frequent realisation group is adverbial clauses of time, which have similar frequencies in English (19 (22.6%)) and Dutch (12 (19.4%)). The group of non-specified, mainly non-finite clauses is larger in English (22 (26.2%) vs. 8 (12.9%)) and the same applies to the adverbial clauses of concession (12 (14.3%) vs. 5 (8.1%)). English contains no instances of adverbial clauses of purpose and reason, whereas Dutch contains a few.

In the short stories genre the adverbial clauses of time form the largest group, with the languages showing similar frequencies (English: 41 (56.2%); Dutch: 51 (54.3%)). The group of adverbial clauses of condition forms the second largest group in Dutch (23 (24.5%) vs. 10 (13.7%)), whereas this position is held by the non-specified adverbial clauses in English (16 (21.9%) vs. 12 (12.8%)).

Finally, in the leaflets genre the adverbial clause of condition forms the largest realisation group in both English and Dutch, but showing higher frequencies in Dutch (English: 46 (50.5%); Dutch: 68 (66.7%)). In Dutch the frequencies of the other realisation groups are fairly evenly spread across the different groups, whereas in English again the non-specified group forms a large group (19 (20.9%)).

In addition to being classified on the basis of their semantic roles, the adverbial clauses in initial position can also be subdivided into finite and non-finite clauses. Table 14 presents the frequencies of both types.

**Table 14**                    **Frequencies of A as a clause in the AC subpattern as finite and non-finite clauses**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Adverbial cl_finite	58 (73.4%)	64 (76.2%)	56 (76.7%)	77 (84.6%)	255 (78.0%)
Adverbial cl_non-finite	21 (26.6%)	20 (23.8%)	17 (23.3%)	14 (15.4%)	72 (22.0%)
<b>Total</b>	<b>79 (100%)</b>	<b>84 (100%)</b>	<b>73 (100%)</b>	<b>91 (100%)</b>	<b>327 (100%)</b>
<b>Dutch</b>					
Adverbial cl_finite	47 (78.3%)	53 (85.5%)	84 (89.4%)	93 (91.2%)	277 (87.1%)
Adverbial cl_non-finite	13 (21.7%)	9 (14.5%)	10 (10.6%)	9 (8.8%)	41 (12.9%)
<b>Total</b>	<b>60 (100%)</b>	<b>62 (100%)</b>	<b>94 (100%)</b>	<b>102 (100%)</b>	<b>318 (100%)</b>

The loglinear analysis showed no significant three-way interaction between language, genre and distribution of finite and non-finite clauses ( $\chi^2(3) = 1.28$ ,  $p = .73$ ). The analysis did show a significant two-way interaction between language and frequencies of finite and non-finite clauses ( $\chi^2(1) = 7.96$ ,  $p < .01$ , Cramer's  $V = .12$ ). The differences between English and Dutch can be found in the higher frequency of non-finite clauses in English when compared to Dutch (72 (22.0%); 41 (12.9%)). It should be noted, however, that in both languages the vast majority of sentence-initial clauses take the form of finite clauses (English: 255 (78.0%); Dutch: 277 (87.1%)).

The non-finite clauses can be of three different types: the non-finite verb can take the form of a present participle, a past participle or an infinitive. Verbless clauses have also been added to the category of non-finite clauses. The frequencies of these different types are presented in Table 15 below. Sentences (23) (present participle), (24) (past participle), (25) (infinitive) and (26) (verbless) provide examples of each of these types of non-finite clauses:

(23) <A><advcl\_nonfin>Stacking a breakfast tray with fruit juice, tea and toast<advcl\_nonfin><A>, I climbed the stairs to the top of the house. <s11241, short stories>

(24) <A><advcl\_nonfin>Vergeleken met hun kindertijd<advcl\_nonfin><A> besteden adolescenten steeds meer tijd aan activiteiten met leeftijdsgenoten zonder dat daar ouders bij aanwezig zijn. <s5940, academic prose>

(<A><advcl\_nonfin>Compared to their childhood<advcl\_nonfin><A> adolescents spend more and more time on activities with their peers without parents being present.)

- (25) <A><advcl\_purp\_nonfin>Om in het gesprek geen vragen te vergeten<advcl\_purp\_nonfin><A>, is het handig om deze van tevoren op te schrijven. <s8655, leaflets>
- <A><advcl\_purp\_nonfin>Not to forget any questions during the interview<advcl\_purp\_nonfin><A>, it is useful to write these down beforehand.)
- (26) <A><advcl\_verbless>With the Reverend Ian Paisley's DUP now the largest party in unionism<advcl\_verbless><A>, he said Sinn Fein was exploring their position. <s515, newspaper articles>

**Table 15<sup>32</sup>      Frequencies of three types of non-finite clauses in the AC subpattern**

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Non-finite present participle	9	8	11	2	30 (41.7%)
Non-finite past participle	8	6	6	2	22 (30.5%)
Non-finite infinitive	4	2	0	6	12 (16.7%)
Non-finite verbless	0	4	0	4	8 (11.1%)
Total	21	20	17	14	72 (100%)
<b>Dutch</b>					
Non-finite present participle	3	1	6	0	10 (24.4%)
Non-finite past participle	8	0	3	4	15 (36.6%)
Non-finite infinitive	1	8	1	5	15 (36.6%)
Non-finite verbless	1	0	0	0	1 (2.4%)
Total	13	9	10	9	41 (100%)

The further subcategorisation of the various types of non-finite clauses and verbless clauses shows that frequencies for the different types vary across the genres. In English, the non-finite clauses with a present participle form the largest group in the academic prose genre, the newspaper genre and the short stories genre. In Dutch, the non-finite clauses with a past participle and an infinitive form the largest groups, the former type occurring predominantly in the academic prose genre and the latter predominantly in the newspaper genre.

<sup>32</sup> In cases where column totals are below 40, only raw figures will be provided and no percentages, with the exception of the column that contains the totals.

### Grammatical realisation of C in the AC subpattern

Similar to the make-up of the C in the main C-pattern, the C in the AC subpattern can either consist of a single, uncoordinated nucleus or of coordinated nuclei. The frequencies of either type are presented in Table 16.

**Table 16** Make-up of C in the AC subpattern

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
C (single, uncoordinated unit)	309 (83.7%)	256 (91.1%)	221 (73.2%)	199 (85.4%)	985 (83.1%)
Ca/b (coordinated nuclei)	60 (16.3%)	25 (8.9%)	81 (26.8%)	34 (14.6%)	200 (16.9%)
Total	369 (100%)	281 (100%)	302 (100%)	233 (100%)	1185 (100%)
<b>Dutch</b>					
C (single, uncoordinated unit)	426 (88.4%)	393 (89.7%)	398 (85.4%)	508 (92.5%)	1725 (89.1%)
Ca/b (coordinated nuclei)	56 (11.6%)	45 (10.3%)	68 (14.6%)	41 (7.5%)	210 (10.9%)
Total	482 (100%)	438 (100%)	466 (100%)	549 (100%)	1935 (100%)

Even though the loglinear analysis just failed to reach significance for the three-way interaction between language, genre and make-up of C in the AC subpattern at the predetermined alpha level of .01 ( $p = .019$ ), it did show a significant two-way interaction between language and make-up of C ( $\chi^2(1) = 20.36$ ,  $p < .001$ , Cramer's  $V = .08$ ). The subtle differences between English and Dutch can be found in the higher frequency of coordinated nuclei in English than in Dutch (200 (16.9%); 210 (10.9%)). In both languages, however, the vast majority of sentences with the AC subpattern contain a single, uncoordinated nucleus (English: 1185 (83.1%); Dutch: 1935 (89.1%)).

Similar to the C in the C-pattern, the uncoordinated C can be realised as an independent clause or a non-independent clause. However, in the AC subpattern the grammatical realisation of C as a non-independent clause hardly occurs. There are no instances of non-independent clauses in the academic prose genre in either language, and only a few in the newspaper genre (English: 3 (1.2%) Dutch: 3 (0.8%)) and the leaflets genre (English: 2 (1.0%); Dutch: 6 (1.2%)). The short stories genre contains a few more instances of non-independent clauses (English: 19 (8.8%); Dutch: 12 (3.1%)). In both languages the vast majority of sentences with the AC subpattern are realised as independent clauses.

Table 17 presents the grammatical realisation of the C when this consists of coordinated nuclei. These coordinated nuclei either take the form of independent clauses or of subclauses, fragments or phrases.

**Table 17 Grammatical realisation of C coordinated nuclei in the AC subpattern**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper article</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Coordinated indep. clauses	39 (65.0%)	18 (72.0%)	69 (85.2%)	30 (88.2%)	156 (78.0%)
Coordinated subcl/phrases/frag	21 (35.0%)	7 (28.0%)	12 (14.8%)	4 (11.8%)	44 (22.0%)
<b>Total</b>	<b>60 (100%)</b>	<b>25 (100%)</b>	<b>81 (100%)</b>	<b>34 (100%)</b>	<b>200 (100%)</b>
<b>Dutch</b>					
Coordinated indep. clauses	42 (75.0%)	40 (88.9%)	64 (94.1%)	34 (82.9%)	180 (85.7%)
Coordinated subcl/phrases/frag	14 (25.0%)	5 (11.1%)	4 (5.9%)	7 (17.1%)	30 (14.3%)
<b>Total</b>	<b>56 (100%)</b>	<b>45 (100%)</b>	<b>68 (100%)</b>	<b>41 (100%)</b>	<b>210 (100%)</b>

A loglinear analysis showed no significant three-way interaction between language, genre and grammatical realisation of coordinated Cs in the AC subpattern ( $\chi^2(3)=3.71$ ,  $p=.29$ ). Nor did the analysis show a significant two-way interaction between language and grammatical realisation at the predetermined alpha level of .01 ( $\chi^2(1)=4.42$ ,  $p=.035$ , Cramer's  $V=.10$ ). A difference in frequency in the distribution of these grammatical realisations was only found between the different genres, irrespective of the languages ( $\chi^2(3)=17.08$ ,  $p<.001$ , Cramer's  $V=.20$ ). In all genres the coordinated C is typically realised as an independent clause. For the academic prose genre the frequencies of the coordinated C realised as non-independent clause are highest when compared to the other genres (35 (30.1%)).

### 5.5.2 Summary

The X-element in the XC pattern can consist of one, two or three elements. As the subpattern in which the X-element consists of one element is by far the most frequent one, the analysis in this chapter has been restricted to this subpattern. The two other subpatterns will be described in more detail in Chapter 6.

The analysis of the AC subpattern showed that the A-element takes the form of either a phrase or a clause. Even though this element typically takes the form of a phrase across the genres in both languages, differences between the

languages can be found in the frequencies of clauses in initial position. Specifically, the frequencies for clauses are higher in the English academic prose genre, the English newspaper genre and the English leaflets genre when compared to these genres in Dutch. The short stories genre is the only genre in which the frequencies of clauses in initial position are similar in both languages.

As for the grammatical realisation of these A-elements, when it takes the form of a phrase, the distribution of the different realisation groups differs significantly between the languages. Even though in both languages across all genres the phrasal A-element is typically realised as an adjunct or PP, this applies even more so to Dutch than to English. In the English academic prose genre, the leaflets genre and especially the short stories genre the phrase also frequently takes the form of a conjunct. In contrast, the Dutch newspaper genre shows a higher frequency of conjuncts when compared to the English newspaper genre.

In cases where the A-element takes the form of a clause, this can either be a finite or non-finite clause. In both languages most clauses are finite clauses, but non-finite clauses occur significantly more often in English than in Dutch, irrespective of genre. Of the different types of non-finite clause, the ones with a present participle occur most frequently in English, whereas most non-finite clauses in Dutch contain a past participle or infinitive. Moreover, as the vast majority of sentence-initial clauses are adverbial clauses, these have also been categorised on the basis of their semantic role. Adverbial clauses of concession are particularly frequent in the English academic prose genre and the newspaper genre, and although clauses of condition also belong to the most frequently occurring group in English, these are particularly frequent in the Dutch academic prose genre, the short stories genre and the leaflets genre. Adverbial clauses of time are particularly frequent in the short stories genre in both languages.

Finally, the nucleus in the AC subpattern predominantly takes the form of a single, uncoordinated nucleus in both languages, although the frequencies for coordinated nuclei are significantly higher in English than in Dutch. When uncoordinated, the nucleus is mainly realised as an independent clause, with the exception of the short stories genre in both languages, in which the frequencies of non-independent clauses are slightly higher. When coordinated, the nuclei are predominantly realised as independent clauses, with the exception of the academic prose genre in both languages, in which the frequency of non-independent clauses is slightly higher.

## 5.6 The CX pattern

Section 5.3.1 already showed that the frequencies for the CX pattern are remarkably similar for both languages, with 17.5% (1410) of all English sentences belonging to this pattern and 16.3% (1419) of all Dutch sentences. The only genre in which a subtle difference in frequency can be found is the leaflets genre, which shows a slightly higher frequency for English than for Dutch (14.5% vs. 11.1%).

In the CX pattern, the sentence consists of a nuclear unit that can be followed by one, two or three appended satellites. As the subpattern in which the X consists of one satellite, the CD subpattern, is by far the most frequent subpattern, this will be described in detail. The two other subpatterns will be described in less detail. Table 18 presents an overview of the three main subpatterns of the XC pattern, with examples of each type.

**Table 18** Subpatterns of the CX pattern

Make-up of X	Subpattern of CX	Example
X = 1 element	CD	<C>But it would probably embarrass the MoD<C> - <D>including Geoff Hoon<D>. <s15, newspaper genre>
X = 2 elements	CDE	<C>It could be argued that measures to reduce the infant and child mortality of the 'residuum' ... would merely be cancelled-out in greater morbidity and mortality rates at higher ages<C>, <D>as these unviable individuals hopelessly struggled to survive<D>, <E>clogging up the nation's labour market with enfeebled and inefficient 'stocks'<E>.<s4721, academic prose>
X = 3+ elements	CDEF(GH)	<C>Some people suffer from panic attacks<C>, <D>experiencing a rapid build-up of overpowering sensations<D>, <E>such as a pounding heart, faintness or shaky limbs<E>, <F>which make them fear that they are going mad, will black out or are having a heart attack<F>. <s7702, leaflets>

The frequencies of the three patterns in which C is followed by one, two or three elements are presented in Table 19.

**Table 19**                      **Frequencies of the CX pattern: X = 1, 2 or 3+ elements**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
CX – X= 1 element	163 (78.7%)	191 (86.4%)	581 (77.4%)	189 (81.8%)	1124 (79.7%)
CX – X = 2 elements	34 (16.4%)	28 (12.7%)	143 (19.0%)	35 (15.2%)	240 (17.0%)
CX – X = 3 elements	10 (4.8%)	2 (0.9%)	27 (3.6%)	7 (3.0%)	46 (3.3%)
<b>Total</b>	<b>207 (100%)</b>	<b>221 (100%)</b>	<b>751 (100%)</b>	<b>231 (100%)</b>	<b>1410 (100%)</b>
<b>Dutch</b>					
CX – X= 1 element	220 (85.6%)	192 (93.2%)	554 (76.2%)	200 (87.3%)	1166 (82.2%)
CX – X = 2 elements	31 (12.1%)	12 (5.8%)	124 (17.1%)	25 (10.9%)	192 (13.5%)
CX – X = 3 elements	6 (2.3%)	2 (1.0%)	49 (6.7%)	4 (1.7%)	61 (4.3%)
<b>Total</b>	<b>257 (100%)</b>	<b>206 (100%)</b>	<b>727 (100%)</b>	<b>229 (100%)</b>	<b>1419 (100%)</b>

The loglinear analysis failed to reach significance for the three-way interaction between language, genre and frequencies of the subpatterns of CX at the predetermined alpha level of .01 ( $p = .04$ ). Nor did it show a significant two-way interaction between language and distribution of the subpatterns of CX ( $\chi^2(2) = 8.14$ ,  $p = .017$ , Cramer's  $V = .05$ ).

Table 19 shows that the subpattern in which the X consists of one element is the most frequent subpattern in both languages (English: 1124 (79.7%); Dutch: 1166 (82.2%)). This is followed by the subpattern in which the X consists of two elements (English: 240 (17.0%); Dutch: 192 (13.5%)) and finally by the pattern in which the X consists of three or more elements (English: 46 (3.3%); Dutch: 61 (4.3%)).

### **5.6.1 The CD subpattern**

Table 19 showed that the pattern in which the nucleus is followed by one appended satellite is the most frequent of the three subpatterns of the CX pattern. In fact, it is one of the more frequent sentence patterns overall, as 14.0% (1124) of all English sentences in the corpus follow this pattern and 13.4% (1166) of all Dutch sentences.

Similar to the nuclei and satellites in the C-pattern and XC pattern, the nuclei and satellites in the CD subpattern can also either take the form of a single unit or of coordinated units. Cases in which both the nucleus and the appended satellite are coordinated are very few, i.e. only six sentences in English and six sentences in Dutch. Table 20 presents the distribution of uncoordinated and coordinated Cs.



**Table 20**      **Make-up of C in the CD subpattern**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper article</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
C (single unit)	141 (86.5%)	172 (90.1%)	488 (84.0%)	168 (88.9%)	969 (86.2%)
Ca/b (coordinated nuclei)	22 (13.5%)	19 (9.9%)	93 (16.0%)	21 (11.1%)	155 (13.8%)
<b>Total</b>	<b>163 (100%)</b>	<b>191 (100%)</b>	<b>581 (100%)</b>	<b>189 (100%)</b>	<b>1124 (100%)</b>
<b>Dutch</b>					
C (single unit)	210 (95.5%)	177 (92.2%)	490 (88.4%)	187 (93.5%)	1064 (91.3%)
Ca/b (coordinated nuclei)	10 (4.5%)	15 (7.8%)	64 (11.6%)	13 (6.5%)	102 (8.7%)
<b>Total</b>	<b>220 (100%)</b>	<b>192 (100%)</b>	<b>554 (100%)</b>	<b>200 (100%)</b>	<b>1166 (100%)</b>

The loglinear analysis showed no significant three-way interaction between language, genre and distribution of uncoordinated and coordinated nuclei in the CD subpattern ( $\chi^2(3) = 4.05$ ,  $p = .25$ ). The analysis did show a significant two-way interaction between language and make-up of C ( $\chi^2(1) = 13.50$ ,  $p < .001$ , Cramer's  $V = .08$ ). Although the vast majority of sentences that belong to the CD subpattern consist of a single, uncoordinated C (English: 969 (86.2%); Dutch: 1064 (91.3%)), the frequencies in which this unit is coordinated are significantly higher for English than for Dutch (155 (13.8%) vs. 102 (8.7%).

**Table 21**      **Make-up of D in the CD subpattern**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
D (single unit)	146 (89.6%)	185 (96.9%)	567 (97.6%)	168 (88.9%)	1066 (94.8%)
Da/b (coordinated nuclei)	17 (10.4%)	6 (3.1%)	14 (2.4%)	21 (11.1%)	58 (5.2%)
<b>Total</b>	<b>163 (100%)</b>	<b>191 (100%)</b>	<b>581 (100%)</b>	<b>189 (100%)</b>	<b>1124 (100%)</b>
<b>Dutch</b>					
D (single unit)	207 (94.1%)	183 (95.3%)	512 (92.4%)	186 (93.0%)	1088 (93.3%)
Da/b (coordinated nuclei)	13 (5.9%)	9 (4.7%)	42 (7.6%)	14 (7.0%)	78 (6.7%)
<b>Total</b>	<b>220 (100%)</b>	<b>192 (100%)</b>	<b>554 (100%)</b>	<b>200 (100%)</b>	<b>1166 (100%)</b>

Table 21 presents the distribution of uncoordinated and coordinated Ds. The loglinear analysis showed a significant three-way interaction between language, genre and distribution of uncoordinated and coordinated D-satellites in the CD subpattern ( $\chi^2(3) = 19.77$ ,  $p < .001$ ). Subsequent chi-square tests showed that this three-way interaction is mainly caused by differences in frequencies of uncoordinated and coordinated Ds between the English and Dutch short stories genre ( $\chi^2(1) = 16.17$ ,  $p < .001$ , Cramer's  $V = .11$ ). No differences between English and

Dutch were found for the academic prose genre ( $\chi^2(1)= 2.65, p= .10, \text{Cramer's } V=.08$ ), the newspaper genre ( $\chi^2(1)= .60, p =.43, \text{Cramer's } V=.04$ ) and the leaflets genre ( $\chi^2(1)= 2.00, p =.15, \text{Cramer's } V=.07$ ).

In both languages, across the genres, the D-satellite consists of a single, uncoordinated unit in the vast majority of cases (English: 1066 (94.8%); Dutch: 1088 (93.3%)). The only genre in which the distribution of uncoordinated and coordinated Ds is significantly different between English and Dutch is the short stories genre, with Dutch containing a higher frequency of coordinated Ds than English (English: 14 (2.4%); Dutch: 42 (7.6%)).

### Grammatical realisation of C in the CD subpattern

Table 22 presents the grammatical realisation of the C when it consists of a single nucleus. This nucleus is realised as either an independent clause or a non-independent clause.

**Table 22** Grammatical realisation of C as a single, uncoordinated unit in the CD subpattern

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Independent clause	135 (95.7%)	168 (97.7%)	364 (74.6%)	146 (86.9%)	813 (83.9%)
Non-independent clause	6 (4.3%)	4 (2.3%)	124 (25.4%)	22 (13.1%)	156 (16.1%)
Total	141 (100%)	172 (100%)	488 (100%)	168 (100%)	969 (100%)
<b>Dutch</b>					
Independent clause	202 (96.2%)	166 (93.8%)	325 (66.3%)	142 (75.9%)	835 (78.5%)
Non-independent clause	8 (3.8%)	11 (6.2%)	165 (33.7%)	45 (24.1%)	229 (21.5%)
Total	210 (100%)	177 (100%)	490 (100%)	187 (100%)	1064 (100%)

The loglinear analysis showed no significant three-way interaction between language, genre and distribution of the grammatical realisation of C as either independent or non-independent clause in the CD subpattern ( $\chi^2(3)= 3.21, p =.35$ ). The analysis did show a significant two-way interaction between language and distribution of independent and non-independent clauses ( $\chi^2(1)= 15.31, p <.001, \text{Cramer's } V=.07$ ), with Dutch containing more Cs that take the form of a non-independent clause. However, it should be noted that there is a two-way interaction that is much more salient, which is the interaction between genre and grammatical realisation of C ( $\chi^2(3)= 171.57, p <.001, \text{Cramer's } V=.29$ ). Irrespective of language, the leaflets genre

and especially the short stories genre show more instances of Cs that are realised as non-independent clauses than the academic prose and newspaper genres. As this study focuses on the differences between English and Dutch, differences between the genres irrespective of language will not be further investigated.

Even though the vast majority of Cs in the CD subpattern are realised as independent clauses in both languages, Dutch shows a higher frequency of Cs taking the form of non-independent clauses than English (English: 156 (16.1%); Dutch: 229 (21.5%)). In both languages, the short stories genre and leaflets genre show the highest frequencies of Cs realised as non-independent clauses. In the short stories genre these non-independent clauses occur predominantly in the simulated dialogue sections, in which the nuclear unit takes the form of an answer to a question, a discourse marker or a vocative. In the leaflets genre the function of the fragments is often to introduce a list of phrases or clauses. These types of fragments have received a particular label, *fragment\_complement\_list*, as these lists often form the complement of an incomplete nuclear sentence (cf. 3.5.1). Sentence (27) presents an example of this particular type of fragment.

(27) <C><fragment\_comp\_list>De bekendste zijn<fragment\_comp\_list><C>:  
<D><appos\_NP\_list>nachtmerries, slaapwandelen en praten of  
tandenknarsen in de slaap<appos\_NP\_list><D>. <s8420, leaflets>

(<C><fragment\_comp\_list>The most familiar ones are<fragment\_comp\_list><C>:  
<D><appos\_NP\_list>nightmares, sleep walking and talking or teeth grinding  
during one's sleep<appos\_NP\_list><D>.)

Table 23 presents the grammatical realisation of the C when this consists of coordinated nuclei. These coordinated nuclei either take the form of independent clauses or of subordinate clauses, fragments or phrases.

**Table 23** Grammatical realisation of C coordinated nuclei in the CD subpattern

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Coordinated indep. Clauses	14	16	80	16	126 (81.3%)
Coordinated subcl/phrases/frag	8	3	13	5	29 (18.7%)
Total	22	19	93	21	155 (100%)
<b>Dutch</b>					
Coordinated indep. Clauses	8	13	53	9	83 (81.4%)
Coordinated subcl/phrases/frag	2	2	11	4	19 (18.6%)
Total	10	15	64	13	102 (100%)

The loglinear analysis showed no significant three-way interaction between language, genre and distribution of the grammatical realisation of coordinated Cs as either independent or non-independent clauses ( $\chi^2(3) = 1.36$ ,  $p = .71$ ). Nor did the analysis show a significant two-way interaction between language and distribution of independent and non-independent clauses ( $\chi^2(1) = .02$ ,  $p = .88$ , Cramer's  $V = .00$ ). In both languages these coordinated Cs typically take the form of coordinated independent clauses (English (126 (81.3%); Dutch: 83 (81.4%)).

#### Grammatical realisation of D in the CD subpattern

When the D-satellite in the CD subpattern consists of a single, uncoordinated satellite, it can take the form of a phrase or a clause. The frequencies of both the phrases and clauses are presented in Table 24.

**Table 24** Grammatical realisation of D as a single, uncoordinated unit in the CD subpattern

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
D = phrase	49 (33.6%)	62 (33.5%)	204 (36.0%)	68 (40.5%)	383 (35.9%)
D = clause	97 (66.4%)	123 (66.5%)	363 (64.0%)	100 (59.5%)	683 (64.1%)
Total	146 (100%)	185 (100%)	567 (100%)	168 (100%)	1066 (100%)
<b>Dutch</b>					
D = phrase	89 (43.0%)	41 (22.4%)	160 (31.3%)	112 (60.2%)	402 (36.9%)
D = clause	118 (57.0%)	142 (77.6%)	352 (68.8%)	74 (39.8%)	686 (63.1%)
Total	207 (100%)	183 (100%)	512 (100%)	186 (100%)	1088 (100%)

The loglinear analysis showed a significant three-way interaction between language, genre and grammatical realisation of the D-satellite in the CD subpattern ( $\chi^2(3)=25.23$ ,  $p < .001$ ). Subsequent chi-square tests showed that this three-way interaction is mainly caused by differences in the distribution of phrases and clauses between the English and Dutch leaflets genre ( $\chi^2(1)=13.76$ ,  $p < .001$ , Cramer's  $V=.19$ ). No differences between English and Dutch were found for the academic prose genre ( $\chi^2(1)=3.20$ ,  $p = .07$ , Cramer's  $V=.09$ ), the newspaper genre ( $\chi^2(1)=5.63$ ,  $p = .02$ , Cramer's  $V=.12$ ) and the short stories genre ( $\chi^2(1)=2.69$ ,  $p = .10$ , Cramer's  $V=.05$ ).

The difference in distribution of phrases and clauses in the leaflets genre between the languages is that in English the majority of sentences take the form of a clause (English: 100 (59.5%); Dutch: 74 (39.8%)), whereas in Dutch the majority of sentences take the form of a phrase (English 68 (40.5%); Dutch: 112 (60.2%)). In the other genres the languages follow a similar pattern, with the D-satellite more often taking the form of a clause than a phrase. Specifically, in the English academic prose genre 97 (66.4%) of the sentences take the form of a clause, compared to 118 (57.0%) sentences in Dutch. For the newspaper genre, this is 123 (66.5%), compared to 142 (77.6%) in Dutch, and for the short stories genre this is 363 (64.0%) in English and 352 (68.8%) in Dutch.

Both the phrasal and clausal element can be further subcategorised. As for the phrasal element, this can be further subcategorised and specified into five main realisation groups. It can take the form of (1) an adjunct or prepositional phrase, (2) an apposition, (3) a phrasal fragment, such as an NP, (4) a discourse marker, vocative, tag or the words *yes* or *no*, or (5) a conjunct or disjunct. Sentences (28) to (32) provide examples of each of the grammatical forms that the appended satellite D can take when it is realised as a phrase.

(28) <Ca>We namen allerhartelijkst afscheid<Ca> <Cb>en ik haastte me het kille steegje uit<Cb>, <D><pp>naar de warme middagzon<pp><D>. <s13685, short stories>

(<Ca>We said our warmest goodbyes<Ca> <Cb>and I quickly left the cold alley<Cb>, <D><pp>to the warm afternoon sun<pp><D>.)

(29) <C>Sharing laughter with others reveals both cultural and emotional attunement with them<C>, <D><appos\_NP\_list>namely a mutuality of interest

in the topic of the laughter and or an interest in the laughter of the others as an affective state in its own right<appos\_NP\_list><D>. <s5817, academic prose>

(30) <C>Laat-ie naar z'n eigen kromme poten kijken<C>, <D><fragment\_NP>de eikel<fragment\_NP><D>! <s14329, short stories>

(<C>He should look at his own crooked legs<C>, <D><fragment\_NP>the jerk<fragment\_NP><D>!)</p>
</div>

(31) <C>He's six years old<C>, <D><dm>for Christ's sake<dm><D>. <s10561, short stories>

(32) <C>The general feeling is that now is not the time to suggest this<C>, <D><conj>however<conj><D>. <s1731, newspaper articles>

The frequencies of these five main realisation groups are provided in Table 25.

**Table 25** Frequencies of grammatical realisations of D as a phrase in the CD subpattern in five realisation groups

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
D = adjunct/PP	19 (38.8%)	23 (37.1%)	44 (21.6%)	17 (25.0%)	103 (26.9%)
D = apposition	29 (59.2%)	31 (50.0%)	58 (28.4%)	40 (58.8%)	158 (41.3%)
D = fragment_phrase	0 (0.0%)	1 (1.6%)	12 (5.9%)	9 (13.2%)	22 (5.7%)
D = dis.mark, vocative, tag, yes/no	0 (0.0%)	1 (1.6%)	77 (37.7%)	0 (0.0%)	78 (20.4%)
D = conjunct/disjunct	1 (2.0%)	6 (9.7%)	13 (6.4%)	2 (2.9%)	22 (5.7%)
Total	49 (100%)	62 (100%)	204 (100%)	68 (100%)	383 (100%)
<b>Dutch</b>					
D = adjunct/PP	38 (42.7%)	20 (48.8%)	64 (40.0%)	22 (19.6%)	144 (35.8%)
D = apposition	49 (55.1%)	19 (46.3%)	39 (24.4%)	81 (72.3%)	188 (46.8%)
D = fragment_phrase	2 (2.2%)	2 (4.9%)	30 (18.8%)	9 (8.0%)	43 (10.7%)
D = dis.mark, vocative, tag, yes/no	0 (0.0%)	0 (0.0%)	22 (13.8%)	0 (0.0%)	22 (5.5%)
D = conjunct/disjunct	0 (0.0%)	0 (0.0%)	5 (3.1%)	0 (0.0%)	5 (1.2%)
Total	89 (100%)	41 (100%)	160 (100%)	112 (100%)	402 (100%)

Due to the fact that certain realisation forms are genre-specific, it causes some cells in Table 25 to have small expected frequencies, which makes it difficult to test differences in frequencies statistically. For this reason, differences in realisation patterns will only be described.

In both the English and the Dutch academic genre, the vast majority of phrases are either realised as appositions or as adjuncts/PPs. In both languages, the group of appositions forms the largest realisation group (English: 29 (59.2%); Dutch:

49 (55.1%)), followed by the group of adjuncts/PPs (English: 19 (38.8%); Dutch: 38 (42.7%)).

This pattern is rather similar for the newspaper genre. In English, the group of appositions again forms the largest realisation group (31 (50.0%)), followed by adjuncts/PPs (23 (37.1%)). In Dutch, the vast majority of phrases are also realised as appositions and adjuncts/PPs, but here the latter type forms the largest group (20 (48.8%)), followed by the former (19 (46.3%)). In English, but not in Dutch, the D-satellite is realised as a conjunct or disjunct in 6 (9.7%) sentences.

In the short stories genre, the languages show some differences in realisation patterns. In English, the largest group is formed by discourse markers, vocatives and tags (77 (37.7%)), followed by appositions (58 (28.4%)) and adjunct/PPs (44 (21.6%)). In 13 (6.4%) sentences it takes the form of a conjunct or disjunct. In Dutch, on the other hand, the adjuncts/PPs form the largest realisation group (64 (40.0%)), followed by appositions (39 (24.4%)). It is realised as a phrasal fragment in 30 (18.8%) sentences and as a discourse marker in 22 (13.8%). Compared to the other genres, this genre shows the widest variety of realisation forms in both languages.

Finally, the leaflets genre follows the same pattern as the academic and newspaper genre, with the appositions again forming the largest realisation group in both languages, especially in Dutch (81 (72.3%); English: 40 (58.8%)). This is followed by the group of adjuncts/PPs (English: 17 (25.0%); Dutch: 22 (19.6%)).

When realised as a clause, this can take various grammatical forms. The clauses have been categorised into six realisation groups: (1) appended clauses, (2) non-restrictive relative clauses, (3) adverbial clauses, (4) reporting clauses, (5) independent clauses and (6) clause fragments. Sentences (33) to (37) provide examples of each type, with the exception of adverbial clauses, as these will be further exemplified below.

(33) <C>Not a day goes by without an attack on US troops or civilians<C> -  
<D><appcl>and usually both<appcl><D>. <s352, newspaper articles>

(34) <C>Er zijn in Nederland veel verschillende dialecten<C>,  
<D><nonrestr\_relcl>die vaak een duidelijk hoorbaar accent in de Nederlandse spraak achterlaten<nonrestr\_relcl><D>. <s6642, academic prose>

(C>There are many different dialects in the Netherlands<C>, <D><nonrestr\_relcl>which often leave a clearly audible accent in Dutch speech<nonrestr\_relcl><D>.)

(35) <C>There is an obvious logic to what the Prime Minister said in the House<C>, <D><reportingcl><fragment>he added<fragment><reportingcl><D>. <s902, newspaper articles>

(36) <C>Iedereen is het er over eens<C>: <D><indepcl>roken en meeroken zijn de meest vermijdbare oorzaak van ziekten en sterfte in Nederland<indepcl><D>. <s9694, leaflets>

(<C>Everyone agrees about it<C>: <D><indepcl>smoking and passive smoking are the most avoidable causes of diseases and death in the Netherlands<indepcl><D>.)

(37) <C>You really wouldn't want to be in his shoes<C>, <D><commcl\_fragment>I'll tell you<commcl\_fragment><D>. <s12061, short stories>

Table 26 presents the frequencies of the six relations groups that the clausal D can take.

**Table 26**                      **Frequencies of grammatical realisations of D as a clause in the CD subpattern in six realisation groups**

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
D = appended clause	2 (2.1%)	23 (18.7%)	1 (0.3%)	1 (1.0%)	27 (4.0%)
D = nonrestrictive rel clause	16 (16.5%)	17 (13.8%)	22 (6.1%)	18 (18.0%)	73 (10.7%)
D = adverbial clause	70 (72.2%)	50 (40.7%)	166 (45.7%)	46 (46.0%)	332 (48.6%)
D = reporting clause	0 (0.0%)	27 (22.0%)	148 (40.8%)	0 (0.0%)	175 (25.6%)
D = independent clause	8 (8.2%)	6 (4.9%)	18 (5.0%)	35 (35.0%)	67 (9.8%)
D = clause fragment	1 (1.0%)	0 (0.0%)	8 (2.2%)	0 (0.0%)	9 (1.3%)
Total	97 (100%)	123 (100%)	363 (100%)	100 (100%)	683 (100%)
<b>Dutch</b>					
D = appended clause	2 (1.7%)	14 (9.9%)	2 (0.6%)	0 (0.0%)	18 (2.6%)
D = nonrestrictive rel clause	36 (30.5%)	39 (27.5%)	27 (7.7%)	22 (29.7%)	124 (18.1%)
D = adverbial clause	47 (39.8%)	48 (33.8%)	74 (21.0%)	29 (39.2%)	198 (28.9%)
D = reporting clause	3 (2.5%)	31 (21.8%)	143 (40.6%)	0 (0.0%)	177 (25.8%)
D = independent clause	30 (25.4%)	10 (7.0%)	103 (29.3%)	22 (29.7%)	165 (24.1%)
D = clause fragment	0 (0.0%)	0 (0.0%)	3 (0.9%)	1 (1.4%)	4 (0.6%)
Total	118 (100%)	142 (100%)	352 (100%)	74 (100%)	686 (100%)



Due to the detailed level of subcategorisation, the expected frequencies of certain cells are too low to test differences in realisation patterns between the languages statistically. For this reason, differences in realisation patterns will only be described.

The academic prose genre in English and Dutch shows differences in realisation patterns. In English, most clausal Ds take the form of adverbial clauses (70 (72.2%)), followed by non-restrictive relative clauses (16 (16.5%)). In a small number of cases, it takes the form of an independent clause (8 (8.2%)). In Dutch, on the other hand, the group of independent clauses forms a relatively large group (30 (25.4%)). However, the largest group is also formed by adverbial clauses (47 (39.8%)), followed by non-restrictive relative clauses (36 (30.5%)).

In the newspaper genre, the group of adverbial clauses again forms the largest realisation group in both languages (English: 50 (40.7%); Dutch: 48 (33.8%)). In English, the second largest group is formed by reporting clauses (27 (22.0%)) and the third by appended clauses (23 (18.7%)). In Dutch, the second largest group is formed by non-restrictive relative clauses (39 (27.5%)) and the third by reporting clauses (31 (21.8%)).

In both the English and Dutch short stories genre the group of reporting clauses forms a large realisation group (English: 148 (40.8%); Dutch: 143 (40.6%)). In English, another very large group is formed by adverbial clauses (166 (45.7%)), whereas in Dutch a large group is formed by, again, independent clauses (103 (29.3%)), but also by adverbial clauses (74 (21.0%)).

Finally, in the leaflets genre the largest realisation group are the adverbial clauses in both languages (English: 46 (46.0%); Dutch: 29 (39.2%)). This is followed by a large group of independent clauses, also in both languages (English: 35 (35.0%); Dutch: 22 (29.7%)). The third largest group is formed by non-restrictive relative clauses, although this has higher frequencies in Dutch than in English (English: 18 (18.0%); Dutch: 22 (29.7%)).

Another way to classify clausal D-satellites is by dividing them into finite or non-finite clauses. Table 27 provides the frequencies of both groups.

**Table 27**      **Frequencies of D as a clause in the CD subpattern as finite and non-finite clauses**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Adverbial cl_finite	32 (45.7%)	26 (52.0%)	45 (27.1%)	31 (67.4%)	134 (40.4%)
Adverbial cl_non-finite	38 (54.3%)	24 (48.0%)	121 (72.9%)	15 (32.6%)	198 (59.6%)
<b>Total</b>	<b>70 (100%)</b>	<b>50 (100%)</b>	<b>166 (100%)</b>	<b>46 9100%</b>	<b>332 (100%)</b>
<b>Dutch</b>					
Adverbial cl_finite	42 (89.4%)	43 (89.6%)	56 (75.7%)	24 (82.8%)	165 (83.3%)
Adverbial cl_non-finite	5 (10.6%)	5 (10.4%)	18 (24.3%)	5 (17.2%)	33 (16.7%)
<b>Total</b>	<b>47 (100%)</b>	<b>48 (100%)</b>	<b>74 (100%)</b>	<b>29 (100%)</b>	<b>198 (100%)</b>

The loglinear analysis showed no significant three-way interaction between language, genre and distribution of finite and non-finite clauses ( $\chi^2(3) = 4.71$ ,  $p = .19$ ). The analysis did show a significant two-way interaction between language and distribution of finite and non-finite clauses ( $\chi^2(1) = 92.10$ ,  $p < .001$ , Cramer's  $V = .41$ ). The difference in distribution of finite and non-finite clauses across all genres is that the group of non-finite clauses is much larger in English than in Dutch (198 (59.6%) vs. 33 (16.7%)).

The non-finite clauses can be of three different types, similar to the non-finite clauses in the AC subpattern above. The non-finite verb can take the form of a present participle, a past participle or an infinitive (see examples (23), (24), (25) and (26) above). The group of verbless clauses have also been added to this group. The frequencies of each of these types are presented in Table 28.

**Table 28**      **Frequencies of four types of non-finite clauses in the CD subpattern**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Non-finite –present participle	25	12	91	11	139 (70.2%)
Non-finite –past participle	11	10	19	2	42 (21.2%)
Non-finite –infinitive	2	2	7	1	12 (6.1%)
Non-finite –verbless	0	0	4	1	5 (2.5%)
<b>Total</b>	<b>38</b>	<b>24</b>	<b>121</b>	<b>15</b>	<b>198 (100%)</b>
<b>Dutch</b>					
Non-finite –present participle	1	0	6	0	7 (21.2%)
Non-finite –past participle	3	3	10	4	20 (60.6%)
Non-finite –infinitive	1	2	1	1	5 (15.2%)
Non-finite –verbless	0	0	1	0	1 (3.0%)
<b>Total</b>	<b>5</b>	<b>5</b>	<b>18</b>	<b>5</b>	<b>33 (100%)</b>

A further categorisation of the various types of non-finite clauses and verbless clause shows that the non-finite clauses with a present participle form by far the largest group in English across all genres, except for the newspaper genre, in which only half of the sentences have a present participle (12 (50.0%)). The second largest group in English is formed by the non-finite clauses with a past participle, especially in the academic prose genre (11 (28.9%) and the newspaper genre (10 (41.7%)).

In Dutch the largest group is formed by the past participle clauses in general (20 (60.6%)), with the short stories genre also showing a few instances of non-finite clauses with a present participle (6 (33.3%)).

The finite adverbial clauses can also be further subcategorised into various semantic classes. The group labelled 'not specified' predominantly contains adverbial clauses introduced by *as* or *zoals*, which can have various functions. The frequencies of the different types of adverbial clauses are provided in Table 29.

**Table 29**                      **Frequencies of semantic roles of adverbial clauses – D in the CD subpattern**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Adverbial cl_time	1	6	12	5	24 (17.9%)
Adverbial cl_condition	2	4	2	4	12 (9.0%)
Adverbial cl_concession	19	9	9	9	46 (34.3%)
Adverbial cl_comparison	0	0	8	0	8 (6.0%)
Adverbial cl_reason	2	2	3	7	14 (10.5%)
Adverbial cl_not specified	5	4	7	4	20 (14.9%)
Adverbial cl_purpose	2	0	0	1	3 (2.2%)
Adverbial cl_result	1	1	4	1	7 (5.2%)
<b>Total</b>	<b>32</b>	<b>26</b>	<b>45</b>	<b>31</b>	<b>134 (100%)</b>
<b>Dutch</b>					
Adverbial cl_time	10	4	22	3	44 (26.7%)
Adverbial cl_condition	4	4	2	7	17 (10.3%)
Adverbial cl_concession	10	4	4	3	17 (10.3%)
Adverbial cl_comparison	1	0	15	0	16 (9.7%)
Adverbial cl_reason	14	11	8	10	43 (26.0%)
Adverbial cl_not specified	3	13	3	1	20 (12.1%)
Adverbial cl_purpose	0	4	2	0	6 (3.7%)
Adverbial cl_result	0	2	0	0	2 (1.2%)
<b>Total</b>	<b>42</b>	<b>43</b>	<b>56</b>	<b>24</b>	<b>165 (100%)</b>

Despite the fact that it is more difficult to describe trends in such a detailed level of subcategorisation, certain genres do seem to contain more instances of one type of adverbial clause than another.

In the English academic prose genre, adverbial clauses of concession form the largest group, whereas in Dutch this is formed by adverbial clauses of reason, followed by adverbial clauses of concession and time.

In the English newspaper genre, the largest group is again that of adverbial clauses of concession, followed by adverbial clauses of time. In Dutch the largest group is formed by non-specified adverbial clauses, followed by adverbial clauses of reason.

In the English short stories genre, adverbial clauses of time form the largest group, closely followed by adverbial clauses of concession and comparison. This pattern is similar in Dutch, in which the largest group is also formed by adverbial clauses of time. These are, however, followed by adverbial clauses of comparison and then reason.

Finally, in the English leaflets genre, the adverbial clauses of concession again form the largest realisation group, followed by clauses of reason, then time and condition. In Dutch, the largest group are adverbial clauses of reason, followed by condition.

Finally, we turn to the grammatical realisation of the D when this consists of coordinated units. These coordinated units either take the form of independent clauses, or of subordinate clauses, fragments or phrases, the frequencies of which are presented in Table 30.

**Table 30 Grammatical realisation of coordinated D in the CD subpattern**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Coordinated indep. clauses	1	2	5	6	14 (24.1%)
Coordinated subcl/phrases/frag	16	4	9	15	44 (75.9%)
<b>Total</b>	<b>17</b>	<b>6</b>	<b>14</b>	<b>21</b>	<b>58 (100%)</b>
<b>Dutch</b>					
Coordinated indep. clauses	8	4	30	6	48 (61.5%)
Coordinated subcl/phrases/frag	5	5	12	8	30 (38.5%)
<b>Total</b>	<b>13</b>	<b>9</b>	<b>42</b>	<b>14</b>	<b>78 (100%)</b>

A loglinear analysis showed no significant three-way interaction between language, genre and grammatical realisation of coordinated Ds in the CD subpattern ( $\chi^2(3)=4.74$ ,  $p=.19$ ). The analysis did show a significant two-way interaction between language and grammatical realisation of coordinated Ds ( $\chi^2(1)=13.17$ ,  $p<.001$ , Cramer's  $V=.37$ ). The main difference between English and Dutch is that the Ds in English are more frequently realised as coordinated phrases and subordinate clauses (44 (75.7%)), whereas in Dutch these are more frequently realised as coordinated independent clauses (48 (61.5%)).

### 5.6.2 The CDE subpattern

The second largest subpattern of the CX pattern, the CDE subpattern, will be described in a similar fashion to the CD subpattern. Compared to the CD subpattern, the CDE subpattern occurs much less frequently, with only 17.0% (240) of all English sentences and 13.5% (192) of all Dutch sentences that belong to the CX main pattern following this pattern.

Similar to the nucleus and satellites in the CD subpattern, the nucleus and satellites in the CDE subpattern can also either take the form of uncoordinated units or of coordinated units. Note that there are no sentences in which all three units, CDE, consist of coordinated elements, and there are only three sentences in English and one sentence in Dutch in which both the D and the E units consist of coordinated elements. Table 31 presents the distributions of single and coordinated Cs.

**Table 31** Make-up of C in the CDE subpattern

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
C (single unit)	34	28	124	31	217 (90.4%)
Ca/b (coordinated nuclei)	0	0	19	3	23 (9.6%)
Total	34	28	143	35	240 (100%)
Dutch					
C (single unit)	28	11	106	23	168 (87.5%)
Ca/b (coordinated nuclei)	3	1	18	2	24 (12.5%)
Total	31	12	124	25	192 (100%)

Due to the small expected frequencies of a number of cells, differences in make-up of C cannot be tested statistically. What Table 31 shows is that the vast majority of

sentences with the CDE subpattern consist of a C that is a single, uncoordinated unit in both English and Dutch (English: 217 (90.4%); Dutch: 168 (87.5%). In English, the academic prose genre and the newspaper genre contain no instances of coordinated Cs at all, and short stories and leaflets contain a few, 19 (13.2%) and 3 (8.6%) sentences respectively. In Dutch the frequency of the coordinated units is also highest in the short stories genre (18 sentences, 14.5%). The other genres contain very few instances of coordinated units.

Table 32 presents the distributions of single and coordinated Ds.

**Table 32** Make-up of D in the CDE subpattern

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
D (single unit)	31	27	142	22	222 (92.5%)
Da/b (coordinated nuclei)	3	1	1	13	18 (7.5%)
Total	34	28	143	35	240 (100%)
<b>Dutch</b>					
D (single unit)	24	12	120	20	176 (91.7%)
Da/b (coordinated nuclei)	7	0	4	5	16 (8.3%)
Total	31	12	124	25	192 (100%)

Due to the small expected frequencies of a number of cells, differences in make-up of D cannot be tested statistically. Similar to the nucleus in the CDE subpattern, the D-satellite also takes the form of a single, uncoordinated unit in most cases in both languages (English: 222 (92.5%); Dutch: 176 (91.7%)). In English, the leaflets genre forms an exception to this, as over half of the sentences contain coordinated Ds (13 (37.1%)).

Table 33 presents the distributions of single and coordinated Es.

**Table 33** Make-up of E in the CDE subpattern

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
E (single unit)	32	26	138	30	226 (94.2%)
Ea/b (coordinated nuclei)	2	2	5	5	14 (5.8%)
Total	34	28	143	35	240 (100%)
<b>Dutch</b>					
E (single unit)	31	10	114	24	179 (93.3%)
Ea/b (coordinated nuclei)	0	2	10	1	13 (6.7%)
Total	31	12	124	25	192 (100%)

Due to the small expected frequencies of a number of cells, differences in make-up of E cannot be tested statistically. For the E-satellite in the CDE subpattern, the distribution between uncoordinated and coordinated Es is similar to the distributions between the uncoordinated and coordinated Cs and Ds. This means that in both English and Dutch most sentences with the CDE subpattern contain single, uncoordinated E satellites (English: 226 (94.2%); Dutch: 179 (93.2%)).

### Grammatical realisation of C in the CDE subpattern

Table 34 presents the grammatical realisation of the C when it consists of a single, uncoordinated nucleus. This nucleus is realised as either an independent clause or a non-independent clause.

**Table 34** Grammatical realisation of single, uncoordinated C in the CDE subpattern

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Independent clause	33	27	78	17	155 (71.4%)
Non-independent clause	1	1	46	14	62 (28.6%)
Total	34	28	124	31	217 (100%)
Dutch					
Independent clause	26	11	67	10	114 (67.9%)
Non-independent clause	2	0	39	13	54 (32.1%)
Total	28	11	106	23	168 (100%)

The loglinear analysis showed no significant three-way interaction between language, genre and realisation of C in the CDE subpattern ( $\chi^2(3) = 1.49$ ,  $p = .68$ ). Nor did it show a significant two-way interaction between language and realisation of C ( $\chi^2(1) = .16$ ,  $p = .67$ , Cramer's  $V = .03$ ). It did, however, show a significant two-way interaction between genre and distribution of the realisation of C ( $\chi^2(3) = 59.60$ ,  $p < .001$ , Cramer's  $V = .35$ ). Table 34 shows that both the academic prose genre and the newspaper genre in both languages contain barely any instances of Cs that are realised as non-independent clauses. The frequencies for the Cs realised as non-independent clauses are much higher in the short stories genre (English: 46 (37.1%); Dutch: 39 (36.8%)) and the leaflets genre (English: 14 (45.2%); Dutch: 13 (56.5%)) in both languages. Sentences (38) and (39) are examples of Cs realised as non-independent clauses in the leaflets and short stories genre respectively.

- (38) (<C><indepcl>Can you get infected your first time<indepcl><C>?<s7851, leaflets>  
<C><fragment\_yes>Yes<fragment\_yes><C>, <D><advcl\_cond>if your partner has a  
STD and you have unsafe sex<advcl\_cond><D>, <E><indepcl>then you can  
become infected<indepcl><E>. <s7852, leaflets>
- (39) <C><reportedcl><frag>Love you too<frag><reportedcl><C>, <D><reportingcl><frag>he  
says<frag><reportingcl><D>, <E><subcl\_advcl\_nonfin>crunching on a wedge of  
crisps<subcl\_advcl\_nonfin><E>. <s11032, short stories>

### Grammatical realisation of D in the CDE subpattern

When the D-satellite in the CDE subpattern consists of a single, uncoordinated nucleus, it can take the form of a phrase or a clause. The frequencies of both the phrases and clauses are presented in Table 35.

**Table 35** Grammatical realisation of D (single, uncoordinated unit) in the CDE subpattern

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
D = phrase	7	8	68	10	93 (41.9%)
D = clause	24	19	74	12	129 (58.1%)
Total	31	27	142	22	222 (100%)
<b>Dutch</b>					
D = phrase	12	4	69	15	100 (56.8%)
D = clause	12	8	51	5	76 (43.2%)
Total	24	12	120	20	176 (100%)

The loglinear analysis showed no significant three-way interaction between language, genre and grammatical realisation of the D-satellite in the CDE subpattern ( $\chi^2(3) = 3.22$ ,  $p = .35$ ). The analysis did show a significant two-way interaction between language and realisation of D ( $\chi^2(1) = 7.58$ ,  $p < .01$ , Cramer's  $V = .15$ ). Table 35 shows that the difference between English and Dutch in the realisation of D is that the number of Ds that take the form of a clause is higher in English than in Dutch (129 (58.1%) vs. 76 (43.2%)).

Similar to the further specification of phrases in the CD pattern in Table 25 above, the phrases in the D of the CDE pattern can also be further specified into five realisation groups. The frequencies for each of these groups are presented in Table 36.



**Table 36**                      **Frequencies of grammatical realisations of D as a phrase in the CDE subpattern in five realisation groups**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
D = adjunct/PP	2	5	13	3	23 (24.7%)
D = apposition	1	2	14	2	19 (20.4%)
D = fragment_phrase	0	1	5	2	8 (8.6%)
D = dis.mark, vocative,tag,yes/no	0	0	29	0	29 (31.2%)
D = conjunct/disjunct	4	0	7	3	14 (15.1%)
<b>Total</b>	<b>7</b>	<b>8</b>	<b>68</b>	<b>10</b>	<b>93 (100%)</b>
<b>Dutch</b>					
D = adjunct/PP	4	1	19	3	27 (27.0%)
D = apposition	8	2	16	9	35 (35.0%)
D = fragment_phrase	0	1	12	3	16 (16.0%)
D = dis.mark, vocative,tag,yes/no	0	0	21	0	21 (21.0%)
D = conjunct/disjunct	0	0	1	0	1 (1.0%)
<b>Total</b>	<b>12</b>	<b>4</b>	<b>69</b>	<b>15</b>	<b>100 (100%)</b>

Due to the fact that a number of realisation groups are genre-specific and thus have low overall frequencies, it is difficult to test differences in frequencies statistically. For this reason, differences in realisation patterns will only be described.

The English academic prose genre contains very few instances of Ds realised as phrases. Of the few occurrences, most take the form of a conjunct/disjunct (4 sentences). In Dutch, on the other hand, phrases typically take the form of an apposition (8 sentences) or an adjunct/PP (4 sentences).

Both the English and especially the Dutch newspaper genre contain very few instances of Ds realised as phrases. In English most of these take the form of an adjunct/PP (5 sentences).

In both the English and Dutch short stories genre, the largest group of phrasal Ds take the form of discourse markers (English: 29 (42.6%); Dutch: 21 (30.4%)). This genre contains examples of most other realisation groups, with adjunct/PPs, appositions and conjuncts/disjuncts forming the largest groups in English and adjuncts/PPs, appositions and fragments the largest groups in Dutch.

The leaflets genre also shows some variation in realisation of the D element. In English it takes the form of adjuncts/PPs, appositions, fragments or conjuncts/disjuncts. In Dutch most phrases are realised as appositions.

When realised as a clause, the D-satellite can also be further subcategorised into four realisation groups, the frequencies of which are provided in Table 37.

**Table 37**      **Frequencies of grammatical realisations of D as a clause in the CDE subpattern in four realisation groups**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
D = nonrestrictive rel clause	1	5	4	2	12 (9.3%)
D = adverbial clause	20	11	28	9	68 (52.7%)
D = reporting clause	1	3	34	0	38 (29.5%)
D = independent clause	2	0	8	1	11 (8.5%)
<b>Total</b>	<b>24</b>	<b>19</b>	<b>74</b>	<b>12</b>	<b>129 (100%)</b>
<b>Dutch</b>					
D = nonrestrictive rel clause	4	2	3	0	9 (11.8%)
D = adverbial clause	5	2	22	3	32 (42.1%)
D = reporting clause	0	3	9	0	12 (15.8%)
D = independent clause	3	1	17	2	23 (30.3%)
<b>Total</b>	<b>12</b>	<b>8</b>	<b>51</b>	<b>5</b>	<b>76 (100%)</b>

Due to the detailed overview of realisation possibilities of the clausal D, the expected values of certain cells are too low to test differences in realisation patterns between the languages statistically. For this reason, differences in realisation patterns will only be described.

In the English academic prose genre most clausal Ds are realised as adverbial clauses (20 (83.3%)). The few occurrences of clausal Ds in the Dutch academic prose genre take the form of a non-restrictive relative clause (4 sentences), an adverbial clause (5 sentences) or an independent clause (3 sentences).

In the English newspaper genre most clausal Ds again take the form of an adverbial clause (11 (57.8%)). The other Ds are realised as nonrestrictive clauses or reporting clauses. In Dutch, the number for clausal Ds is lower and the frequencies for the various realisation groups are evenly divided across the groups.

In the English short stories genre, most clausal Ds take the form of a reporting clause (34 (45.9%)) or an adverbial clause (28 (37.8%)). In Dutch the largest realisation groups are formed by adverbial clauses (22 (43.1%)) and independent clauses (17 (33.3%)).

Finally, in the English leaflets genre the largest group is again formed by adverbial clauses (9 sentences). In Dutch the number for clausal Ds are much lower and the few instances take the form of adverbial clauses or independent clauses. The largest group of clauses, the adverbial clauses, can be further subcategorised into finite and non-finite clauses. Table 38 provides the frequencies of both groups.

**Table 38**      **Frequencies of D as a clause in the CDE subpattern as finite and non-finite clauses**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Adverbial cl_finite	12	5	12	5	34 (50.0%)
Adverbial cl_non-finite	8	6	16	4	34 (50.0%)
<b>Total</b>	<b>20</b>	<b>11</b>	<b>28</b>	<b>9</b>	<b>68 (100%)</b>
<b>Dutch</b>					
Adverbial cl_finite	4	2	19	2	27 (84.4%)
Adverbial cl_non-finite	1	0	3	1	5 (15.6%)
<b>Total</b>	<b>5</b>	<b>2</b>	<b>22</b>	<b>3</b>	<b>32 (100%)</b>

The loglinear analysis showed no significant three-way interaction between language, genre and distribution of finite and non-finite clauses ( $\chi^2(3) = 2.05$ ,  $p = .56$ ). The analysis did show a significant two-way interaction between language and distribution of finite and non-finite clauses ( $\chi^2(1) = 12.17$ ,  $p < .001$ , Cramer's  $V = .32$ ). The difference in distribution of finite and non-finite clauses across all genres is that non-finite clauses occur more frequently in English than in Dutch (34 (50.0%) vs. 5 (15.6%)).

The non-finite clauses can be of three different types, similar to the non-finite clauses in the CD subpattern above. The frequencies of each of these types are presented in Table 39.

**Table 39**      **Frequencies of four types of non-finite clauses in the CDE subpattern**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Non-finite –present participle	5	4	13	2	24 (70.6%)
Non-finite –past participle	3	2	2	2	9 (26.5%)
Non-finite –infinitive	0	0	0	0	0 (0.0%)
Non-finite –verbless	0	0	1	0	1 (2.9%)
<b>Total</b>	<b>8</b>	<b>6</b>	<b>16</b>	<b>4</b>	<b>34 (100%)</b>
<b>Dutch</b>					
Non-finite –present participle	0	0	1	0	1 (20%)
Non-finite –past participle	1	0	2	1	4 (80%)
Non-finite –infinitive	0	0	0	0	0 (0.0%)
Non-finite –verbless	0	0	0	0	0 (0.0%)
<b>Total</b>	<b>1</b>	<b>0</b>	<b>3</b>	<b>1</b>	<b>5 (100%)</b>

A further categorisation of the various types of non-finite clauses and verbless clauses shows that the non-finite clauses with a present participle form by far the largest group in English across all genres (24 (70.5%)). The second largest group is formed by non-finite clauses with a past participle (9 (26.4%)). In Dutch the few instances of non-finite clauses contain a past participle (4 sentences) and 1 sentence contains a present participle. Sentences (40) and (41) below provide examples of a non-finite clause with a present participle and one with a past participle respectively.

(40) <C><indepcl>There is no increase in time spent alone over the three later age groups<indepcl><C>, <D><subcl\_advcl\_nonfin>suggesting that by age 7-8 some children are already solitary in the playground and that this may not increase in the primary school years<subcl\_advcl\_nonfin><D>, <E><subcl\_advcl\_conces>although other aspects of exclusion may do so<subcl\_advcl\_conces><E>. <s5303, academic prose>

(41) <C><fragment\_NP>Twee jaar van zijn leven<fragment\_NP><C>, <D><subcl\_advcl\_nonfin>gefossiliseerd tot papier<subcl\_advcl\_nonfin><D>, <E><subcl\_advcl\_nonfin>ergens verblekend achter het glas van een archiefkast<subcl\_advcl\_nonfin><E>. <s16320, short stories>

(<C><fragment\_NP>Two years of his life<fragment\_NP><C>, <D><subcl\_advcl\_nonfin>fossilised into paper<subcl\_advcl\_nonfin><D>, <E><subcl\_advcl\_nonfin>fading behind the glass of a filing cabinet somewhere<subcl\_advcl\_nonfin><E>.)

The finite adverbial clauses can also be further subcategorised into seven semantic classes, the frequencies of which are presented in Table 40.

With such a refined subcategorisation of a small number of sentences, it is difficult to find a trend in what type of adverbial clause occurs particularly frequently. What can be seen, however, is that the adverbial clause of concession is the most frequently occurring adverbial clause in English (13 (38.2%)), followed by the adverbial clauses of time, the non-specified ones and condition. In Dutch this pattern is different, as the adverbial clauses of time occur most frequently (9 (33.3%)), followed by adverbial clauses of reason/purpose and condition.

**Table 40**      **Frequencies of semantic roles of adverbial clauses – D in the CDE subpattern**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Adverbial cl_time/place	2	1	1	1	5 (14.7%)
Adverbial cl_condition	0	1	1	2	4 (11.8%)
Adverbial cl_concession	7	3	2	1	13 (38.2%)
Adverbial cl_comparison	1	0	0	0	1 (2.9%)
Adverbial cl_reason/purpose	1	0	3	0	4 (11.8%)
Adverbial cl_result	0	0	2	0	2 (5.9%)
Adverbial cl_not specified	1	0	3	1	5 (14.7%)
<b>Total</b>	<b>12</b>	<b>5</b>	<b>12</b>	<b>5</b>	<b>34 (100%)</b>
<b>Dutch</b>					
Adverbial cl_time/place	1	1	7	0	9 (33.3%)
Adverbial cl_condition	0	1	3	1	5 (18.5%)
Adverbial cl_concession	0	0	2	0	2 (7.4%)
Adverbial cl_comparison	0	0	2	0	2 (7.4%)
Adverbial cl_reason/purpose	2	0	5	0	7 (26.0%)
Adverbial cl_result	0	0	0	1	1 (3.7%)
Adverbial cl_not specified	1	0	0	0	1 (3.7%)
<b>Total</b>	<b>4</b>	<b>2</b>	<b>19</b>	<b>2</b>	<b>27 (100%)</b>

**Grammatical realisation of coordinated E in the CDE subpattern**

Only a small number of Es in the CDE subpattern are coordinated (see Table 32 above), 7.5% (18) of the English sentences with this pattern and 8.3% (16) of the Dutch sentences. In English, 17 of these 18 sentences are realised grammatically as a list of coordinated phrases or subordinate clauses and mainly occur in the leaflets genre. In Dutch some of these coordinated units are realised as independent clauses, particularly in the short stories genre, but the majority are realised as coordinated phrases and subordinate clauses (10 out of 16). Sentence (42) presents an example of coordinated Ds in the English leaflets. This sentence structure is typical of the leaflets genre, in which a nuclear unit is followed by a list of coordinated phrases.

- (42) <Ca>People have used sleeping Tablets for many years<Ca>,  
 <Cb><fragment\_comp\_list>but we now know that they<fragment\_comp\_list><Cb>:  
 <Da><coord\_a\_emb\_asyn\_list>don't work for very  
 long<coord\_a\_emb\_asyn\_list><Da>

<Db><coord\_b\_emb\_asyn\_list>Leave you tired and irritable the next  
 day<coord\_b\_emb\_asyn\_list><Db>  
 <Dc><coord\_c\_emb\_asyn\_list>Lose their effect quite  
 quickly<coord\_c\_emb\_asyn\_list><Dc>, <E><subcl\_advcl\_result>so you have to take  
 more and more to get the same effect<subcl\_advcl\_result><E>. <s7384, leaflets>

### Grammatical realisation of E in the CDE subpattern

The E-satellite takes the form of a single, uncoordinated unit in 94.2% of the cases in English and in 93.2% of the cases in Dutch. The coordinated E-satellite occurs mainly in the short stories and leaflets genre in English and in the short stories genre in Dutch. Similar to the D-satellite, the E-satellite can be realised as a clause or a phrase. Table 41 presents the frequencies of E when realised as a phrase or clause.

**Table 41 Grammatical realisation of E (single, uncoordinated unit) in the CDE subpattern**

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
E = phrase	10 (31.3%)	9 (34.6%)	45 (32.6%)	7 (23.3%)	71 (31.4%)
E = clause	22 (68.8%)	17 (65.4%)	93 (67.4%)	23 (76.7%)	155 (68.6%)
Total	32 (100%)	26 (100%)	138 (100%)	30 (100%)	226 (100%)
Dutch					
E = phrase	16 (51.6%)	3 (30.0%)	36 (31.6%)	16 (66.7%)	71 (39.7%)
E = clause	15 (48.4%)	7 (70.0%)	78 (68.4%)	8 (33.3%)	108 (60.3%)
Total	31 (100%)	10 (100%)	114 (100%)	24 (100%)	179 (100%)

The loglinear analysis showed no significant three-way interaction between language, genre and grammatical realisation of the E-satellite in the CDE subpattern at the predetermined alpha level of .01 ( $\chi^2(3) = 10.18, p = .02$ ). Nor did it show a significant two-way interaction between language and realisation of D ( $\chi^2(1) = 2.86, p = .09$ , Cramer's  $V = .08$ ). Table 41 shows that, on the whole, in both the English and Dutch the majority of Es take the form of a clause (English: 155 (68.6%); Dutch: 108 (60.3%)).

Similar to the further specification of D in this subpattern, E can also be further specified into four realisation groups, the frequencies of which are provided in Table 42.

**Table 42**                      **Frequencies of grammatical realisations of E as a phrase in the CDE subpattern in four realisation groups**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
E = adjunct/PP	6	2	17	2	27 (38.0%)
E = apposition	4	7	14	5	30 (42.3%)
E = fragment_phrase	0	0	5	0	5 (7.0%)
E = dis.mark, vocative,tag,yes/no	0	0	9	0	9 (12.7%)
<b>Total</b>	<b>10</b>	<b>9</b>	<b>45</b>	<b>7</b>	<b>71 (100%)</b>
<b>Dutch</b>					
E = adjunct/PP	1	1	8	4	14 (19.7%)
E = apposition	11	2	9	11	33 (46.5%)
E = fragment_phrase	4	0	15	1	20 (28.2%)
E = dis.mark, vocative,tag,yes/no	0	0	4	0	4 (5.6%)
<b>Total</b>	<b>16</b>	<b>3</b>	<b>36</b>	<b>16</b>	<b>71 (100%)</b>

Due to the fact that certain realisation forms are genre specific, it caused the expected frequencies of a number of cells in Table 42 to be too low to be tested statistically. For this reason, differences in realisation patterns will only be described.

In the newspaper articles genre and the leaflets genre in English most phrasal Es take the form of an apposition and this applies to most of the phrases in all four genres in Dutch. In English the second largest group is formed by adjuncts/PPs, whereas in Dutch this is the phrasal fragment. The short stories genre contains a relatively high frequency of discourse markers in both languages.

When realised as a clause, the E-satellite can also be further subcategorised into four realisation groups, the frequencies of which are provided in Table 43.

**Table 43**      **Frequencies of grammatical realisations of E as a clause in the CDE subpattern in four realisation groups**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
E = nonrestrictive rel clause	2	6	7	4	19 (12.3%)
E = adverbial clause	13	9	56	10	88 (56.7%)
E = reporting clause	0	1	18	0	19 (12.3%)
E = independent clause	7	1	12	9	29 (18.7%)
<b>Total</b>	<b>22</b>	<b>17</b>	<b>93</b>	<b>23</b>	<b>155 (100%)</b>
<b>Dutch</b>					
E = nonrestrictive rel clause	4	5	10	3	22 (20.4%)
E = adverbial clause	5	1	23	2	31 (28.7%)
E = reporting clause	0	1	9	0	10 (9.3%)
E = independent clause	6	0	36	3	45 (41.6%)
<b>Total</b>	<b>15</b>	<b>7</b>	<b>78</b>	<b>8</b>	<b>108 (100%)</b>

The detailed subcategorisation of the clausal Es in the CDE subpattern produces too many cells with low expected frequencies, because of which differences in frequencies between the different realisation groups cannot be tested statistically.

In the English academic prose genre, most clausal Es take the form of an adverbial clause, with the second largest group being formed by independent clauses. Dutch follows the reverse pattern, with independent clauses forming the largest group, followed by adverbial clauses.

In the English newspaper genre, the adverbial clauses again form the largest groups, followed by non-restrictive relative clauses. In Dutch this latter group forms the largest group.

In the English short stories genre, the adverbial clauses yet again form the largest group, followed by reporting clauses, rather particular to this genre. In Dutch the largest group is formed by independent clauses, followed by adverbial clauses.

Finally, in the English leaflets genre, the adverbial clauses again form the largest group, closely followed by independent clauses. The Dutch leaflets genre only contains a few instances of clausal Es, which are realised as non-restrictive, independent or adverbial clauses.

Sentence (43) presents an example of a clausal E realised as an independent clause, the most frequently occurring realisation group in Dutch.



- (43) <C><indepcl>De kinderen uit de controlegroep lieten een geheel andere reactie zien<indepcl><C>: <D><indepcl>90% van deze kinderen voelde zich wel geroepen om de proefleider op de hoogte te stellen van de verandering in de situatie die had plaatsgevonden in zijn afwezigheid<indepcl><D>; <E><indepcl>de helft van de kinderen deed dit reeds direct bij zijn terugkeer<indepcl><E>. <s6165, academic prose>

(<C><indepcl>The children of the control group showed a completely different reaction<indepcl><C>: <D><indepcl>90% of these children did feel the urge to inform the experiment leader of the change in the situation that had occurred in his absence<indepcl><D>; <E><indepcl>half of the children did so immediately at his return<indepcl><E>.)

The largest group of clauses in English, adverbial clauses, can be further subcategorised into finite and non-finite clauses. Table 44 provides the frequencies of both groups.

**Table 44**                      **Frequencies of E as a clause in the CD subpattern as finite and non-finite clauses**

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Adverbial cl_finite	7	4	19	3	33 (37.5%)
Adverbial cl_non-finite	6	5	37	7	55 (62.5%)
Total	13	9	56	10	88 (100%)
<b>Dutch</b>					
Adverbial cl_finite	3	1	15	2	21 (67.7%)
Adverbial cl_non-finite	2	0	8	0	10 (32.3%)
Total	5	1	23	2	31 (100%)

The loglinear analysis showed no significant three-way interaction between language, genre and distribution of finite and non-finite clauses ( $\chi^2(3) = 2.46$ ,  $p = .48$ ). The analysis did show a significant two-way interaction between language and distribution of finite and non-finite clauses ( $\chi^2(1) = 8.94$ ,  $p < .01$ , Cramer's  $V = .27$ ). The difference in distribution of finite and non-finite clauses on the whole is that non-finite clauses occur significantly more frequently in English than in Dutch (55 (62.5%) vs. 10 (32.2%)).

The non-finite clauses can be of three different types, similar to the non-finite clauses in the CD subpattern above. The frequencies of each of these types, including the verbless clauses, are presented in Table 45.

**Table 45**            **Frequencies of four types of non-finite clauses in the CDE subpattern**

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Non-finite –present participle	5	4	33	4	46 (83.7%)
Non-finite –past participle	1	1	3	1	6 (10.9%)
Non-finite –infinitive	0	0	0	1	1 (1.8%)
Non-finite –verbless	0	0	1	1	2 (3.6%)
Total	6	5	37	7	55 (100%)
<b>Dutch</b>					
Non-finite –present participle	0	0	5	0	5 (50.0%)
Non-finite –past participle	0	0	2	0	2 (20.0%)
Non-finite –infinitive	2	0	1	0	3 (30.0%)
Non-finite –verbless	0	0	0	0	0 (0.0%)
Total	2	0	8	0	10 (100%)

A further categorisation of the various types of non-finite clauses and verbless clauses shows that non-finite clauses with a present participle form by far the largest group in English across all genres (46 (83.6%)). The second largest group is formed by non-finite clauses with a past participle (6 (10.9%)). In Dutch a few instances of non-finite clauses contain a present participle (5 sentences) and 3 sentences contain an infinitive. Sentence (44) provides an example of an English sentence from the short stories genre that follows the CDE subpattern, where E is realised as a non-finite clause with a present participle.

- (44) <C><reportedcl><indepcl>I'm going<indepcl><reportedcl><C>,  
 <D><reportingcl><fragment>she said suddenly<fragment><reportingcl><D>,  
 <E><subcl\_advcl\_nonfin>striding off the pier on to the heaving  
 esplanade<subcl\_advcl\_nonfin><E>. <s11390, short stories>

### **Grammatical realisation of coordinated E in the CDE subpattern**

Only a very small number of Es in the CDE subpattern are coordinated (see Table 33 above), 5.8% (14) of the English sentences with this pattern and 6.8% (13) of the Dutch sentences. In English, 10 of these are realised grammatically as a list of coordinated phrases or subordinate clauses and mainly occur in the short stories (4) and leaflets genre (5). In Dutch these coordinated units are realised as independent clauses in 12 out of 13 sentences.

### 5.6.3 The CDEF+ subpattern

Table 19 above already showed that the sentence pattern in which the nucleus is followed by three or more satellites is very infrequent in both languages. Of the English sentences that belong to the CX pattern only 3.3% (46) belong to the CDEF+ subpattern, while this applies to 4.3% (61) of the Dutch sentences.

For English, this pattern mainly occurs in the academic prose genre, in which 4.8% of all the sentences that belong to the CX main pattern take the form of the CDEF+ subpattern. For Dutch, this pattern mainly occurs in the short stories genre, in which 6.7% of the sentences belong to this pattern. Sentence (45) presents an example of an English sentence taken from the academic prose genre and sentence (46) is taken from the Dutch short stories genre.

(45) <C><indepcl>There was a strong contrast with Venice and Florence<indepcl><C>, <D><nonrestr\_relcl>where the councils were constituted and conducted according to clearly defined regulations<nonrestr\_relcl><D>, <E><pp>in an ethos in which political faction was officially condemned<pp><E>, <F><subcl\_advcl\_cond\_verbless>if often present<subcl\_advcl\_cond\_verbless><F>. <s4145, academic prose>

(46) <C><indepcl>Ik probeer steeds al op zijn naam te komen<indepcl><C> ... <Da><coord\_a\_subcl>die ligt, <1><subcl\_advcl>zoals dat zo mooi heet<subcl\_advcl><1>, op het puntje van mijn tong<coord\_a\_subcl><Da> ... <Db><coord\_b\_subcl>maar misschien is het daarmee als met het centrum van het oog<coord\_b\_subcl><Db> ... <E><indepcl>daarin lopen de zenuwen dood<indepcl><E> ... <F><indepcl>er valt niets meer te zien<indepcl><F> ...<s13621, short stories>

<C><indepcl>I keep trying to think of his name<indepcl><C> ... <Da><coord\_a\_subcl>which is, <1><subcl\_advcl>as the saying goes<subcl\_advcl><1>, on the tip of my tongue<coord\_a\_subcl><Da> ... <Db><coord\_b\_subcl>but perhaps it is the same thing as with the centre of the eye<coord\_b\_subcl><Db> ... <E><indepcl>that's where nerves reach a dead end<indepcl><E> ... <F><indepcl>there's nothing left to see<indepcl><F> ...)

### 5.6.4 Summary

The X in the CX pattern can consist of one, two or three+ elements. The subpattern in which the nucleus is followed by one satellite is by far the most frequent in both languages. The analysis in this chapter has been restricted to the subpatterns in which the nucleus is followed by one or two satellites, as the frequencies for sentences in which it is followed by three or more satellites are very low.

The nucleus in both the CD and CDE subpattern is usually uncoordinated in both languages, with English showing a higher frequency of coordinated nuclei than Dutch in the CD subpattern. In both languages these coordinated nuclei typically take the form of independent clauses in the CD subpattern. The same applies to the CDE subpattern, although the Cs in the short stories and leaflets genres in both languages are realised as non-independent clauses in over half of the cases. When uncoordinated, although the nucleus is typically realised as an independent clause in both languages, the frequencies for cases in which the nucleus is realised as a non-independent clause are higher for Dutch than for English.

The D-satellite in the CD and CDE subpattern can either consist of uncoordinated or coordinated satellites. In both languages it predominantly consists of uncoordinated satellites, although for the CD subpattern the English short stories genre shows a slightly higher frequency of coordinated Ds and for the CDE subpattern this applies to the English leaflets genre. When coordinated, in English the coordinated elements are predominantly realised as phrases or subordinate clauses in both the CD and CDE subpatterns, whereas in Dutch they are predominantly realised as independent clauses in the CD subpattern, but also as phrases in the CDE subpattern. When uncoordinated, in both languages the D more often takes the form of a clause than a phrase, with the exception of the Dutch leaflets genre in the CD subpattern, in which this pattern is reversed.

When the D-satellite is realised as a clause, the different genres in English and Dutch show variation in its grammatical realisation. In English the largest group is formed by adverbial clauses, with the short stories genre also containing a high instance of reporting clauses and, in the CD subpattern, the leaflets genre showing a high instance of independent clauses. In Dutch adverbial clauses also form a large group, but in the CD subpattern non-restrictive clauses also occur frequently, especially in the academic prose genre, the newspaper genre and leaflets genre. In both the CD and CDE subpatterns independent clauses also form a large group, especially in the academic prose genre, short stories genre and leaflets genre. When realised as an adverbial clause, this more often takes the form of a non-finite clause in English than in Dutch, and then mainly of non-finite clauses with a present

participle. When finite, in English the adverbial clauses of concession form the largest group, especially in the academic prose genre, the newspaper genre, and the leaflets genre in the CD subpattern. In Dutch, on the other hand, the largest groups are formed by adverbial clauses of time and reason/purpose.

When realised as a phrase, in both languages these typically take the form of either an apposition or an adjunct/PP, although there is some variation between the different genres. In the academic prose genre in both languages appositions form the largest group in the CD subpattern, which also applies to the Dutch CDE subpattern. In the CD subpattern appositions also form the largest group in the English newspaper genre, but in Dutch the group of adjuncts/PPs is slightly larger. In the CDE subpattern, phrases are very infrequent in the newspaper genre. In English short stories genre the discourse markers form the largest group, whereas in Dutch a large group is also formed by adjuncts/PPs. In the leaflets genre in both languages the group of appositions again form the largest group in the CD subpattern, which also applies to Dutch in the CDE subpattern.

The E-satellite of course only occurs in the CDE subpattern. In both languages this mainly consists of uncoordinated satellites. The few coordinated satellites mainly occur in the short stories and leaflets genres, and are in English mainly realised as coordinated phrases or subordinate clauses and in Dutch mainly as coordinated independent clauses. When uncoordinated, it can either take the form of a phrase or a clause, with the latter being more frequent in both languages across the different genres.

When realised as a clause, the largest group is formed by adverbial clauses in English and by independent clauses in Dutch, with the exception of the Dutch newspaper genre. The group of independent clauses forms the second largest group in English, especially in the academic prose genre and leaflets genre. When realised as an adverbial clause, this more often takes the form of a non-finite clause in English than Dutch, most of which contain a present participle.

When realised as a phrase, this often takes the form of an apposition across all genres in Dutch and in the English this applies particularly to the newspaper genre and leaflets genre. In English, the second largest group is formed by adjuncts/PPs, whereas in Dutch this is formed by the phrasal fragment.

## 5.7 The XCX pattern

Table 3 at the start of this chapter showed that the XCX pattern is the least frequent sentence pattern. Only 4.6% of all English sentences belong to this pattern, compared to 5.2% of all Dutch sentences. In both languages, it occurs predominantly in the academic prose genre (English: 116 (8.1%); Dutch: 126 (7.2%)), followed by the short stories genre (English: 154 (4.9%); Dutch: 181 (5.7%)).

As both the overviews of the subpatterns of the XC pattern and the CX pattern already showed, the X elements in these patterns can consist of one to three satellites that precede the nucleus and one to three, and in exceptional cases four or five, elements that follow the nucleus. These various possibilities create nine potential subpatterns of the XCX main pattern. Table 46 presents an example of each of these subpatterns and Table 47 presents the frequencies of the patterns.

**Table 46** Potential subpatterns of the XCX pattern

Subpattern of XCX	Example
X (1 el) – C – X (1 el)	<C><zz>Instead<zz> they ordered a second referendum a year later<C> - <D>when voters eventually backed it<D>. <s900, newspaper articles>
X (1 el) – C – X (2 el)	<A>If you ask me<A> <C>they're scared of talking bollocks<C>, <D>which pisses me off<D>, <E>because once I was sat here for a while I wanted the bollocks<E>. <s12115, short stories>
X (1 el) – C – X (3 el)	<A>In nineteenth-century New Zealand<A>, <C>Otago petitioned several times for full separation<C>, <D>sometimes trying to take the rest of South Island with it<D>, <E>while Auckland also attempted full separation<E>, <F>no fewer than five times<F>. <s4384, academic prose>
X (2 el) – C – X (1 el)	<C><zz>On this form of the gender task, <1>which more closely reflects the requirements of real-world gender processing<1><zz>, the majority of children are successful somewhat later<C>, <D>at 36 months<D>. <s5785, academic prose>
X (2 el) – C – X (2 el)	<A>Unfortunately<A>, <B>however<B>, <C>this is far from straightforward<C>, <D>because everyday inferences are global<D>: <Ea>whether a conclusion follows typically depends not just on a few circumscribed 'premises'<Ea>, <Eb>but on arbitrarily large amounts of general world knowledge<Eb>. <s5714, academic prose>

- X (2 el) – C – X (3 el) <A>When they had walked into the bungalow<A>, <B>at five o'clock in the afternoon<B>, <Ca>it had smelled the same<Ca>, <1>the smell she dreamed of sometimes when she was at home<1>, <2>an intense, many-layered smell<2>, <Cb>and each room had a slight variation<Cb> - <D>in the bathroom<D>, <E>traces of antiseptic from the solution in which Elsa's syringes were steeped<E>, <F>in the front room<F>, <G>a faint staleness, and the clothiness of the moquette<G>. <s12021, short stories><sup>33</sup>
- X (3 el) – C – X (1 el) <A>If complete abolition is deemed impracticable in the short term<A>, <C><zz>then<zz> <zz2><sup>34</sup>at the very least<zz2> Europe should commit itself at once to the complete abolition of all export subsidies<C>, <D>direct and indirect<D>. <s1268, newspaper articles>
- X (3 el) – C – X (2 el) <A>All the same<A>, <B>if I indulged the fancy of the child I'd seen in the mirror<B>, <zz>perhaps<zz> <C>I might engage the fancy of her mother<C>, <Da>who must be still young enough to enjoy the caress of a bearskin bedstead<Da>; <Db>and not, <1>I'd be bound<1>, inimical to poetry<Db>, <E>either<E>. <s11440, short stories>
- X (3 el) – C – X (3 el) <A>Klein<A>, <B>blond<B>, <Z>met wat sjokkende tred<Z>, <C>kwam ze mijn richting uit<C>, <D>meegevoerd in de stoet die langs de in massa's uitgestalde vis, kreeft en mosselen trok<D>, <E>de zon in haar gezicht<E>, <F>haar dochtertje, <1>onvermijdelijk<1>, aan haar hand<F>. <s14217, short stories><sup>35</sup>  
(<A>Little<A>, <B>blond<B>, <Z>at a leisurely pace<Z>, <C>she came towards me<C>, <D>carried along by the crowd that passed the heaps of displayed fish, lobster and mussels<D>, <E>the sun in her face<E>, <F>her daughter, <1>inevitably<1>, by the hand<F>.)

<sup>33</sup> This sentence is the only sentence in the English corpus that follows this pattern.

<sup>34</sup> Note that this sentence can also be classified as following an ABZCD pattern. See Chapter 2.4.3 on the <zz> label, which mainly reflects a difference in punctuation practice between English and Dutch.

<sup>35</sup> This sentence is the only sentence in the corpus that follows this pattern. Note also that the <A>, <B> and <Z> satellites could also be analysed as constituting an instance of asyndetic coordination of phrases.

**Table 47**      **Frequencies of subpatterns of the XCX pattern**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
X (1 el) – C – X (1 el)	83 (71.6%)	35 (74.5%)	111 (72.1%)	41 (83.7%)	270 (73.8%)
X (1 el) – C – X (2 el)	16 (13.8%)	6 (12.8%)	26 (16.9%)	2 (4.1%)	50 (13.7%)
X (1 el) – C – X (3 el)	3 (2.6%)	2 (4.3%)	5 (3.2%)	4 (8.2%)	14 (3.8%)
X (2 el) – C – X (1 el)	10 (8.6%)	2 (4.3%)	7 (4.5%)	2 (4.1%)	21 (5.7%)
X (2 el) – C – X (2 el)	3 (2.6%)	0 (0.0%)	3 (1.9%)	0 (0.0%)	6 (1.6%)
X (2 el) – C – X (3 el)	0 (0.0%)	0 (0.0%)	1 (0.6%)	0 (0.0%)	1 (0.3%)
X (3 el) – C – X (1 el)	1 (0.9%)	2 (4.3%)	0 (0.0%)	0 (0.0%)	3 (0.8%)
X (3 el) – C – X (2 el)	0 (0.0%)	0 (0.0%)	1 (0.6%)	0 (0.0%)	1 (0.3%)
X (3 el) – C – X (3 el)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
<b>Total</b>	<b>116 (100%)</b>	<b>47 (100%)</b>	<b>154 (100%)</b>	<b>49 (100%)</b>	<b>366 (100%)</b>
<b>Dutch</b>					
X (1 el) – C – X (1 el)	98 (77.8%)	64 (88.9%)	115 (63.5%)	67 (90.5%)	344 (75.9%)
X (1 el) – C – X (2 el)	21 (16.7%)	3 (4.2%)	29 (16.0%)	6 (8.1%)	59 (13.0%)
X (1 el) – C – X (3 el)	4 (3.2%)	0 (0.0%)	19 (10.5%)	0 (0.0%)	23 (5.1%)
X (2 el) – C – X (1 el)	3 (2.4%)	4 (5.6%)	10 (5.5%)	1 (1.4%)	18 (4.0%)
X (2 el) – C – X (2 el)	0 (0.0%)	1 (1.4%)	4 (2.2%)	0 (0.0%)	5 (1.1%)
X (2 el) – C – X (3 el)	0 (0.0%)	0 (0.0%)	1 (0.6%)	0 (0.0%)	1 (0.2%)
X (3 el) – C – X (1 el)	0 (0.0%)	0 (0.0%)	1 (0.6%)	0 (0.0%)	1 (0.2%)
X (3 el) – C – X (2 el)	0 (0.0%)	0 (0.0%)	1 (0.6%)	0 (0.0%)	1 (0.2%)
X (3 el) – C – X (3 el)	0 (0.0%)	0 (0.0%)	1 (0.6%)	0 (0.0%)	1 (0.2%)
<b>Total</b>	<b>126 (100%)</b>	<b>72 (100%)</b>	<b>181 (100%)</b>	<b>74 (100%)</b>	<b>453 (100%)</b>

Due to the rare occurrence of several subpatterns of the XCX main pattern, the expected values of certain cells are too low to test differences in frequencies between the languages statistically. What becomes clear by looking at Table 47 is that the pattern in which the nucleus is both preceded and followed by one satellite (X(1 el) – C – X(1el)) is by far the most frequent subpattern in both languages across all genres (English: 270 (73.8%); Dutch: 344 (75.9%)). Although far less frequent, this is followed by the subpattern in which the nucleus is preceded by one and followed by two satellites (X (1 el) – C – X (2 el)), which have similar frequencies in both languages in the academic prose genre (English: 16 (13.8%); Dutch: 21 (16.7%)) and the short stories genre (English: 26 (16.9%); Dutch: 29 (16.0%)), but has higher frequencies in the English newspaper genre (6 (12.8%)) when compared to Dutch (3 (4.2%)).

Because of the low frequencies of all but the X(1 el) – C – X(1el) subpattern, this section will be restricted to a detailed analysis of this subpattern. The sentence patterns in which the nucleus is preceded by two or three elements will be described separately in Chapter 6.



### 5.7.1 The ACD subpattern

Table 47 above showed that the ACD (X(1 el) – C - X(1 el)) subpattern is the most frequent one of the XCX subpatterns. This section will describe the make-up of the nucleus and appended satellite and will provide insight into the grammatical realisation of all three elements.

#### A in the ACD subpattern

Similar to the A in the other subpatterns, the A-element in the ACD subpattern can take the form of a phrase or a clause.<sup>36</sup> The frequencies of both the phrases and clauses are presented in Table 48.

**Table 48 Grammatical realisation of A in the ACD subpattern**

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
A = phrase	74 (89.2%)	25 (71.4%)	80 (72.1%)	28 (68.3%)	207 (76.7%)
A = clause	9 (10.8%)	10 (28.6%)	31 (27.9%)	13 (31.7%)	63 (23.3%)
Total	83 (100%)	35 (100%)	111 (100%)	41 (100%)	270 (100%)
Dutch					
A = phrase	92 (93.9%)	55 (85.9%)	94 (81.7%)	59 (88.1%)	300 (87.2%)
A = clause	6 (6.1%)	9 (14.1%)	21 (18.3%)	8 (11.9%)	44 (12.8%)
Total	98 (100%)	64 (100%)	115 (100%)	67 (100%)	344 (100%)

The loglinear analysis showed no significant three-way interaction between language, genre and grammatical realisation of A in the ACD subpattern ( $\chi^2(3)=1.438$ ,  $p=.69$ ). The analysis did show a significant two-way interaction between language and grammatical realisation of the A-element ( $\chi^2(1)=12.00$ ,  $p<.001$ , Cramer's  $V=.13$ ). The difference between English and Dutch is that in English the A-element more often takes the form of a clause when compared to Dutch (English: 63 (23.3%); Dutch: 44 (12.8%)).

The phrasal element can be further subcategorised and specified into four main realisation groups. They can take the form of (1) an adjunct or prepositional phrase, (2) a conjunct, (3) a disjunct or subjunct, and (4) a discourse marker, questions

<sup>36</sup> Note that, similar to the remark in footnote 31 above, the A-element can take the form of either an uncoordinated or a coordinated satellite. As the latter applies to only one sentence, this has not been taken into account in the present analysis.

word or vocative. To this last realisation group a very small number of fronted adjective phrases and fragments have been added. The frequencies of these four main realisation groups are provided in Table 49.

**Table 49**                    **Frequencies of grammatical realisations of A as a phrase in the ACD subpattern in four realisation groups**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
A/zz = adjunct/PP	30 (40.5%)	17 (68.0%)	37 (46.2%)	19 (67.9%)	103 (49.8%)
A/zz = conjunct	38 (51.4%)	7 (28.0%)	16 (20.0%)	8 (28.6%)	69 (33.3%)
A/zz = disjunct/subjunct	6 (8.1%)	1 (4.0%)	8 (10.0%)	1 (3.6%)	16 (7.7%)
A/zz = discourse marker, question word, vocative, yes/no	0 (0.0%)	0 (0.0%)	19 (23.8%)	0 (0.0%)	19 (9.2%)
<b>Total</b>	<b>74 (100%)</b>	<b>25 (100%)</b>	<b>80 (100%)</b>	<b>28 (100%)</b>	<b>207 (100%)</b>
<b>Dutch</b>					
A/zz = adjunct/PP	56 (60.9%)	33 (60.0%)	68 (72.3%)	38 (64.4%)	195 (65.0%)
A/zz = conjunct	30 (32.6%)	18 (32.7%)	5 (5.3%)	16 (27.1%)	69 (23.0%)
A/zz = disjunct/subjunct	4 (4.3%)	4 (7.3%)	5 (5.3%)	2 (3.4%)	15 (5.0%)
A/zz = discourse marker, question word, vocative, yes/no	2 (2.2%)	0 (0.0%)	16 (17.0%)	3 (5.1%)	21 (7.0%)
<b>Total</b>	<b>92 (100%)</b>	<b>55 (100%)</b>	<b>94 (100%)</b>	<b>59 (100%)</b>	<b>300 (100%)</b>

Due to the rare occurrence of several grammatical realisations in some of the genres, the expected values of certain cells are too low to test differences in frequencies between the languages statistically.

In the English academic prose genre conjuncts form the largest realisation group (38 (51.4%)), followed by the group of adjuncts/PPs (30 (40.5%)). Dutch follows the reverse pattern, with the group of adjuncts/PPs forming the largest realisation group (56 (60.9%)), followed by the group of conjuncts (30 (32.6%)).

In the newspaper genre the group of adjuncts/PPs forms the largest realisation group in both languages (English: 17 (68.0%); Dutch: 33 (60.0%)), again followed by conjuncts (English: 7 (28.0%); Dutch: 18 (32.7%)).

In the English short stories genre, adjuncts/PPs form the largest realisation group (37 (46.2%)), followed by discourse markers (19 (23.8%)) and conjuncts (16 (20.0%)). In Dutch adjuncts/PPs also form by far the largest group (68 (72.3%)), followed by discourse markers (16 (17.0%)).

Finally, the leaflets genre in both languages shows a similar distribution of the realisation groups, with the adjuncts/PPs forming the largest group in both

languages (English: 19 (67.9%); Dutch: 38 (64.4%)), followed by conjuncts (English: 8 (28.6%); Dutch: 16 (27.1%)).

The clausal element can also be further specified and subcategorised into six realisation groups: (1) adverbial clauses of time, (2) adverbial clauses of condition, (3) adverbial clauses of concession, (4) adverbial clauses of purpose or reason, (5) adverbial clauses that have not been further specified into a semantic class (i.e. mainly non-finite clauses), and (6) other types of clauses, such as comment clauses. The frequencies of the various clause types can be found in Table 50.

**Table 50**                      **Frequencies of grammatical realisations of A as a clause in the ACD subpattern in six realisation groups**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Advcl_time	2	3	20	2	27 (42.9%)
Advcl_condition	1	2	1	7	11 (17.5%)
Advcl_concession	1	2	0	2	5 (7.9%)
Advcl_purpose/reason	0	0	0	1	1 (1.6%)
Advcl_not specified	5	3	8	1	17 (27.0%)
Other clauses	0	0	2	0	2 (3.1%)
<b>Total</b>	<b>9</b>	<b>10</b>	<b>31</b>	<b>13</b>	<b>63 (100%)</b>
<b>Dutch</b>					
Advcl_time	3	2	8	1	14 (31.8%)
Advcl_condition	1	5	8	4	18 (40.9%)
Advcl_concession	0	0	0	1	1 (2.3%)
Advcl_purpose/reason	0	2	1	2	5 (11.4%)
Advcl_not specified	2	0	4	0	6 (13.6%)
Other clauses	0	0	0	0	0 (0.0%)
<b>Total</b>	<b>6</b>	<b>9</b>	<b>21</b>	<b>8</b>	<b>44 (100%)</b>

Due to the refined subcategorisation of clausal A-elements, the expected frequencies of a large number of cells is too low to test differences in frequencies between the various grammatical realisations statistically. In English the adverbial clauses of time form the largest group (27 (42.9%)), particularly in the short stories genre (20 (64.5%)). The second largest group in English is formed by the non-specified (non-finite) adverbial clause (17 (27.0%)) and the third is formed by adverbial clauses of condition (11 (17.5%)), particularly in the leaflets genre (7 (53.8%)).

In Dutch the adverbial clauses of condition form the largest group (18 (40.9%)), especially in the newspaper (5) and leaflets genre (4). This is followed by adverbial clauses of time (14 (31.8%)).

The largest group of clauses, the adverbial clauses, can be further subcategorised into finite and non-finite clauses. Table 51 provides the frequencies of both groups.

**Table 51**                      **Frequencies of A as a clause in the ACD subpattern as finite and non-finite clauses**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Adverbial cl_finite	3	7	22	11	43 (70.5%)
Adverbial cl_non-finite	6	3	7	2	18 (29.5%)
<b>Total</b>	<b>9</b>	<b>19</b>	<b>29</b>	<b>13</b>	<b>61 (100%)</b>
<b>Dutch</b>					
Adverbial cl_finite	4	9	18	6	37 (84.1%)
Adverbial cl_non-finite	2	0	3	2	7 (15.9%)
<b>Total</b>	<b>6</b>	<b>9</b>	<b>21</b>	<b>8</b>	<b>44 (100%)</b>

The loglinear analysis showed no significant three-way interaction between language, genre and distribution of finite and non-finite clauses ( $\chi^2(3) = 3.67$ ,  $p = .29$ ). Nor did it show a significant two-way interaction between language and distribution of finite and non-finite clauses ( $\chi^2(1) = 2.72$ ,  $p = .09$ , Cramer's  $V = .15$ ).

The non-finite clauses can be of four different types, similar to the non-finite clauses in the CDE subpattern above. In English, of the small number of cases, the majority have a present participle (10 sentences), which occur mainly in the academic prose genre (5) and the short stories genre (5). A few sentences take an infinitive verb (4) or are verbless (3). In Dutch, 5 of the 7 non-finite clauses take an infinitive verb.

### **C in the ACD subpattern**

The C in the ACD subpattern can take the form of a single, uncoordinated unit or it can be coordinated with other Cs. Table 52 gives the frequencies of the make-up of the C in the ACD subpattern.

**Table 52**      **Make-up of C in the ACD subpattern**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
C (single unit)	70 (84.3%)	34 (97.1%)	94 (84.7%)	35 (85.4%)	233 (86.3%)
Ca/b (coordinated nuclei)	13 (15.7%)	1 (2.9%)	17 (15.3%)	6 (14.6%)	37 (13.7%)
<b>Total</b>	<b>83 (100%)</b>	<b>35 (100%)</b>	<b>111 (100%)</b>	<b>41 (100%)</b>	<b>271 (100%)</b>
<b>Dutch</b>					
C (single unit)	90 (91.8%)	59 (92.2%)	96 (83.5%)	66 (98.5%)	311 (90.4%)
Ca/b (coordinated nuclei)	8 (8.2%)	5 (7.8%)	19 (16.5%)	1 (1.5%)	33 (9.6%)
<b>Total</b>	<b>98 (100%)</b>	<b>64 (100%)</b>	<b>115 (100%)</b>	<b>67 (100%)</b>	<b>344 (100%)</b>

A loglinear analysis showed that there is no significant three-way interaction between language, genre and distribution of coordinated and uncoordinated Cs ( $\chi^2(3) = 8.77$ ,  $p = .03$ ). There was also no significant two-way interaction between language and coordinated/uncoordinated Cs ( $\chi^2(1) = 1.60$ ,  $p = .20$ , Cramer's  $V = .06$ ).

The uncoordinated C can either be grammatically realised as an independent clause or a non-independent clause. As the instances in which it takes the form of a non-independent clause are very rare, the results will not be presented in table form. In English, 219 of the 233 sentences (94.0%) are realised as independent clauses. The non-independent clauses occur predominantly in the short stories genre (7 (7.4%)) and the leaflets genre (6 (17.1%)). In Dutch, 302 of the 311 sentences are realised as independent clauses (97.1%), with the non-independent clauses occurring predominantly in the short stories genre (7 (7.3%)).

The coordinated Cs can also either be grammatically realised as coordinated independent clauses or coordinated non-independent clauses. Again, most coordinated Cs take the form of independent clauses (English: 29 (80.6%); Dutch: 29 (87.9%)). In both languages, the few non-independent clauses mainly occur in the academic prose genre (English: 4 out of 12 sentences, Dutch 3 out of 8 sentences).

### **D in the ACD subpattern**

Similar to the Cs in CDE, the Ds can also be uncoordinated or coordinated. Because of the very low frequencies of coordinated Ds, the results will not be presented in table form. In English, 256 of the 270 sentences (94.8%) take the form of uncoordinated Ds and in Dutch this applies to 329 of the 344 sentences (95.6%). The various genres in both languages all have a similar distribution, except for the leaflets genre in English, which contains a slightly higher frequency of coordinated

Ds (7 (17.1%)). These coordinated Ds mainly take the form of a list of coordinated subordinate clauses, which is rather typical of this genre (see example 47).

- (47) <zz>In the following pages<zz> <C>we explain<C>:  
 - <Da><coord\_a\_emb\_asyn\_list>what ULDs are<coord\_a\_emb\_asyn\_list><Da>;  
 - <Db><coord\_b\_phr\_asyn\_list>their symptoms<coord\_b\_phr\_asyn\_list><Db>;  
 - <Dc><coord\_c\_emb\_asyn\_list>how you can avoid  
 them<coord\_c\_emb\_asyn\_list><Dc>;  
 - <Dd><coord\_d\_emb\_list>and what you can do to help<coord\_d\_emb\_list><Dd>.  
 <s7551, leaflets>

The uncoordinated Ds can either take the form of a phrase or a clause, the frequencies of which are provided in Table 53.

**Table 53 Grammatical realisation of uncoordinated D in the ACD subpattern as phrase or clause**

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
D = phrase	26 (32.1%)	13 (39.4%)	41 (38.0%)	12 (35.3%)	92 (35.9%)
D = clause	55 (67.9%)	20 (60.6%)	67 (62.0%)	22 (64.7%)	164 (64.1%)
Total	81 (100%)	33 (100%)	108 (100%)	34 (100%)	256 (100%)
<b>Dutch</b>					
D = phrase	38 (40.4%)	11 (17.7%)	43 (39.8%)	34 (52.3%)	126 (38.3%)
D = clause	56 (59.6%)	51 (82.3%)	65 (60.2%)	31 (47.7%)	203 (61.7%)
Total	94 (100%)	62 (100%)	108 (100%)	65 (100%)	329 (100%)

The loglinear analysis showed no significant three-way interaction between language, genre and grammatical realisation of the D-satellite in the ACD subpattern at the predetermined alpha level of .01 ( $\chi^2(3) = 9.01$ ,  $p = .02$ ). Nor did the analysis show a significant two-way interaction between language and realisation of D ( $\chi^2(1) = .43$ ,  $p = .51$ , Cramer's  $V = .02$ ). In both languages the majority of Ds take the form of a clause (English: 164 (64.1%); Dutch: 203 (61.7%)). In English, the distribution is similar across all genres.

The phrasal element can be further subcategorised and specified into four main realisation groups. They can take the form of (1) an adjunct or PP, (2) an apposition, (3) a phrasal fragment, such as an NP, (4) a discourse marker, vocative, tag or the words *yes* or *no*. Note that there was only one instance of a conjunct in the English short stories genre, which has been added to the fragment category. The frequencies of these four main realisation groups are given in Table 54.

**Table 54**                    **Frequencies of grammatical realisations of D as a phrase in the ACD subpattern in four realisation groups**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
D = adjunct/PP	7	4	13	3	27 (29.3%)
D = apposition	18	9	14	8	49 (53.3%)
D = fragment_phrase	1	0	6	1	8 (8.7%)
D = dis.mark, vocative,tag,yes/no	0	0	8	0	8 (8.7%)
<b>Total</b>	<b>26</b>	<b>13</b>	<b>41</b>	<b>12</b>	<b>92 (100%)</b>
<b>Dutch</b>					
D = adjunct/PP	11	6	19	10	46 (36.5%)
D = apposition	24	4	16	20	64 (50.8%)
D = fragment_phrase	3	1	6	4	14 (11.1%)
D = dis.mark, vocative,tag,yes/no	0	0	2	0	2 (1.6%)
<b>Total</b>	<b>38</b>	<b>11</b>	<b>43</b>	<b>34</b>	<b>126 (100%)</b>

Because certain realisation forms are genre-specific, it causes the expected frequencies of some cells in Table 54 to be too low to test differences in frequencies statistically. For this reason, differences in realisation patterns will be only be described.

In English, across all genres, most phrases are realised as an apposition (53.3%). The second largest group is formed by adjuncts/PPs (29.3%), also across all genres. The short stories genre is the only genre that shows some more variation in grammatical realisation of D, containing some instances of phrasal fragments and discourse markers.

In Dutch it is only in the academic prose genre (63.2%) and leaflets genre (58.8%) that the appositions form the largest group and the adjuncts/PPs the second largest group. In the newspaper and short stories genre the situation is reversed, with adjuncts/PPs forming the largest group, followed by appositions.

When realised as a clause, this can take various grammatical forms. Table 55 presents the frequencies of D realised as a clause, categorised into six realisation groups.

**Table 55**                      **Frequencies of grammatical realisations of D as a clause in the ACD subpattern in six realisation groups**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
D = appended clause	1 (1.8%)	1 (5.0%)	0 (0.0%)	0 (0.0%)	2 (1.2%)
D = nonrestrictive rel clause	14 (25.5%)	3 (15.0%)	7 (10.4%)	2 (9.1%)	26 (15.9%)
D = adverbial clause	38 (69.1%)	13 (65.0%)	35 (52.2%)	14 (63.6%)	100 (61.0%)
D = reporting clause	0 (0.0%)	1 (5.0%)	24 (35.8%)	0 (0.0%)	25 (15.2%)
D = independent clause	2 (3.6%)	2 (10.0%)	1 (1.5%)	6 (27.3%)	11 (6.7%)
D = complement clause	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
<b>Total</b>	<b>55 (100%)</b>	<b>20 (100%)</b>	<b>67 (100%)</b>	<b>22 (100%)</b>	<b>164 (100%)</b>
<b>Dutch</b>					
D = appended clause	0 (0.0%)	4 (7.8%)	1 (1.5%)	0 (0.0%)	5 (2.5%)
D = nonrestrictive rel clause	17 (30.4%)	13 (25.5%)	10 (15.4%)	7 (22.6%)	47 (23.2%)
D = adverbial clause	21 (37.5%)	18 (35.3%)	23 (35.4%)	15 (48.4%)	77 (37.9%)
D = reporting clause	1 (1.8%)	10 (19.6%)	9 (13.8%)	0 (0.0%)	20 (9.9%)
D = independent clause	13 (23.2%)	6 (11.8%)	22 (33.8%)	9 (29.0%)	50 (24.6%)
D = complement clause	4 (7.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	4 (2.0%)
<b>Total</b>	<b>56(100%)</b>	<b>51 (100%)</b>	<b>65 (100%)</b>	<b>31 (100%)</b>	<b>203 (100%)</b>

Due to the detailed subcategorisation of the clausal Ds, the expected frequencies of various cells is too low to test differences in realisation pattern statistically. For this reason differences in frequencies will only be described.

In English, the group of adverbial clauses forms the largest realisation group across all genres (100 (61.0%). In the academic prose genre and newspaper genre, this is followed by the group of non-restrictive clauses (14 (25.5%) and 3 (15.0%)). In the short stories genre the reporting clauses form the second largest group (24 (35.8%)) and in the leaflets genre the independent clauses (6 (27.3%)).

In Dutch the group of adverbial clauses also forms the largest group (77 (37.9%)), but are closely followed by independent clauses (50 (24.6%)) and non-restrictive relative clauses (47 (23.2%)). The frequency of independent clauses is roughly similar in the academic prose genre (13 (23.2%)), the short stories genre (22 (33.8%)) and the leaflets genre (9 (29.0%)), and somewhat lower in the newspaper genre (6 (11.8%)). In this latter genre, the group of reporting clauses forms a fairly large group (10 (19.6%)).

The finite adverbial clauses can also be further subcategorised into seven semantic classes, the frequencies of which are presented in Table 56.



**Table 56**                      **Frequencies of semantic roles of adverbial clauses – D in the ACD subpattern**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Adverbial cl_time/place	3	4	1	1	9 (9.0%)
Adverbial cl_condition	2	1	1	1	5 (5.0%)
Adverbial cl_concession	10	2	0	1	13 (13.0%)
Adverbial cl_reason	3	0	1	0	4 (4.0%)
Adverbial cl_not specified	18	6	30	10	64 (64.0%)
Adverbial cl_result/purpose	2	0	2	1	5 (5.0%)
<b>Total</b>	<b>38</b>	<b>13</b>	<b>35</b>	<b>14</b>	<b>100 (100%)</b>
<b>Dutch</b>					
Adverbial cl_time/place	1	4	4	0	9 (11.7%)
Adverbial cl_condition	2	2	2	2	8 (10.4%)
Adverbial cl_concession	9	1	0	2	12 (15.6%)
Adverbial cl_reason	3	5	0	5	13 (16.9%)
Adverbial cl_not specified	3	4	12	5	24 (31.1%)
Adverbial cl_result/purpose	3	2	5	1	11 (14.3%)
<b>Total</b>	<b>21</b>	<b>18</b>	<b>23</b>	<b>15</b>	<b>77 (100%)</b>

Due to the refined subcategorisation of the various types of adverbial clauses, the expected frequencies of too many cells are too low to test differences in frequencies statistically. What Table 56 shows is that in English the group of adverbial clauses that has not been further specified according to semantic role forms the largest group across all genres, but particularly in the short stories genre (30 (85.7%)) and the leaflets genre (10 (71.4%)). This group of clauses mainly consists of non-finite adverbial clauses (see Table 57 below). In the academic prose genre, the second largest group is formed by clauses of concession (10 (26.3%)) and by adverbial clauses of time in the newspaper genre, although it should be noted that this concerns only 4 sentences. Dutch, on the other hand, does not show a real preference for one type of adverbial clause, but shows variation between the different genres. In the academic prose genre, for instance, the adverbial clauses of concession form the largest group with 9 sentences. In the newspaper genre this is the group of adverbial clauses of reason, although this only applies to 5 sentences. In the short stories genre it is the non-specified group with 12 sentences (52.2%) and in the leaflets genre the adverbial clauses of reason and the non-specified ones form the largest groups (5 sentences each).

In addition to categorising them on the basis of their semantic role, the adverbial clauses can also be divided into finite and non-finite clauses. Tables 57 and 58 provide the frequencies of both groups and a further subcategorisation of different types of non-finite clauses respectively.

**Table 57**                    **Frequencies of D as a clause in the ACD subpattern as finite and non-finite clauses**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Adverbial cl_finite	20	7	7	10	44 (44.0%)
Adverbial cl_non-finite	18	6	28	4	56 (56.0%)
<b>Total</b>	<b>38</b>	<b>13</b>	<b>35</b>	<b>14</b>	<b>100 (100%)</b>
<b>Dutch</b>					
Adverbial cl_finite	20	16	16	10	62 (80.5%)
Adverbial cl_non-finite	1	2	7	5	15 (19.5%)
<b>Total</b>	<b>21</b>	<b>18</b>	<b>23</b>	<b>15</b>	<b>77 (100%)</b>

The loglinear analysis showed no significant three-way interaction between language, genre and distribution of finite and non-finite clauses at the predetermined alpha level of .01 ( $\chi^2(3) = 8.62$ ,  $p = .03$ ). The analysis did show a significant two-way interaction between language and distribution of finite and non-finite clauses ( $\chi^2(1) = 25.16$ ,  $p < .001$ , Cramer's  $V = .36$ ). The difference in distribution of finite and non-finite clauses across all genres is that non-finite clauses occur more frequently in English than in Dutch (56 (56.0%) vs. 15 (19.5%)).

**Table 58**                    **Frequencies of four types of non-finite clauses in the ACD subpattern**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Non-finite –present participle	13	4	21	3	41 (73.2%)
Non-finite –past participle	4	2	6	1	13 (23.2%)
Non-finite –infinitive	0	0	1	0	1 (1.8%)
Non-finite –verbless	1	0	0	0	1 (1.8%)
<b>Total</b>	<b>18</b>	<b>6</b>	<b>28</b>	<b>4</b>	<b>56 (100%)</b>
<b>Dutch</b>					
Non-finite –present participle	0	1	4	0	5 (33.3%)
Non-finite –past participle	0	1	1	4	6 (40.0%)
Non-finite –infinitive	1	0	2	0	3 (20.0%)
Non-finite –verbless	0	0	0	1	1 (6.7%)
<b>Total</b>	<b>1</b>	<b>2</b>	<b>7</b>	<b>5</b>	<b>15 (100%)</b>

Due to the detailed subcategorisation of the different types of non-finite adverbial clauses, the expected frequencies of a number of cells are too low to test differences in frequencies statistically. What Table 58 shows is that in English, the non-finite clauses with a present participle are by far the most frequently occurring ones across all genres (41 (73.2%)). For Dutch, Table 57 above already showed that non-finite clauses occur much less frequently when compared to English, the few instances mainly take the form of clauses with a present participle (5 in total), particularly in the short stories genre (4) or clauses with a past participle (6 in total), particularly in the leaflets genre (4).

### 5.7.2 Summary

As the ACD subpattern is the most frequently occurring subpattern of the nine potential subpatterns of the XCX pattern, the analysis here has been restricted to this subpattern.

Similar to the A in the XC pattern, the A-element in the ACD subpattern can take the form of a phrase or a clause, with cases in which it takes the form of a clause being more frequent in English than in Dutch. In both languages, the vast majority of these clauses are realised as adverbial clauses of various types, with adverbial clauses of time followed by non-finites being most frequent in English and adverbial clauses of condition followed by time being most frequent in Dutch. The small number of sentence-initial non-finite clauses are mainly realised as clauses with a present participle in English.

When the A-element takes the form of a phrase, in both English and Dutch the groups of adjuncts/PPs and conjuncts form the largest realisation group, showing some variation in frequencies between the different genres. The group of conjuncts forms the largest group in the English academic prose genre and the adjunct/PPs forms the largest group in all other genres in both languages, including the Dutch academic prose genre.

As for the nucleus in the ACD subpattern, in both languages this is predominantly realised as a single, uncoordinated nucleus. This is realised as an independent clause in the vast majority of cases, with slightly higher frequencies of non-independent clauses in the short stories genre in both languages.

Finally, the D-satellite can also be coordinated or uncoordinated. In both languages it is uncoordinated in the vast majority of cases, with the English leaflets genre showing a slightly higher frequency of coordinated Ds. When uncoordinated, it can either take the form of a phrase or a clause, with the latter being more

frequent in both languages. In English, most of these clauses take the form of adverbial clauses and in Dutch this is also the largest group, but this is closely followed by independent clauses and non-restrictive clauses. When realised as adverbial clauses, most of these take the form of non-finite clauses with a present participle in English and in Dutch the non-specified adverbial clauses forms a large group, closely followed by adverbial clauses of time and reason.

When realised as a phrase, the realisation groups of appositions and adjunct/Ps are the largest ones in both languages. In English, the group of appositions is the largest one across all genres; in Dutch this only applies to the academic prose genre and leaflets genre, with the group of adjunct/Ps being the largest in the short stories and newspaper genre.

## 5.8 Conclusion

The aim of this chapter was to present an analysis of the main sentencing patterns in English and Dutch in four different genres. Similar to the results throughout the chapter, in this conclusion section the order in which the information will be presented is by first discussing all three-way interaction, which will be followed by a second that provides an overview of all significant two-way interactions. The conclusion section will end with an overview of structural similarities between the two languages.

### **Characteristics of the different genres following from three-way interactions and informal comparisons**

The following tables contain all significant three-way interactions (marked with an asterisk) and a few structural differences between the languages at the level of the genres that could not be tested due to low expected frequencies (marked without an asterisk).; Table 60 the main results for the newspaper genre; Table 61 the main results for the short stories genre, and Table 62 the main results for the leaflets genre.

*Academic prose*

Table 59 below presents the main results for the English and Dutch academic prose genre.

**Table 59 Overview of all significant three-way interactions and informal comparisons: academic prose genre**

	English	Dutch
<b>Academic prose</b>		
Sentence length	longer sentences (+ 31 words cat)*	shorter sentences (+ 1-10 & + 11-20 words cat)*
<b>C-pattern</b>		
Ca/b	more coordinated nuclei (26.6%)*	less coordinated nuclei (15.3%)*
<b>XC pattern</b>		
<i>AC pattern</i>		
A: gram. form	more clauses (21.4%)*	fewer clauses (12.4%)*
A – phrase	lower frequency of adjunct/PP (54.8%)* higher frequency of conjuncts (37.9%)*	higher frequency of adjunct/PP (69.9%)* lower frequency of conjuncts (26.6%)*
A- clause	higher frequency of advcl concession (32.9%) lower frequency of advcl condition (10.1%)	lower frequency of advcl concession (11.1%) higher frequency of advcl condition (23.3%)
<b>CX pattern</b>		
<i>CD pattern</i>		
D – clause	most clausal Ds adv clauses (72.2%) adv cl of concession most freq (59.4%) lower freq of indepcl (8.2%) & nonres cl (16.5%)	fewer adv clauses (39.8%) adv cl time (35.7%) & reason (33.3%) most freq higher freq of indep cl (25.4%) & nonres cl (30.5%)
<i>CDE pattern</i>		
D – phrase	few instances, mainly conj/disj	few instances, mainly apposition
D – clause	most clausal Ds adv cl (83.3%), few nonres (4.2%)	fewer adv clauses (41.7%), more nonres (33.3%)

In the academic prose genre, sentences in English are significantly longer than sentences in Dutch, with English containing more sentences in the 31+ words and above category and Dutch containing more sentences in the 1-10 words and 11-20 words categories.

The main difference for the most frequent sentence pattern, the C-pattern, is that English contains more sentences in which the nuclei are coordinated than Dutch, although it should be noted that in both languages the vast majority of sentences contain uncoordinated nuclei.

For the AC subpattern, a difference between English and Dutch is that cases in which the A-satellite takes the form of a clause are more frequent in English than in Dutch, although most A-satellites in both languages take the form of a phrase. When realised as a phrase, in both languages the group of adjunct/PPs forms the largest realisation group, but this occurs significantly more often in Dutch than in English. The second largest group is formed by conjuncts, which occur significantly more often in English than in Dutch. When realised as a clause, English shows a higher frequency of adverbial clauses of concession, whereas Dutch shows a higher frequency of adverbial clauses of condition.

The differences between English and Dutch for the main subpattern of the CX pattern, the CD subpattern, are that in English the D-satellite, when realised as a clause, takes the form of an adverbial clause in the vast majority of cases, whereas in Dutch the frequencies for independent clauses and non-restrictive clauses are much higher when compared to English. The D-satellite in the CDE subpattern shows a similar pattern, with adverbial clauses being the most frequently occurring types of clauses in English, whereas in Dutch the numbers are divided across the categories of adverbial clauses and non-restrictive relative clauses.

*Newspaper articles*

Table 60 below presents the main results for the English and Dutch academic prose genre.

**Table 60 Overview of all significant three-way interactions and informal comparisons: newspaper genre**

	English	Dutch
<b>Newspaper articles</b>		
Sentence length	longer sentences ( + 31 words cat)*	shorter sentences (+ 11-20 words cat)*
<b>C-pattern</b>	more frequent (68.5%)*	less frequent (57.6%)*
<b>XC pattern</b>	less frequent (16.9%)*	more frequent (26.6%)*
<i>AC pattern</i>		
A: gram. form	more clauses (29.9%)*	fewer clauses (14.2%)*
A = phrase	lower frequency of conjuncts (17.3%)*	higher frequency of conjuncts (26.6%)*
A = clause	lower frequency advcl condition (34.5%) higher frequency advcl concession (14.3%)	higher frequency advcl condition (43.5%) lower frequency advcl concession (8.1%)
<b>CX- pattern</b>		
<i>CD pattern</i>		
D – clause	lower freq nonres cl (13.8%) adv cl conces (34.6%) & time (23.0%) most freq	higher freq nonres cl (27.5%) adv cl not spec (30.2%) & reason (25.6%) most freq
<i>CDE pattern</i>		
D – clause	higher freq adv cl (57.9%)	lower freq adv cl (25.0%)

In the newspaper genre, English sentences are again longer than Dutch sentences, with English containing significantly more sentences in the 31+ words category and Dutch containing more sentences in the 11-20 words category.

The main difference between the languages for the main C-pattern is that the frequencies for this pattern are significantly higher for English than for Dutch. In contrast, the frequencies for the XC pattern are significantly higher for Dutch than for English. In the main subpattern of the XC pattern, the AC subpattern, cases in which the A-satellite takes the form of a clause are significantly more frequent for English than Dutch. In both languages, this clause often takes the form of an adverbial clause of condition, but this occurs more frequently in Dutch than in English, whereas clauses of concession occur more frequently in English than in Dutch. Moreover, when realised as a phrase, Dutch shows a significantly higher frequencies of conjuncts than English.

In the CX pattern, the main differences between English and Dutch can be found in the grammatical realisation of the D-satellite. Specifically, in the CD subpattern, adverbial clauses of concession and time are particularly frequent in English, whereas non-specified adverbial clauses (mainly introduced by *zoals*) and adverbial clauses of reason are particularly frequent in Dutch. Moreover, Dutch also shows a higher frequency for non-restrictive relative clauses. In the CDE subpattern, English again shows a higher frequency of clausal Ds taking the form of adverbial clauses than Dutch.

### Short stories

Table 61 below presents the main results for the English and Dutch academic prose genre.

**Table 61** Overview of all significant three-way interactions and informal comparisons: short stories genre

	English	Dutch
<b>Short stories</b>		
<b>C-pattern</b>	more frequent (60.3%)*	less frequent (55.2%)*
C: gram. form	more fragments (=reporting cl) (11.8%)*	fewer fragments (3.9%)*
<b>XC pattern</b>	less frequent (11.1%)*	more frequent (16.0%)*
<i>AC pattern</i>		
A:= phrase	lower frequency of adjunct/PP (58.5%)*	higher frequency of adjunct/PP (80.1%)*
	higher frequency of conjuncts (16.6%)*	lower frequency of conjuncts (6.2%)*
A = clause	lower frequency advcl of condition (13.7%)	higher frequency advcl of condition (24.5%)
<b>CX pattern</b>		
<i>CD pattern</i>		
Da/b	fewer coordinated Ds (2.4%)	more coordinated Ds (7.6%)
D – phrase	disc. mark. (37.7%) most frequent	adj/PPs (40.0%) most frequent
D – clause	higher freq adv cl (45.7%)	lower freq adv cl (21.0%)
	adv cl time (26.7%) & conces (20.0%) most freq	adv cl time (39.2%) & compar (26.8%) most freq
	lower freq indep cl (5.0%)	higher freq indep cl (29.3%)
<i>CDE pattern</i>		
D – clause	reporting cl (45.9%) & adv cl (37.8%) most freq	adv cl (43.1%) & indepcl (33.3%) most freq



The frequencies for the main C-pattern are significantly higher for English than for Dutch, whereas the frequencies for the XC pattern are significantly higher for Dutch than for English. As for the main subpattern of the XC pattern, the AC subpattern, when the A-satellite takes the form of a phrase, the frequencies for adjuncts/PPs are significantly higher for Dutch than for English, whereas English shows a significantly higher frequency of conjuncts than Dutch. When this element takes the form of a clause, Dutch shows a higher frequency of adverbial clauses of condition than English.

For the CX pattern, the main differences between English and Dutch can again be found for the D-satellite. Specifically, in the CD subpattern, although the vast majority of sentences contain uncoordinated Ds, cases in which D is coordinated are significantly more frequent in Dutch than in English. When uncoordinated and when it takes the form of a phrase, the realisation group of discourse markers, vocatives and questions words forms the largest realisation group in English, whereas Dutch shows a higher frequency of adjuncts/PPs. When realised as a clause, English again shows a higher frequency of D-elements that take the form of an adverbial clause than Dutch, with adverbial clauses of time and concession forming particularly large groups in English and adverbial clauses of time and comparison (clauses introduced by *zoals*) forming large groups in Dutch. Furthermore, Dutch shows a much higher frequency of independent clauses than English. In the CDE subpattern, the clausal D-satellite often takes the form of an adverbial clause in both languages, but frequencies for reporting clauses are particularly high in English and independent clauses particularly high in Dutch.

*Leaflets*

Table 62 below presents the main results for the English and Dutch leaflets genre.

**Table 62 Overview of all significant three-way interactions and informal comparisons: leaflets genre**

	English	Dutch
<b>Leaflets</b>		
Sentence length	longer sentences (+21-30 & +31 words cat)*	shorter sentences (+1-10 words cat)*
<b>C-pattern</b>	more frequent (65.7%)*	less frequent (56.9%)*
C: gram. form	fewer fragments (3.5%)*	more fragments (9.7%)*
Ca/b	more coordinated nuclei (19.0%)*	fewer coordinated nuclei (14.2%)*
<b>XC pattern</b>	less frequent (16.7%)*	more frequent (28.3%)*
<i>AC pattern</i>		
A: gram. form	more clauses (39.1%)*	fewer clauses (18.6%)*
A = phrase	lower frequency of adjunct/PP (59.9%)*	higher frequency of adjunct/PP (70.9%)*
	higher frequency of conjuncts (36.6%)*	lower frequency of conjuncts (23.3%)*
A = clause	lower frequency advcl of condition (50.5%)	higher frequency advcl of condition (66.7%)
<b>CX pattern</b>	more frequent (65.7%)*	less frequent (56.9%)*
<i>CD pattern</i>		
D – gram. form	more clauses (59.5%)	more phrases (60.2%)
D – clause	lower freq nonres cl (18.0%) adv cl conces (29.0%) & reas (22.6%) most freq	higher freq nonres cl (29.7%) adv cl reas (41.7%) & cond (29.2%) most freq
<i>CDE pattern</i>		
D – phrase	low frequency apposition (20.0%)	high frequency apposition (60.0%)

Sentences in the leaflets genre are again significantly longer in English than in Dutch, with English containing more sentences in the 21-30 and 31+ words categories and Dutch containing more sentences in the 1-10 words category.

The C-pattern is again significantly more frequent in English than in Dutch. When the nucleus is uncoordinated, cases in which this is grammatically realised as a fragment are significantly more frequent in Dutch than in English. Coordinated nuclei, on the other hand, occur significantly more often in English than in Dutch.

The AC subpattern occurs significantly more often in Dutch than in English. The A-satellite is significantly more often realised as a clause in English when compared to Dutch. In Dutch, this clause is particularly often realised as an adverbial clause of condition. When realised as a phrase, Dutch shows significantly

higher frequencies of adjuncts/PPs than English, whereas English shows significantly higher frequencies of conjuncts than Dutch.

The D-satellite in the main subpattern of the CX pattern, the CD subpattern, is significantly more often realised as a clause in English when compared to Dutch. In English cases in which this clause takes the form of an adverbial clause of concession and time are particularly frequent, whereas in Dutch cases in which this clause takes the form of a non-restrictive relative clause or adverbial clause of reason or condition are particularly frequent. As for the phrasal D in the CDE subpattern, in Dutch this particularly often takes the form of an apposition when compared to English.

### **Characteristics of the languages following from two-way interactions and informal comparisons**

In the AC subpattern, when the A-satellite takes the form of a clause, this clause is significantly more often non-finite in English than in Dutch. Moreover, in English the majority of these non-finite clauses contain a present participle, whereas in Dutch they mainly contain a past participle or an infinitive. As for the nucleus in this subpattern, cases in which this is coordinated are significantly more frequent in English than in Dutch.

In addition to being preceded by one element, the nucleus can also be preceded by two or three elements. Cases in which the nucleus is preceded by two or three elements are significantly more frequent in English than in Dutch.

In the CD subpattern, when the nucleus is uncoordinated, this is significantly more often realised as a non-independent clause in Dutch than in English. English, on the other hand, contains significantly more instances of coordinated nuclei than Dutch. The D-satellite in the CD subpattern is significantly more often realised as a non-finite clause in English than Dutch, and in English these are mainly non-finite clauses with a present participle, whereas in Dutch the few instances mainly contain a past participle. When finite, English shows a high frequency of adverbial clauses of concession, whereas Dutch shows higher frequencies for adverbial clauses of time and reason. When it consists of coordinated Ds, these D satellites are significantly more often realised as coordinated non-independent clauses in English than in Dutch.

In the CDE subpattern, the D-satellite can take the form of a phrase or a clause. Cases in which it takes the form of a clause are significantly more frequent in English than in Dutch. Similar to the D-satellite in the CD subpattern, this clause

is significantly more often non-finite in English than in Dutch, mainly taking a present participle, whereas in Dutch the few instances, again, take a past participle. When finite, on the other hand, English again shows a higher frequency of adverbial clauses of concession, whereas Dutch shows a higher frequency of adverbial clauses of time and reason. The E-satellite in the CDE subpattern can also take the form of a clause or a phrase. When it takes the form of a clause, in English the largest group is formed by adverbial clauses, whereas in Dutch the largest group is formed by independent clauses. When adverbial, English shows a significantly higher frequency of non-finite clauses than Dutch, mainly containing a present participle.

In the ACD subpattern, the A-satellite can take the form of a phrase or a clause. In English, this satellite significantly more often takes the form of a clause than Dutch. In English, this clause is mainly realised as a non-finite clause with a present participle or a finite adverbial clause of time, whereas Dutch shows a particular high frequency for adverbial clauses of condition and time. The clausal D-satellite in the ACD subpattern is significantly more often realised as a non-finite clause, mainly with a present participle, in English than in Dutch. Dutch shows a higher frequency for independent clauses in this position.

**Table 63** Overview of all significant two-way interactions and informal comparisons

	English	Dutch
<b>XC pattern</b>		
<i>AC pattern</i>		
A = clause	advcl non-finite more frequent (22.0%)* nonfin cl: mainly present participle	advcl non-finite less frequent (12.9%)* nonfin cl: mainly past participle & infinitive
C – coordinated nuclei	higher frequency (16.9%)*	lower frequency (10.9%)*
X= 2 elements	more frequent (12.3%)*	less frequent (6.5%)*
X= 3 elements	more frequent (1.7%)*	less frequent (0.3%)*
<b>CX pattern</b>		
<i>CD pattern</i>		
C – gram. form	lower frequency fragments (16.1%)*	higher frequency fragments (21.5%)*
C – coordinated nuclei	higher frequency (13.8%)*	lower frequency (8.7%)*
D – clause fin/non-fin	higher frequency non-fin cl (59.6%)*	lower freq non-fin cl (16.7%)*
D – clause non-fin	mainly present participle (70.2%)	few instances mainly past participle (60.6%)
D – clause adv	relatively high freq conces (34.3%) time (17.9%) & reason (10.4%) smaller grps	lower freq conces (10.3%) time (26.7%) & reason (26.0%) largest grps
Da/b – gram. form	high freq coord of non-indepcl (75.7%)*	high freq of coord indep cl (61.5%)*
<i>CDE pattern</i>		
D – gram. form	higher frequency clauses (58.1%)*	lower frequency clauses (43.2%)*
D – clause fin/non-fin	higher frequency non-fin cl (50.0%)*	lower frequency non-fin cl (15.6%)*
D – clause non-fin	mainly present participle (70.5%)	few instances, mainly past participle
D – clause adv	higher freq conces (38.2%)	higher freq time (33.3%) & reas (25.9%)
E – clause	adverbial cl (56.8%) most freq	indep cl (41.7%) most freq
E – clause fin/non-fin	higher frequency non-fin cl (62.5%)*	lower frequency non-fin cl (32.2%)*
E – clause non-fin	mainly present participle (83.6%)	(only short stories genre) present participle (5x)
<b>XCX pattern</b>		
<i>ACD pattern</i>		
A – gram. form	higher frequency clauses (23.3%)*	lower frequency clauses (12.8%)*
A – clause adv	time (42.9%) & non-fin (27.0%) most freq	condition (40.9%) & time (31.8%) most freq
A – clause non-fin	mainly present participle (55.6%)	few occurrences, mainly infinitive
D – clause	indep cl lower frequency (6.7%)	indep cl higher frequency (24.6%)
D – clause fin/non-fin	higher frequently non-fin cl (56.0%)*	lower frequency non-fin cl (19.5%)*
D – clause non-fin	mainly present participle (73.2%)	few instances, mainly past & present participle

### **Structural similarities between the different genres of English and Dutch**

One of the most notable similarities between English and Dutch, across the different genres, is that in both languages most sentences belong to the C-pattern and that only a small number of sentences belong to the XCX pattern. In fact, in both languages, most sentences belong to a restricted number of subpatterns of the main sentence patterns. Specifically, the largest number of sentences belong to the C-pattern in which the nuclei are uncoordinated. Two other large subpatterns are the AC subpattern and the CD subpattern, the former of which is more frequent in Dutch than in English. Finally, although far less frequent when compared to the other subpatterns, the ACD subpattern forms the most frequent subpattern of the XCX main pattern in both languages.

#### *Academic prose*

With respect to the distribution of sentences across the four main sentence patterns, English and Dutch show no significant differences in this genre. In both languages, the majority of sentences belong to the C-pattern, followed by the XC pattern, then followed by the CX pattern and finally the XCX pattern. In the most frequent pattern, the C-pattern, hardly any sentences in either language are grammatically realised as non-independent clauses, meaning that almost all sentences in this genre are independent clauses. Another similarity between the languages concerns the grammatical realisation of the phrasal D in the CD and ACD subpatterns, which in both languages typically takes the form of an apposition or an adjunct/PP. It should be noted, however, that the D-satellite in the ACD subpattern typically takes the form of a clause in both languages.

#### *Newspaper articles*

With respect to the distribution of the main sentence patterns, English and Dutch show similar frequencies for the CX pattern, including the main subpattern of this sentence pattern, the CD subpattern. Similar to the phrasal D in the CD subpattern in the academic prose genre, in the newspaper genre the phrasal D is also predominantly realised as an apposition or adjunct/PP in both languages. As for the A-satellite in the ACD subpattern, in both languages this typically takes the form of an adjunct/PP, with the second largest group being formed by conjuncts. The D-satellite in this pattern again predominantly takes the form of a clause in both languages, similar to the academic prose genre.

### *Short stories*

The short stories genre is the only one of the four genres in which no differences in sentence length between the two languages were found. In both languages, over 80% of the sentences do not exceed 20 words. This is also the genre in which the CX pattern is relatively more frequent (23.4%) when compared to the other genres (12.9%). Another characteristic of this genre that applies to both languages is that the nucleus in the C-pattern, the CD subpattern and the CDE subpattern is realised as a non-independent clause in a large number of cases, especially when compared to the other genres. Furthermore, in the AC subpattern the A-satellite takes the form of an adverbial clause of time in a large number of cases in both languages. The analyses throughout the chapter have shown that the grammatical realisation group of discourse markers, vocatives, question words and tags is particularly characteristic of the short stories genre in both languages. The frequency of this realisation group is particularly high in the CDE subpattern, where the phrasal D-satellite takes this form. Finally, again similar to the other genres discussed so far, the D-satellite in the ACD subpattern much more often takes the form of a clause than a phrase in both languages.

### *Leaflets*

The main similarities between the leaflets genres in English and Dutch can be found in the CX and XCX patterns. In the CD subpattern, the nucleus takes the form of a non-independent clause in a relatively high number of cases in both languages. The phrasal D-satellite in this subpattern, but also in the ACD subpattern, predominantly takes the form of an apposition in both languages, followed by adjunct/PPs. In the CD subpattern, the clausal D is realised as an independent clause in quite a high number of cases (32.8%) in both languages. Finally, similar to the newspaper genre, the phrasal A-element in the ACD subpattern typically takes the form of an adjunct/PP in both languages, followed by conjuncts.

**Table 64** Overview of structural similarities between English and Dutch genres

English & Dutch	
<b>Academic prose</b>	
main sentence patterns	no differences (C: 47.3%; XC: 30.5%; CX: 14.6%; XCX: 7.6%)
<b>C-pattern</b>	
gram. form C	hardly any fragments (0.7%)
<b>CX pattern</b>	
D – phrase	most phrasal Ds realised as adjuncts/PPs (41.3%) and appositions (56.5%)
<b>XCX pattern</b>	
<i>ACD pattern</i>	
D – gram. form	clause most frequent, phrase less frequent
D – phrase	appositions largest group (65.6%), followed by adj/PPs (28.1%)
<b>Newspaper articles</b>	
<b>C-pattern</b>	
Ca/b	similar frequencies (14.2%)
<b>CX pattern</b>	
D – phrase	similar frequencies (11.9%)
	most phrasal Ds realised as adjuncts/PPs and appositions
<b>XCX pattern</b>	
<i>ACD pattern</i>	
A – phrase	adjunct/PP most frequent (62.5%), conjunct 2nd (31.3%)
D – gram. form	clause more frequent, phrase less frequent
<b>Short stories</b>	
Sentence length	no differences (81.7% of sentences 1-20 words)
<b>C-pattern</b>	
gram. form C	high frequency non-independent clauses (22.7%)
Ca/b	similar frequencies (23.5%)
<b>XC pattern</b>	
<i>AC pattern</i>	
A = clause	advcl time largest group & similar frequencies (55.0%)
<b>CX pattern</b>	
<i>CD pattern</i>	similar frequencies (23.4%)
C – gram. form	high frequency non-independent clauses (Eng: 25.4%; Du: 33.7%)
<i>CDE pattern</i>	
C – gram. form	high frequency non-independent clauses (40.0%)
D – phrase	high frequency of discourse markers (36.5%)
<b>XCX pattern</b>	
<i>ACD pattern</i>	
D – gram. form	high frequency of clauses (61.1%), phrases less frequent
<b>Leaflets</b>	
<b>CX pattern</b>	
<i>CD pattern</i>	
D – phrase	most phrasal Ds realised as appositions
D – clause	relatively high frequency of indep cl (32.8%)
<i>CDE pattern</i>	
C – gram. form	high frequency of non-independent clauses (50.0%)
<b>XCX pattern</b>	
<i>ACD pattern</i>	
A – phrase	adjunct/PP most frequent (65.5%), conjuncts 2nd (27.6%)
D – phrase	apposition largest group (60.9%), followed by adjunct (28.3%)



**Case studies: complex beginnings, interpolated satellites and punctuational devices**

Following this main results chapter, three areas will be further investigated in the case studies that follow this chapter. The first case study will focus on the sentence patterns in which the nucleus is preceded by two or three prepended satellites. The present chapter has shown that the pattern in which the nucleus is preceded by one satellite is more frequent in Dutch and the pattern in which it is preceded by two or more satellites more frequent in English.

The second case study will focus on the sentence patterns formed by interpolated satellites, which can occur in the nucleus and in prepended/appended satellites. It will provide a detailed analysis of these subpatterns, comparing and contrasting their frequencies between the different languages and the four genres.

The third and final case study will look at a number of punctuation marks in more detail, i.e. the colon, the semi-colon and the dash. The chapter will focus on the type of units that these punctuation marks typically link by looking into both the discourse and grammatical status of these units. It will compare and contrast the patterns formed by these punctuation marks in the two languages and four genres.

## 6. Complex beginnings

### 6.1 Introduction

Chapter 5 (5.3.1 & 5.5) already showed that differences between English and Dutch can be found with respect to the start of sentences. Dutch showed significantly more instances of the XC main sentence pattern in three of the four genres, with the academic prose genre being the only exception. It was also shown that sentences can start with one, two or three satellites that precede the nucleus. Despite the fact that the XC pattern as a whole is more frequent in the Dutch newspaper, short stories and leaflets genres than in English, the subpatterns in which the nucleus is preceded by two or by three satellites occur significantly more frequently in English than in Dutch (cf. 5.5, Table 10). This latter result in itself is not surprising, as Dutch is a verb-second language, which means that the finite verb is typically placed in second position and no more than one element can be in sentence-initial position (Haeseryn et al. 1997: 1261, see also Smits 2002: 22). However, despite this characteristic of Dutch, the language does allow for sentences that start with two or three satellites that precede the nucleus, as Section 5.5 showed. This is in line with Smits' finding, who, in her study of complex beginnings in English and the English produced by Dutch learners, found that Dutch allows for the same adverbial clusters in sentence-initial position as English does, albeit with a lower frequency (2002: 168-169).

The present chapter will take a closer look at sentences that start with two or three sentence-initial satellites. Specifically, it will look at the range of subpatterns that the different combinations of sentence-initial satellites form. Similar to Smits' characterisation, these types of beginnings will be referred to as *complex beginnings*, although it should be noted that she restricted this label to the occurrence of multiple adverbials in sentence-initial position (2002: 16), whereas this study uses this label to refer to those cases in which the subject is preceded by two or more elements, irrespective of their grammatical realisation. Furthermore, the chapter will describe how the different subpatterns can be distinguished from each other and what functions and grammatical forms the different sentence-initial satellites have in these patterns. The focus will be on how the two languages differ

from each other with respect to the occurrence of these patterns and on what the role of genre is in this respect.

The chapter will start by providing an overview of the different types of complex beginning<sup>37</sup>. It will then look at each of these different types of complex beginnings in more detail, starting with sentences that start with two satellites that occur before the nucleus and concluding with sentences in which the nucleus is preceded by three satellites.

## 6.2 Overview and exemplification of complex beginnings

A closer analysis of the complex beginnings in this study shows that the elements in sentence-initial position form a number of recurring combinations with each other, creating a range of subpatterns. This section will present an overview of these different subpatterns, exemplifying them and showing how they can be distinguished from each other. In presenting this overview, a distinction will be made between sentences that start with two sentence-initial elements and sentences that start with three sentence-initial elements. As the sentences with two elements in initial position are far greater in number than the sentences with three elements in initial position, these will receive the main focus. The subpatterns formed by the three sentence-initial elements will be further described and exemplified in Section 6.5 below.

The two elements in sentence-initial position can form the following combinations of discourse units: A1AC(X)<sup>38</sup>, A1C(X) and ABC(X). In the A1AC(X) pattern the nucleus is preceded by the satellite A that is interrupted by an interpolated satellite <1>. Sentence (1) presents an example of this particular

---

<sup>37</sup> It should be noted that this chapter provides an overview of *all* complex beginnings in the two languages, meaning that the complex beginnings can occur in both the XC main pattern and the XCX main pattern (see Chapter 5 for these main sentence patterns).

<sup>38</sup> The label <1> indicates that two of the three subpatterns under consideration contain an interpolated satellite. As one of the case studies, presented in Chapter 8, is devoted to the subpatterns formed by interpolated satellites, these two patterns could also have been included in that chapter. However, as the present chapter focuses on sentence beginnings, a choice was made to include them in this chapter.

sentence pattern, which will be further exemplified and described in Section 6.4 below.

- (1) <A>Where, <1>as was frequently the case<1>, a testator died with minority-aged children<A>, <C>a system of support was required that provided for their maintenance and did not disrupt this and future anticipated stages of the life course<C>. <s4497, academic prose>

In the A1C(X) and ABC(X) patterns the nucleus is also preceded by two satellites. To explain the difference between the A1C(X) and ABC(X) patterns, reference will be made to the study on complex beginnings in native and learner English carried out by Smits (2002). In this study, Smits distinguishes between two main types of complex beginnings: sentences starting with two adverbials in sentence-initial position in which the second adverbial does not ground the first adverbial and sentences starting with two adverbials in which the second adverbial does ground the first adverbial (2002: 75ff).<sup>39</sup> In the present study, a similar distinction will be made, where the grounding function of the second initial element is indicated by the label <1> in the A1C(X) pattern and the non-grounding function is indicated by the label <B> in the ABC(X) pattern. Unlike the <1> in the A1AC<X> pattern, the <1> in the A1C<X> pattern does not literally interrupt the A-satellite, but actually follows the A-satellite, occurring *between* the A and the C satellites. The A1C<X> and the ABC<X> patterns can thus be distinguished from each other on the basis of the functions of the <1> or <B> satellites respectively. Sentences (2) and (3) provide examples of the A1C(X) and ABC(X) patterns.

- (2) <A/zz>Met deze vraagstukken<A/zz><sup>40</sup>, <1>die soms van levensbelang waren<1>, <C>zagen velen in het vroegmoderne Europa zich geconfronteerd<C>. <s3662, academic prose>

(<A/zz>With these questions<A/zz>, <1>which sometimes of life importance were<1>, <C>saw many in the early modern Europe themselves confronted<C>.)

---

<sup>39</sup> For the purposes of her study Smits further distinguishes between several subclasses of both main types. The analysis of complex beginnings in the present study has, however, been restricted to a subdivision into the two main categories.

<sup>40</sup> Note that the first sentence-initial element receives the discourse label <A> if it also constitutes a punctuation unit and the label <zz> if it does not constitute a punctuation unit (cf. 2.4.3 & 5.5).

(*<A/zz>With these questions<A/zz>, <1>which were sometimes of life importance<1>, <C>many saw themselves confronted in early modern Europe<C>.*)

- (3) *<A>However<A>, <B>in all but the perfectly matched situation<B>, <C>the success rates of these transplants are equal to those of related donors<C>.*  
*<s7080, leaflets>*

As exemplified by sentence (2), the information contained in the interpolated satellite refers back to the information provided in the A/zz-element. More specifically, it further specifies it or grounds it. The information contained in the B-satellite in sentence (3), on the other hand, does not further specify or ground the A-element. Instead, it provides a separate orientation for the information contained in the nucleus.

Table 1 presents the frequencies of the different types of complex beginning: the A1AC(X) pattern, the A1C(X) pattern, the ABC(X) pattern and the sentences starting with three elements in sentence-initial position.<sup>41</sup>

---

<sup>41</sup> In line with Smits' study (2002: 39-41), sentences in which the satellites are realised as correlatives, such as *if...then*, as discourse markers or vocatives, or as coordinating conjunctions are excluded from the A1C(X) and ABC(X) patterns, and from the sentences starting with three satellites. With respect to the latter category, *yet* and *so* are included in the present analysis as having a conjunct-status when occurring in sentence-initial position (cf. 3.4.1 on the gradient between subordination and coordination). The exclusion of these sentences from the present analysis means the numbers as presented in this chapter deviate from the numbers as presented in Tables 10 and 47 of Chapter 5. If the numbers of Tables 10 and 47 are added up, sentences that start with two elements amount to 198 in English and 159 in Dutch, and sentences that start with three elements to 28 in English and 10 in Dutch. This means that for the present chapter 20 English sentences that start with two elements and 9 with three elements are excluded from the analysis, and the same applies to 37 Dutch sentences that start with two elements and 5 that start with three elements. It should furthermore be noted that cases in which the interpolated satellite in the A1C(X) pattern is realised as an apposition are included in the present study, whereas Smits restricts her analyses to the adverbials occurring in sentence-initial position (2002: 32-40).

**Table 1**      **Frequencies of complex beginnings: A1C(X), ABC(X), A1AC(X) and sentences starting with 3 elements**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
A1-C(X)	27 (31.4%)	19 (63.3%)	32 (66.7%)	14 (42.4%)	92 (46.7%)
AB-C(X)	39 (45.3%)	8 (26.7%)	12 (25.0%)	12 (36.4%)	71 (36.0%)
A1A-C(X)	9 (10.5%)	2 (6.7%)	1 (2.1%)	3 (9.1%)	15 (7.6%)
3 elements	11 (12.8%)	1 (3.3%)	3 (6.2%)	4 (12.1%)	19 (9.6%)
<b>Total</b>	<b>86 (100%)</b>	<b>30 (100%)</b>	<b>48 (100%)</b>	<b>33 (100%)</b>	<b>197 (100%)</b>
<b>Dutch</b>					
A1-C(X)	22 (57.9%)	14 (77.8%)	23 (69.7%)	8 (57.1%)	67 (65.0%)
AB-C(X)	8 (21.1%)	2 (11.1%)	3 (9.1%)	4 (28.6%)	17 (16.5%)
A1A-C(X)	6 (15.8%)	2 (11.1%)	5 (15.2%)	1 (7.1%)	14 (13.6%)
3 elements	2 (5.3%)	0 (0.0%)	2 (6.1%)	1 (7.1%)	5 (4.9%)
<b>Total</b>	<b>38 (100%)</b>	<b>18 (100%)</b>	<b>33 (100%)</b>	<b>14 (100%)</b>	<b>103 (100%)</b>

Table 1 shows that there are 197 complex beginnings in English in total, which means that 2.5% of all English sentences (8040) start with a complex beginning. In Dutch, on the other hand, there are 103 complex beginnings in total, which means that 1.2% of all Dutch sentences (8708) start with two or more satellites. When looking at these percentages per genre, it becomes clear that the English academic prose genre (1438 sentences) contains the highest percentage of sentences that start with a complex beginning (6.0%). Of the various genres in Dutch, the academic prose genre (1740) is also the genre that shows the highest percentage (2.2%). In English, the leaflets genre (1589 sentences) contains the second largest percentage (2.0%), followed by the newspaper genre (1844 sentences) (1.6%), and finally the short stories genre (1844 sentences) (1.5%). Dutch, on the other hand, shows different frequencies, with both the newspaper (1754 sentences) and the short stories genre (3154 sentences) containing 1.0% of sentences that start with a complex beginning. The Dutch leaflets genre (2060) shows the lowest percentage (0.7%) of sentences starting with a complex beginning.

The loglinear analysis showed no significant three-way interaction between language, genre and the frequencies of the various types of complex beginnings ( $\chi^2(9) = 5.94$ ,  $p = .74$ ), but it did show a significant two-way interaction between language and the distribution of the complex beginnings ( $\chi^2(3) = 16.31$ ,  $p < .001$ , Cramer's  $V = .24$ ). The main differences between the languages can be found in the occurrence of the ABC(X) pattern, which is more frequent in English than in Dutch (71 (36.0%) vs. 17 (16.5%)).

The following sections will look into the make up of each of these complex beginnings in more detail and illustrate them by providing a wide range of example sentences taken from the different genres.

## 6.3 ABC(X) and AIC(X) patterns

In addition to the discourse structure of the sentence-initial elements, the grammatical realisation of these elements also provides insight into the nature of these complex beginnings. This will look into the grammatical realisation of both sentence-initial satellites in the A1C(X) pattern (6.3.1) and the ABC(X) pattern (6.3.2).

### 6.3.1 Grammatical realisation of A and 1 in A1C(X)

#### Grammatical realisation of A in A1C(X)

In the A1C(X) pattern, the A-satellite can take the form of a phrase or a clause. When realised as a phrase, it can take the form of an adjunct or PP, a conjunct or a disjunct, and the clauses in initial position are all realised as adverbial clause of various types. Table 2 provides the frequencies of the different types of grammatical realisation of the A-satellite.

**Table 2** Grammatical realisation of A in A1C(X)

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Adjunct/PP	19	18	23	10	70 (76.0%)
Conjunct/disjunct	4	0	4	0	8 (8.7%)
Adverbial clause	4	1	5	4	14 (15.3%)
Total	27	19	32	14	92 (100%)
<b>Dutch</b>					
Adjunct/PP	17	14	21	5	57 (85.0%)
Conjunct/disjunct	0	0	0	0	0 (0.0%)
Adverbial clause	5	0	2	3	10 (15.0%)
Total	22	14	23	8	67 (100%)

Due to the fact that the expected frequencies of a number of cells are too low, no statistical tests can be carried out. Table 2 shows that the A-satellite takes the form of a phrase, realised as an adjunct or PP, in the vast majority of sentences in both languages (English: 70 (76.1%); Dutch: 57 (85.1%)) (cf. 5.5.1 & 5.7.1 for similar results in the AC and ACD subpatterns respectively). In both languages this applies especially to the newspaper genre (English: 18 out of 19; Dutch 14 out of 14). The other genres show somewhat more variation in the realisation of A, with the English academic prose and short stories genres also containing some instances of conjuncts and disjuncts (ac prose: 4 of 27; short stories: 4 of 32). Both these genres and the leaflets genre also have a number of adverbial clauses in this position. The English academic prose genre contains four adverbial clauses of different types, two of which are adverbial clauses of concession. The adverbial clauses that occur in the other three genres in English are all realised as adverbial clauses of condition. Dutch, on the other hand, shows no instances of conjuncts/disjuncts in this position, but it does show some instances of adverbial clauses, the vast majority of which are realised as adverbial clauses of condition.

#### **Grammatical realisation of 1 in A1C(X)**

Similar to the A-satellite, the interpolated satellite in the A1C(X) pattern can be realised as either a phrase or a clause. When realised as a phrase, it can take the form of (1) an adjunct or PP, (2) a conjunct or disjunct or (3) an apposition. It should be noted that the distinction between adjuncts/PPs on the one hand and non-nominal appositions on the other hand is not always clear-cut. This means that in certain cases the adjuncts/PPs could also have been classified as instances of non-nominal appositions (cf. 3.4.3 for a discussion on the gradient between these categories and for a definition of appositions in the present study). Besides being realised as a phrase, the interpolated satellite can also take the form of a clause. More specifically, it can take the form of a non-restrictive relative clause, an adverbial clause or an independent clause. Table 3 presents the frequencies of the different types of grammatical realisation of 1 in A1C(X).



**Table 3** Grammatical realisation of 1 in A1C(X)

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Adjunct/PP	6	6	9	3	24 (26.0%)
Conjunct/disjunct	9	0	3	8	20 (21.8%)
Apposition	4	3	3	1	11 (12.0%)
Non-restrictive relative clause	3	1	0	2	6 (6.5%)
Adverbial clause	5	8	17	0	30 (32.6%)
Independent clause	0	1	0	0	1 (1.1%)
Total	27	19	32	14	92 (100%)
<b>Dutch</b>					
Adjunct/PP	6	3	13	2	24 (35.8%)
Conjunct/disjunct	0	1	0	0	1 (1.5%)
Apposition	9	6	0	5	20 (29.8%)
Non-restrictive relative clause	3	3	1	1	8 (12.0%)
Adverbial clause	3	1	7	0	11 (16.4%)
Independent clause	1	0	2	0	3 (4.5%)
Total	22	14	23	8	67 (100%)

With respect to the realisation of the interpolated satellite, the languages show some differences. Specifically, in English a large group is formed by adverbial clauses (30 out of 92), most of which are non-finite (17). The second largest realisation group is formed by adjuncts/PPs (24 out of 92) and the third largest by conjuncts/disjuncts (20 out of 92). In Dutch, on the other hand, the largest realisation groups are formed by adjuncts/PPs (24 out of 67) and appositions (20 out of 67). Of the interpolated satellites that are realised as clauses in Dutch, 11 out of 67 take the form of adverbial clauses (of time, comparison or non-finite) and 8 out of 67 the form of non-restrictive relative clauses. Dutch contains only one instance of an interpolated satellite that takes the form of a conjunct/disjunct (see sentence (7) below).

With respect to the realisation of the interpolated satellite in the different genres, in the English academic prose genre the largest group is formed by conjuncts/disjuncts (9 out of 27), followed by adjuncts/PPs (6 out of 27). In the Dutch academic prose genre, on the other hand, most satellites are realised as appositions (9 out of 22), followed by adjuncts/PPs (6).

In the English newspaper genre, the majority of interpolated satellites take the form of an adverbial clause (8 out of 19), most of which are realised as non-finite adverbial clauses (5). The second largest realisation groups are formed by adjuncts/PPs (6) and appositions (3). In the Dutch newspaper genre the largest

realisation group is formed by appositions (6 out of 14), followed by adjuncts/PPs (3 out of 14) and non-restrictive relative clauses (3 out of 14).

In the English short stories genre the majority of interpolated satellites are realised as adverbial clauses (17 out of 32), most of which take the form of non-finite adverbial clauses (9) or adverbial clauses of time (7). The second largest group is formed by adjuncts/PPs (9). In Dutch, on the other hand, the situation is reversed, with adjuncts/PPs forming the largest realisation group (13 out of 23) and adverbial clauses (7), mostly of time (3) and non-finite (3), forming the second largest group.

Finally, in the English leaflets genre the majority take the form of a conjunct/disjunct (8 out of 14), followed by adjuncts/PPs (3). In Dutch, on the other hand, most of these satellites are realised as appositions (5 out of 8) or adjuncts/PPs (2).

Sentences (4) to (11) present examples of a variety of the most common syntactic realisations of both sentence-initial elements in the different genres of both languages.

- (4) (In the United States Joel Tarr, William Cronon, Martin Melosi and others have championed a city-oriented variant of a subdiscipline long dominated by a deep preoccupation with wilderness <4693, academic prose>.)  
 <A><adj>In Britain<adj><A>, <1><conj>by contrast<conj><1>, <C>powerfully established traditions in economic and social history have militated against the academic autonomy of a subject<C>, <D>which, <1>languishing as a minority interest<1>, remains wedded to predominantly scientific rather than unequivocally historical objectives<D>. <s4694, academic prose>
- (5) <A><advcl\_place>Waar het specifieke groepen kinderen betreft<advcl\_place><A>, <1><appos\_NP>zoals kinderen met leer-of gedragsproblemen of kinderen met een mentale of fysieke handicap<appos\_NP><1>, <C>wordt zelfconcept gekoppeld aan kwetsbaarheid, aangepastheid en geestelijke gezondheid<C> (Bracken, 1996; Van der Meulen, 1993; Veerman & Straathof, 1993, Verschuieren, Marcoen, & Schoefs, 1996). <s5981, academic prose>
- <A><advcl\_place>Where it specific groups of children concerns<advcl\_place><A>, <1><appos\_NP>such as children with educational or behavioural problems or children with a mental of physical handicap<appos\_NP><1>, <C>is self-image linked to vulnerability,

adaptability and mental health <C> (Bracken, 1996; Van der Meulen, 1993; Veerman & Straathof, 1993, Verschueren, Marcoen, & Schoefs, 1996).<sup>42</sup>

(<A><advcl\_place>Where it concerns specific groups of children<advcl\_place><A>, <1><appos\_NP>such as children with educational or behavioural problems or children with a mental of physical handicap<appos\_NP><1>, <C>is self- image linked to vulnerability, adaptability and mental health <C> (Bracken, 1996; Van der Meulen, 1993; Veerman & Straathof, 1993, Verschueren, Marcoen, & Schoefs, 1996).

- (6) <A><pp>By saying that the government is cutting industrial investment and aiming for economic expansion of 7% this year<pp><A>, <1><advcl\_nonfin>compared with over 9% growth last year<advcl\_nonfin><1>, <C>Mr Wen is trying to slow the country's breakneck speed and spread the fruits of prosperity beyond the coastal urban hotspots<C>. <s1145, newspaper articles>

- (7) <A/zz><adj>Tegelijk<adj><A/zz> <1/zz2><conj>echter<conj><1/zz2><sup>43</sup> <C>moeten ook de piloten erkennen dat een marshal een extra veiligheidsgarantie is<C>, <D>zowel voor passagiers als voor bemanning<D>. <s2044, newspaper articles>

(<A/zz><adj>At the same time<adj><A/zz> <1/zz2><conj>however<conj><1/zz2> <C>should also the pilots acknowledge that a marshal an extra safety guarantee is<C>, <D>both for the passengers and the crew<D>.)

(<A/zz><adj>At the same time<adj><A/zz> <1/zz2><conj>however<conj><1/zz2> <C>the pilots too should acknowledge that a marshal is an extra safety guarantee <C>, <D>both for the passengers and the crew<D>.)

- (8) <A/zz><adj>In the evening<adj><A/zz>, <1><advcl\_time>while I sagged feebly on a kitchen chair<advcl\_time><1>, <Ca>she breaded cutlets<Ca>, <Cb>and sliced tomato and cucumber on to a glass dish<Cb>. <s12269, short stories>

---

<sup>42</sup> The majority of the Dutch examples receive just one English translation. An exception to this is formed by cases in which differences between English and Dutch, for instance with respect to the position of the finite verb in the sentence, are relevant to the present discussion. In these cases the Dutch example will first be translated literally, which will be followed by a more idiomatic English translation.

<sup>43</sup> Similar to the <zz> label, the <zz2> label has been used to annotate those elements that occur in sentence-initial position but are not marked off by punctuation marks (cf. 2.4.3). With respect to two subpatterns, this pattern belongs to the A1C(X) pattern.

- (9) <A><pp>In de kleedkamer<pp><A>, <1><pp>onder de douche<pp><1>, <C>had hij zijn vraag herhaald<C>. <s15095, short stories>
- (<A><pp>In the locker room<pp><A>, <1><pp>in the shower<pp><1>, <C>had he his question repeated<C>.)
- (<A><pp>In the locker room<pp><A>, <1><pp>in the shower<pp><1>, <C>he had repeated his question<C>.)
- (10) <zz><adj>Maar tijdens de hele sollicitatieprocedure<adj><zz><C> - <1><pp>vanaf de advertentie tot en met het sluiten van de arbeidsovereenkomst<pp><1> - mogen mensen met een handicap of chronische ziekte niet ongelijk worden behandeld<C>. <s9151, leaflets>
- (<zz><adj>But during the whole application procedure<adj><zz><C> - <1><pp>from the advertisement to the signing of the contract<pp><1> - may people with a disability or chronic disease not unequally be treated<C>.)
- (<zz><adj>But during the whole application procedure<adj><zz><C> - <1><pp>from the advertisement to the signing of the contract<pp><1> - people with a disability or chronic disease may not be treated unequally <C>.)
- (11) <A/zz><pp>Despite their concerns<pp><A/zz>, <1><conj>however<conj><1>, <C>less than 1 in 10 of these companies had an official policy on mental health<C>. <s7242, leaflets>

These examples show that the grounding function of the interpolated satellite in the A1C(X) pattern often involves a further specification or clarification of the information contained in the preceding A-satellite, which explains the high frequency of interpolated satellites that take the form of adjunct/Ps, appositions and certain types of adverbial clauses or non-restrictive relative clauses. However, particularly in English, the interpolated satellite also takes the form of a conjunct in a large number of cases, which has the function of relating the information contained in both the A-satellite and the nucleus that it precedes to the preceding discourse (cf. Smits 2002: 89 on a special subtype of the A1C(X) pattern she identifies and labels 'complex orientations').

### 6.3.2 Grammatical realisation of A and B in ABC(X)

#### Grammatical realisation of A in ABC(X)

In the ABC(X) pattern, the A-satellite can take the form of a phrase or a clause. When realised as a phrase, this can take the form of (1) an adjunct or PP, or (2) a conjunct or disjunct. When realised as a clause, it can take the form of an adverbial clause of concession, condition, time or place, or a non-finite clause. Table 4 provides the frequencies of the different types of grammatical realisation of the A-satellite.

**Table 4** Grammatical realisation of A in ABC(X)

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Adjunct/PP	3	1	4	4	12 (16.9%)
Conjunct	25	4	6	6	41 (57.8%)
Disjunct	3	2	0	0	5 (7.0%)
Adverbial clause	8	1	2	2	13 (18.3%)
Total	39	8	12	12	71 (100%)
Dutch					
Adjunct/PP	0	0	2	0	2 (11.8%)
Conjunct	7	1	1	2	11 (64.8%)
Disjunct	1	0	0	0	1 (5.9%)
Adverbial clause	0	1	0	2	3 (17.5%)
Total	8	2	3	4	17 (100%)

Due to the detailed level of categorisation and the infrequent occurrence of the ABC(X) pattern in Dutch, the expected frequencies of a number of cells are too low to be tested statistically. Table 4 shows that in both languages the majority of A-satellites are realised as phrases that take the form of conjuncts (English: 41 of 71; Dutch: 11 of 17). In both languages this especially applies to the academic prose genre, in which 25 of 39 English sentences have conjuncts as the A-satellite and 7 of 8 Dutch sentences. In addition to conjuncts, the English academic prose genre also contains various types of adverbial clauses in this position (8), particularly those of concession (4), whereas Dutch shows no instances of clauses in this position in this genre. Besides the conjunct, the other genres show no particular preferences for any other grammatical category in this position. Because of the fact that the pattern has more occurrences in English than in Dutch, English shows more variation, but the numbers of the other categories are too low to detect any patterns.

### Grammatical realisation of B in ABC(X)

The B-satellite in the ABC(X) pattern can also take the form of a phrase or a clause. When realised as a phrase, this can take the form of (1) an adjunct or PP, or (2) a conjunct or disjunct. When realised as a clause, it can take the form of an adverbial clause of various types. Table 5 provides the frequencies of the different types of grammatical realisation of the B-satellite.

**Table 5** Grammatical realisation of B in ABC(X)

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Adjunct/PP	22	5	8	9	44 (62.0%)
Conjunct/disjunct	1	0	0	0	1 (1.4%)
Adverbial clause	16	3	4	3	26 (36.6%)
Total	39	8	12	12	71 (100%)
Dutch					
Adjunct/PP	7	1	2	2	12 (70.6%)
Conjunct/disjunct	0	0	0	1	1 (5.9%)
Adverbial clause	1	1	1	1	4 (23.5%)
Total	8	2	3	4	17 (100%)

In both languages the B-satellite takes the form of an adjunct or PP in the majority of cases across the different genres (English: 44 of 71; Dutch: 12 of 17). The ABC(X) pattern has the highest number of occurrences in the English academic prose genre and this is also the genre that shows most variation with respect to the realisation of the B-satellite. In addition to adjuncts/PPs, it is realised as an adverbial clause of various types in 16 out of 39 cases in this genre. These types range from clauses of concession (4), to condition (4), time (2), comparison (2) and non-finite clauses (4). In the other English genres, besides the adjuncts/PPs, no other grammatical categories dominate. Of the various clause types, the adverbial clause of concession and condition occur most frequently.

With the ABC(X) pattern being rather infrequent in Dutch, each of the categories shows low frequencies, with the adverbial clause category only containing four instances, i.e. one per genre, three of which are realised as adverbial clauses of condition.

Sentences (12) to (19) provide examples of the ABC pattern in English and Dutch in various genres.

(12) <A><conj>Similarly<conj><A>, <B><adj>in the actual crisis of May 1940<adj><B>, <C>it was Amery rather than Salisbury who occupied the centre-stage<C>, <D>as even Witherell's narrative tends to confirm<D>. <s4348, academic prose>

(13) <A><disj>Indeed<disj><A>, <B><advcl\_conces\_verbless>while to a degree true of all societies<advcl\_conces\_verbless><B>, <C>the inadequacy of national 'boxes' would seem to be especially and emphatically true of the 'neo-Britains'<C>. <s4366, academic prose>

(14) <A><conj>Integendeel<conj><A>, <B/zz><pp>met hun van binnenuit geformuleerde waarden<pp><B/zz>, <C>gedroeg deze groep zich eerder als een buffer<C>. <s3944, academic prose>

(<A><conj>On the contrary<conj><A>, <B/zz><pp>with their internally formulated values<pp><B/zz>, <C>acted this group more like a buffer<C>.)

(<A><conj>On the contrary<conj><A>, <B/zz><pp>with their values that were formulated internally<pp><B/zz>, <C>this group acted more like a buffer<C>.)

(15) <A><conj>Echter<conj><A>, <B><advcl\_cond>als er vaak en veel woede en frustratie bij die conflicten komt kijken<advcl\_cond><B>, <C>heeft dit negatieve effecten voor kinderen<C>. <s5966, academic prose>

(<A><conj>However<conj><A>, <B><advcl\_cond>if there often and much anger and frustration involved in these conflicts is<advcl\_cond><B>, <C>has this negative effects on the children<C>.)

(<A><conj>However<conj><A>, <B><advcl\_cond>if there is often much anger and frustration involved in these conflicts <advcl\_cond><B>, <C>this has negative effects on the children<C>.)

(16) <A><advcl\_conces>Hoeveel waardering er ook is in brede lagen van de samenleving voor het werk van Oudkerk<advcl\_conces><A>, <B/zz><pp>door de beschadigingen aan zijn persoon en zijn functie<pp><B/zz> <C>is het onontkoombaar dat hij aftreedt als wethouder<C>. <s2183, newspaper articles>

(<A><advcl\_conces>However much appreciation there is in large sections of the population for Oudkerk's work<advcl\_conces><A>, <B/zz><pp>because of the damage to him as a person and his position<pp><B/zz> <C>is it inevitable that he steps down as an alderman<C>.)

(<A><advcl\_conces>However much appreciation there is in large sections of the population for Oudkerk's work<advcl\_conces><A>, <B/zz><pp>because of

the damage to him as a person and his position<pp><B/zz> <C>it is inevitable that he steps down as an alderman<C>.)

(17) <A/zz><conj>Yet<conj><A/zz> <B/zz2><adj>soon<adj><B/zz2> <C>this scene would be annihilated by history<C>. <s13164, short stories>

(18) <A><advcl\_place>Where your employees are likely to be exposed to the second or peak action level or above<advcl\_place><A>, <B><advcl>so far as is reasonably practicable<advcl><B>, <C>reduce their exposure to noise in ways other than by providing hearing protection<C>. <6818, leaflets>

(19) <A><conj>Immers<conj><A>, <B><advcl\_cond>als uw keuze niet in het Donorregister staat<advcl\_cond><B>, <C>moet uw familie na uw overlijden een beslissing nemen<C>. <s8939, leaflets>

(<A><conj>After all<conj><A>, <B><advcl\_cond>if your choice is not in the Donor Registry listed<advcl\_cond><B>, <C>has to your family after your death make a decision<C>.)

(<A><conj>After all<conj><A>, <B><advcl\_cond>if your choice is not listed in the Donor Registry <advcl\_cond><B>, <C>your family has to make a decision after your death<C>.)

Sentences (12) to (19) can be distinguished from sentences (4) to (11) above on the basis of the function of the second initial satellite. Instead of grounding the first satellite, the second initial satellite in sentences (12) to (19) provides an orientation for the event as described in the nucleus. In all these examples the second initial satellite and the main clause fall within the scope of the sentence-initial satellite, which positions the sentence as a whole in the rhetorical structure of the text (cf. Smits 2002: 81, who refers to this subtype of the ABC(X) pattern as ‘stepwise orientations’). This rhetorical function is predominantly expressed by either conjuncts or disjuncts, hence the high frequency of the former category (see Table 4 above), or by adverbial clauses in initial position.

In a very small number of cases the first initial satellite does not provide an orientation for the second satellite and the information contained in the nucleus. Instead, both initial elements provide a separate orientation for the nucleus and could be analysed as being asyndetically coordinated (cf. Smits 2002: 81ff, who labels this subtype ‘compound orientations’). Sentence (20) provides an example, taken from the Dutch short stories genre.



(20) <A><np>Handen in de zakken<np><A>, <B><advcl\_nonfin>kraag  
opgeslagen<advcl\_nonfin><B>, <C>loopt hij de donkere weg naar het dorp<C>.  
<s15085, short stories>

(<A><np>Hands in pockets<np><A>, <B><advcl\_nonfin>collar turned  
up<advcl\_nonfin><B>, <C>walks he the dark road to the village<C>.)

(<A><np>Hands in pockets<np><A>, <B><advcl\_nonfin>collar turned  
up<advcl\_nonfin><B>, <C>he walks along the dark road to the village<C>.)

## 6.4 The A1AC(X) pattern

As described in Section 6.2 above, another type of complex beginning is formed by the sentence pattern in which the sentence-initial element, A or zz, is interrupted by an interpolated satellite, creating the A1AC(X) pattern. Although the wide range of sentence patterns created by interpolated satellites will be discussed in more detail in Chapter 8, this particular pattern will be described in the present chapter as it concerns the start of sentences. Table 6 repeats the frequencies of this particular pattern, as presented in Table 1 above.

**Table 6** Frequencies of A1AC(X) pattern

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
A1A-C(X)	9	2	1	3	15
<b>Dutch</b>					
A1A-C(X)	6	2	5	1	14

As Table 1 above already showed, the A1AC(X) pattern has a low frequency in both languages (English: 15; Dutch: 14). In English, this particular pattern makes up 7.6% of all sentences that start with a complex beginning (197) and in Dutch it makes up 13.6% of these sentences (103) (see Table 1). As Table 6 shows, most instances of the A1AC(X) pattern can be found in the academic prose genre in both languages (English: 9 out of 15; Dutch: 6 out of 14), with the Dutch short stories genre also showing a relatively high frequency (5).

The following sections will look at the grammatical realisation of both sentence-initial elements in the A1AC(X) pattern.

### 6.4.1 Grammatical realisation of A and 1 in the A1AC(X) pattern

#### Grammatical realisation of A in A1AC(X)

As has already been shown in 5.5.1, which describes the AC-subpattern, the A-satellite can be realised as either a phrase or a clause. With respect to the A1AC(X) pattern, when realised as a phrase, it can take the form of an adjunct or PP and when realised as a clause, it can take the form of various types of adverbial clauses. Table 7 presents the frequencies of both types of grammatical realisation of the A-satellite in the A1AC(X) pattern.

**Table 7** Grammatical realisation of A in A1AC(X) pattern

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Adjunct/PP	1	0	0	2	3
Adverbial clause	8	2	1	1	12
Total	9	2	1	3	15
<b>Dutch</b>					
Adjunct/PP	1	2	1	0	4
Adverbial clause	5	0	4	1	10
Total	6	2	5	1	14

In both languages the vast majority of A-satellites take the form of a clause (English: 12 out of 15; Dutch: 4 out of 14). In English, most of these clauses take the form of adverbial clauses of condition (5 out of 12). The English academic prose genre is the only genre that shows a relatively high frequency of adverbial clauses of concession (4 out of 8). Dutch, on the other hand, shows no instances of adverbial clauses of condition, but predominantly contains adverbial clauses of time and place (5 out of 10) or reason (3). The Dutch academic prose genre does contain two adverbial clauses of concession.

#### Grammatical realisation of 1 in A1AC(X)

The interpolated satellite in the A1AC(X) pattern can also be realised as a phrase or a clause. When realised as a phrase, it can take the form of (1) an apposition, (2) an adjunct/PP or (3) a disjunct. It should be noted, however, that the PPs could in certain cases also be analysed as appositions if they further specify or exemplify a phrase in the A-satellite that they interrupt (see similar note in Section 6.3.1 above). When realised as a clause, it can take the form of adverbial clause of various types

or of an independent clause. Table 8 presents the frequencies of these different grammatical realisations.

**Table 8** Grammatical realisation of 1 in A1AC(X) pattern

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Apposition	2	0	1	1	4 (26.7%)
Adjunct/PP	3	0	0	1	4 (26.7%)
Disjunct	1	0	0	0	1 (6.6%)
Adverbial clause	3	2	0	1	6 (40.0%)
Independent clause	0	0	0	0	0 (0.0%)
Total	9	2	1	3	15 (100%)
<b>Dutch</b>					
Apposition	3	1	0	1	5 (35.7%)
Adjunct/PP	3	0	3	0	6 (42.9%)
Disjunct	0	0	0	0	0 (0.0%)
Adverbial clause	0	1	1	0	2 (14.3%)
Independent clause	0	0	1	0	1 (7.1%)
Total	6	2	5	1	14 (100%)

In both languages the majority of interpolated satellites are realised as phrases instead of clauses (English: 9 out of 15; Dutch: 11 out of 14). These phrases either take the form of appositions or of adjuncts/PPs, although the distinction between these is not always clear-cut. English contains one instance of a phrase that is realised as a disjunct in the academic prose genre. When realised as a clause, this takes the form of a non-finite clause in the majority of cases in English (4 out of 6). Dutch only shows three instances of clauses, one of which occurs in the newspaper genre (adverbial clause of concession), and two of which in the short stories genre (one non-finite adverbial clause and one independent clause).

In addition to their grammatical realisation, the interpolated satellites can also be classified on the basis of their position with respect to the finite verb of the satellite that they interrupt, if that satellite is realised as a finite clause (cf. 2.5.3 on the position of interpolated satellites). As for the position of these interruptions with respect to the finite verb in the A-satellite, of the 12 A-satellites that are realised as finite adverbial clauses in English (see Table 7 above), 6 interpolated satellites occur before the finite verb the A-satellite and 6 after the finite verb. In Dutch, on the other hand, 8 of the interpolated satellites that interrupt a clausal A occur before the finite verb, whereas only 2 occur after the finite verb.

Sentences (21) to (25) provide examples of the A1AC(X) pattern taken from different genres, exemplifying a range of grammatical realisations of both the A and the 1 satellites.

- (21) <A><advcl\_conces>And although there seems, <1\_postfv><advcl\_compar>as Tosh argues<advcl\_compar><1\_postfv>, to have been a 'flight from domesticity' in the late Victorian era<advcl\_conces><A>, <C>Springthorpe's diaries demonstrate that many men still sustained their masculinity through an ordered and affirming private sphere<C>. <s4543, academic prose>
- (22) <Aa><advcl\_conces>Hoewel niet alle niet-rationele, automatische en onbewuste aspecten van het menselijk gedrag, <1\_prefv><appos\_NP>zoals de neiging negatieve lange-termijnevolgen van eigen gedrag te negeren<appos\_NP><1\_prefv>, vanuit evolutionair perspectief direct te verklaren zijn<advcl\_conces><Aa>, <Ab><advcl\_conces>en hoewel voor verschillende hier besproken aspecten de directe empirische evidentie voor evolutionaire verklaring ontbreekt<advcl\_conces><Ab>, <Ca>biedt een evolutionaire benadering wel een kader om op termijn beter te begrijpen waarom en welk gedrag meer en minder automatisch verloopt<Ca>, <Cb>en door onbewuste en ogenschijnlijk niet rationele factoren wordt bepaald<Cb>. <s6579, academic prose>

(<Aa><advcl\_conces>Although not all non-rational, automatic and subconscious aspects of human behaviour, <1\_prefv><appos\_NP>such as the tendency negative long term effects of one's own behaviour to ignore <appos\_NP><1\_prefv>, from an evolutionary perspective directly to explain are <advcl\_conces><Aa>, <Ab><advcl\_conces>and although for various presently described aspects the direct empirical evidence for evolutionary explanation is lacking<advcl\_conces><Ab>, <Ca>offers an evolutionary approach a framework to eventually better understand why and what behaviour more and less automatically occurs<Ca>, <Cb>and by subconscious and seemingly non-rational factors is determined<Cb>.)

(<Aa><advcl\_conces>Although not all non-rational, automatic and subconscious aspects of human behaviour, <1\_prefv><appos\_NP>such as the tendency to ignore negative long term effects of one's own behaviour <appos\_NP><1\_prefv>, can be explained from an evolutionary perspective <advcl\_conces><Aa>, <Ab><advcl\_conces>and although the direct empirical evidence for an evolutionary explanation is lacking for various of the currently described aspects<advcl\_conces><Ab>, <Ca>an evolutionary approach does provide a framework for a better understanding of why and what behaviour occurs more automatically and what behaviour occurs less

automatically <Ca>, <Cb>and is determined by subconscious and seemingly non-rational factors<Cb>.)

- (23) <A><advcl\_cond>If the figures - <1><advcl\_nonfin>published in a new Oxfam report, <i><appos\_NP>Dumping on the World</i><appos\_NP><i><sup>44</sup>, this week<advcl\_nonfin><1> - were applied to any other industry<advcl\_cond><A>, <C>they would be laughed out of court<C>. <s1257, newspaper articles>

- (24) <A><advcl\_reas>Aangezien de tijd waarin zij voor het eerst in het café verscheen, <1><adj>zo'n twee maanden geleden<adj><1>, precies overeenkwam met die van de aanschaf van haar flat<advcl\_reas><A>, <C>moest zij rond die periode gescheiden zijn<C>. <s15444, short stories>

(<A><advcl\_reas>Because the time in which she for the first time in the cafe appeared, <1><adj>about two months ago<adj><1>, exactly corresponded with that of the purchase of her flat<advcl\_reas><A>, <C>had to be she around that time divorced<C>.)

(<A><advcl\_reas>Because the time in which she appeared for the first time in the cafe, <1><adj>about two months ago<adj><1>, exactly corresponded with that of the purchase of her flat<advcl\_reas><A>, <C> she had to be divorced around that time<C>.)

- (25) <A><advcl\_cond>If your investigations (<1\_prefv><advcl\_nonfin>known as a risk assessment<advcl\_nonfin><1\_prefv>) show that there is a problem<advcl\_cond><A>, <C>the following section provides some helpful suggestions for reducing the risks<C>. <s7582, short stories>

## 6.5 Three elements in sentence-initial position

In addition to sentences starting with two elements, a very small number of sentences start with three elements in sentence-initial position, forming a range of subpatterns. A close analysis of the sentences with three sentence-initial elements yields the following six subpatterns: AB1BC(X), A12AC(X), A12C(X), A1BC(X), ABZC(X) and AB1C(X), and one sentence that is particularly complex because it has an A1a1b1c(X) subpattern (see example (32) below). Each of these subpatterns will

---

<sup>44</sup> Note that this sentence contains an interpolated satellite within an interpolated satellite. As these sentences are very infrequent, they will not be described or presented as constituting separate subpatterns.

be exemplified and further analysed on the basis of example (26) to (32) below. It should be noted that the frequencies of the various subpatterns are very low, with the exception of the ABZ pattern, to which 8 of the total number of 24 sentences belong. Table 9 below repeats the frequencies of the sentences that have three sentence-initial satellites that were already presented in Table 1 at the start of this chapter.

**Table 9**                      **Frequencies of sentences with 3 elements in sentence-initial position**

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
3 sentence-initial elements	11	1	3	4	19
<b>Dutch</b>					
3 sentence-initial elements	2	0	2	1	5

This table shows that sentences that start with three elements in sentence-initial position are very rare in both languages. This applies especially to Dutch, which has only 5 occurrences, compared to the 19 occurrences in English. In English most of these sentences occur in the academic prose genre (11 of 19), followed by the leaflets genre, which contains 4 instances. In Dutch the academic prose genre contains 2 examples; the short stories genre also contains 2 examples and the leaflets genre one. In both languages the newspaper genre shows the lowest frequencies, with only one occurrence in English and none in Dutch.

Sentences (26) to (32) below exemplify different subpatterns of the sentences that start with three elements, taken from different genres.

- (26) <A><adj>Immediately following devaluation, <1><advcl\_time>when Roy Jenkins, <i><appos\_NP>the new Chancellor<appos\_NP><i>, presented to senior ministers his proposals for further cuts in overseas defence spending<1> (<2><pp>with the full support of Wilson<pp><2>)<adj><A>, <Cr>he pushed home his point by remarking<Cr>: <s4255, academic prose> <Ca>We had come to the point of defeat on the economic road<Ca>, <Cb>and unless we took measures of the kind that he was proposing he saw no prospect of success for many years ahead<Cb>. <s4256, academic prose>
- (27) <A><conj>Thus<conj><A>, <B><advcl\_cond>if a child correctly positions a toy bottle near the mouth of a doll<advcl\_cond><B>, <1><advcl\_nonfin>as if feeding the doll<advcl\_nonfin><1>, <C>this is functional play<C>. <s5394, academic prose>

Both sentences (26) and (27) are taken from the English academic prose genre. Sentence (26) could be analysed as following an A1i12C(X) pattern, which is categorised with the four other sentences that follow an A12C(X) pattern. It starts with an A-satellite that is followed by an interpolated satellite, which has a grounding function, similar to the interpolated satellites in the A1C(X) pattern described above (cf. 6.3.1). In this particular example, the interpolated satellite is itself also interrupted by an interpolated satellite that is lower in hierarchy, indicated by <i>, and realised as an apposition. The interpolated satellite <1> is followed by another element that could be analysed as a second interpolated satellite, mainly because it appears to further ground the first interpolated satellite. Furthermore, sentence (27) is analysed as having an AB1C(X) pattern, which is again similar to the A1C(X) pattern, but instead of grounding the A-satellite, the interpolated satellite in this pattern grounds the B-satellite. There are three sentences that belong to this pattern.

Now consider sentence (28):

- (28) <A><conj>Kortom<conj><A>, <B><advcl\_conces>ook al was Nederland een 'oude' natie<advcl\_conces><B>, <zz>in het Interbellum evenals rond 1900<Z/zz> <C>viel een sterke etnonationale impuls aan te treffen<C>. <s5234, academic prose>

(<A><conj>In short<conj><A>, <B><advcl\_conces>even though was the Netherlands an 'old' nation<advcl\_conces><B>, <zz>in the Interbellum and around 1900<Z/zz> <C>could be a strong ethno-national impulse found<C>.)

(<A><conj>In short<conj><A>, <B><advcl\_conces>even though the Netherlands was an 'old' nation<advcl\_conces><B>, <zz>in the Interbellum and around 1900<Z/zz> <C>a strong ethno-national impulse could be found<C>.)

This example is taken from the Dutch academic prose genre and is classified as having an ABZ/zzC(X) pattern, which is the most frequently occurring subpattern (8 instances). In (28) the A-satellite is realised as a conjunct, the B-satellite as an adverbial clause of concession and the Z/zz-element as an adjunct/PP, situating the event in time. Sentence (29) below is also classified as following the ABZ/zzC(X) pattern, and is taken from the Dutch short stories genre:

- (29) <A><adj>Tergend langzaam<adj><A>, <B><NP>het bekken naar voren<NP><B>, <Z><NP>het hoofd in de nek<NP><Z>, <C>naderde hij de stier<C>, <Da>die niet bewoog<Da>, <Db>alleen maar ademde<Db>. <s15942, short stories>

(<A><adj>Painfully slowly<adj><A>, <B><NP>the pelvis forward<NP><B>,  
<Z><NP>head in the neck<NP><Z>, <C>approached he the bull<C>,  
<Da>which didn't move<Da>, <Db>just breathed<Db>.)

(<A><adj>Painfully slowly<adj><A>, <B><NP>the pelvis forward<NP><B>,  
<Z><NP>head in the neck<NP><Z>, <C>he approached the bull<C>,  
<Da>which didn't move<Da>, <Db>just breathed<Db>.)

Even though sentences (28) and (29) can be classified as following the same discourse structure, there is a difference in the way in which the sentence-initial elements are related to each other and to the information contained in the nucleus they precede. In the examples as presented in (28), to which most ABZ/zzC(X) sentences belong, it is typically the function of the first sentence-initial satellite to indicate how this sentence as a whole fits in the rhetorical structure of the text. In line with this function, the A-satellite often takes the syntactic form of a conjunct (6 out of 8). In sentence (29), on the other hand, all three sentence-initial satellites present orientations for the information contained in the nucleus. In this particular example, the satellites describe the manner in which the subject of the sentence approaches the bull. In this respect, all three elements are related to the nucleus in a similar way (Smits refers to this particular pattern as 'compound orientation' (2002: 81ff)). Note that the discourse pattern of this type of sentence could also have been analysed as following an AaAbAcC(X) pattern, in which the three satellites are analysed as being asyndetically coordinated to each other. Only two sentences belong to this pattern, both of which occur in the Dutch short stories genre.

Consider sentence (30), which presents an example of the AB1BC(X) subpattern. This is one of two sentences that follow this pattern.

(30) <A><conj>Yet<conj><A> <B><pp>with one opposition party, <1><appos>the Conservatives<appos><1>, already led by a 62-year-old QC<pp><B>, <C>the case for choosing a second, <1>Sir Menzies<1>, is weak<C>. <s1466, newspaper articles>

Note that sentence (30) presents the only example in the newspaper genre of a sentence that starts with three sentence-initial elements. The A-satellite is realised as a conjunct, the B-satellite as a PP and the interpolated satellite in B as an apposition.



Sentences (31) below follows the A/zz1BC(X) subpattern, to which three sentences in total belong. It is taken from the English leaflets genre.

- (31) <A/zz><adj>Often<adj><A/zz> <1><conj>however<conj><1>, <B><pp>by working together<pp><B>, <C>teachers, parents and pupils are able to resolve the incidents themselves much more quickly and satisfactorily than if the police or courts were involved<C>. <s7533, leaflets>

Sentence (32) below presents the only example of a sentence that could be classified as following an A-1a-1b-1c-i-1c-C(X) pattern, with the interpolated satellites forming a list of three, in which the third element is interrupted by an interpolated satellite at a lower level of the hierarchy, <i>.

- (32) <A><advcl\_cond>Als je risico loopt<advcl\_cond><A>, <1a><advcl\_reas>bijvoorbeeld omdat je vaste partner hepatitis-B heeft<advcl\_reas><1a>, <1b><pp>bij intraveneus drugsgebruik<pp><1b>, <1c><advcl\_cond>of als je veel wisselende (<i><premod>homo<premod><i>) seksuele contacten hebt<advcl\_cond><1c>, <C>is er de mogelijkheid tot vaccinatie<C>. <s9917, leaflets>

(<A><advcl\_cond>If you're at risk<advcl\_cond><A>, <1a><advcl\_reas>for example because your partner has hepatitis-B<advcl\_reas><1a>, <1b><pp>with intravenous drug use<pp><1b>, <1c><advcl\_cond>or if you many changing (<1><premod>homo<premod><1>) sexual contacts have<advcl\_cond><1c>, <C>is there the possibility of vaccination<C>.)

(<A><advcl\_cond>If you're at risk<advcl\_cond><A>, <1a><advcl\_reas>for example because your partner has hepatitis-B<advcl\_reas><1a>, <1b><pp>with intravenous drug use<pp><1b>, <1c><advcl\_cond>or if you have many changing (<1><premod>homo<premod><1>) sexual contacts<advcl\_cond><1c>, <C>there is the possibility of vaccination<C>.)

Sentence (32) could be considered particularly complex, not only because of the three elements in sentence-initial position, but mainly because the list of coordinated elements that make up the interpolated satellite do not form a parallel structure. Specifically, the first coordinate is realised as an adverbial clause of reason, the second as a PP and the third as an adverbial clause of condition. This last coordinate is further interrupted by an interpolated satellite lower in hierarchy (i), which is realised as a premodifier, a type of interruption that occurs predominantly in Dutch (cf. Chapter 8, Section 8.3).

Finally, sentence (33) presents one of two examples of the A12AC(X) pattern, taken from the Dutch academic prose genre. This pattern closely resembles the A1AC(X) pattern as described in Section 6.4 above, except for the extra interpolated satellite that interrupts the A-satellite. In (33) both interpolated satellites are realised as appositions.

- (33) <A><advcl\_compar>Zoals NIOD-directeur en eindverantwoordelijke voor het rapport, <1><appos>J.C.H. Blom<appos><1>, in de inleiding ('<2><appos>proloog<appos><2>') stelt<advcl\_compar><A>, <C>kon de onderzoeksgroep grotendeels zelf bepalen welke elementen zij relevant achtte<C>. <s3812, newspaper articles>

(<A><advcl\_compar>As NIOD-director and the person responsible for the report, <1><appos>J.C.H. Blom<appos><1>, in the introduction ('<2><appos>prologue<appos><2>') states<advcl\_compar><A>, <C>could the research group largely determine what elements they considered relevant<C>.)

(<A><advcl\_compar>As NIOD-director and the person responsible for the report, <1><appos>J.C.H. Blom<appos><1>, states in the introduction ('<2><appos>prologue<appos><2>') <advcl\_compar><A>, <C>the research group could largely determine what elements they considered relevant<C>.)

## 6.6 Conclusion

A closer analysis of the sentences starting with either two or three satellites before the nucleus showed that these satellites form a range of subpatterns. Sentences starting with two satellites can be further categorised into three subpatterns: A1C(X), ABC(X) and A1AC(X). Sentences starting with three satellites follow these same patterns, but allow for more variation because of the extra satellite. Both sentences that start with two and three satellites are significantly more frequent in English than in Dutch and both languages show more instances of sentences that start with two satellites than sentences that start with three satellites.

With respect to the distribution of the different subpatterns, the main difference between the languages can be found in the frequency of the ABC(X) pattern, which occurs significantly more frequently in English than in Dutch. It should, however, be noted that this characteristic of Dutch does not merely involve a limit on the number of elements that precede the finite verb, but also on the type

of elements and the hierarchical relation between them. As the interpolated satellite in the A1C(X) pattern actually just functions as in interruption of the A-satellite, but then one that is positioned at the very end of the satellite instead of literally interrupting it like in the A1AC(X) pattern (cf. 6.4), the combination A1 can be seen as constituting one complex A-satellite instead of two elements (see Haeseryn et al. 1997: 1297 for a similar analysis). It is for this reason not surprising that Dutch easily allows for this sentence pattern (cf. see 6.2). The occurrence of the ABC(X) pattern in Dutch is in this respect more remarkable, as this really consists of two separate satellites that do not function as one unit. However, the occurrence of this pattern does not seem to be solely dependent on language, but also on the genre within the language.

In both languages it is the academic prose genre that shows by far the highest relative frequency of sentences starting with complex beginnings and this applies especially to the English academic prose genre. In English, although showing a much lower frequency than the academic prose genre, the leaflets genre is the second genre to contain most instances of complex beginnings, further followed by the newspaper genre and finally by the short stories genre. Dutch presents a different order, with the newspaper genre following the academic prose genre and the leaflets genre containing the lowest number of complex beginnings.

As for the grammatical realisation of the different satellites in the A1C(X) and ABC(X) patterns, differences between the languages were mainly found for interpolated satellite in the A1C(X) pattern. In the ABC(X) pattern the A-satellite is predominantly realised as a conjunct and the B-satellite is typically as an adjunct/PP in both languages. As the English academic prose genre contains most instances of this pattern, it is not surprising that this also shows more variation in the grammatical realisation of both satellites. In the A1C(X) pattern, on the other hand, the A-satellite is predominantly realised as an adjunct/PP in both languages, particularly in the newspaper genre, with the English academic prose genre and short stories genre also showing a few instances of conjuncts and disjuncts in this position. The interpolated satellite, which has a grounding function in this pattern, is in English predominantly realised as a non-finite adverbial clause, followed by adjuncts/PPs and thirdly by conjuncts/disjuncts. In Dutch, on the other hand, it is predominantly realised as either an adjunct/PP or an apposition, showing only one instance of a conjunct. Differences can, however, be found between the various genres. Both the English academic prose and leaflets genres show a relatively high frequency of conjuncts/disjuncts, whereas these genres in Dutch show a high frequency of appositions in this position. In the English newspaper and short stories

genre, the largest realisation group is formed by adverbial clauses, whereas in Dutch this is again the apposition in the newspaper genre and adjuncts/PPs in the leaflets genre.

As for the third subpattern of complex beginnings consisting of two satellites, the A1AC(X) pattern shows few occurrences in both languages, with the academic prose genre showing the highest frequency. In both languages the A-satellite is predominantly realised as an adverbial clause, in English mainly of condition and in Dutch mainly of reason or time. The interpolated satellite, on the other hand, is mainly realised as a phrase in both languages, taking the form of an apposition or adjunct/PP. With respect to the position of the interruption in relation to the finite verb in the clausal A-satellites, in Dutch these occur mainly before the finite verb, whereas in English the distribution between those occurring before and those occurring after the finite verb is even.

Finally, sentences that start with three satellites are rare in both languages, but more frequent in English than in Dutch, again particularly in the academic prose genre. The different satellites form a range of seven different subpatterns, many of which are extended forms of the patterns formed by two satellites. Of the different subpatterns, the ABZC(X) pattern is the one that occurs most frequently.



## 7. Sentence patterns and punctuation

### 7.1 Introduction

As punctuation formed one of the main criteria, in addition to syntactic and semantic ones, for identifying Sentence Information Units (SIUs, cf. 2.3.3), the use and occurrence of a wide range of punctuation marks has been closely analysed in this study. Situations in which the use of particular punctuation marks led to certain issues or difficulties in the segmentation of discourse into units or in the grammatical categorisation of these units have already been addressed in Chapters 2 and 3. Some of these constituted more straightforward cases, such as examples in which a comma is used serially to separate a number of phrases and therefore not considered to mark unit boundaries (cf. 2.4.3), whereas others constituted more complex cases, for instance when the hierarchical status of a particular discourse unit is partly dependent on the punctuation unit by which it is introduced. Consider in this respect example (1) below, which was presented earlier in Chapter 2 as (75):

- (1) <C>Mr Blair's case is this<C>: <D>Universities need more money<D>. <s215, newspaper articles>

This example is analysed in this study as consisting of two SIUs, the first of which has nuclear status and the second satellite status. This decision is largely based on the fact that these two sentences are separated by means of a colon (cf. Chapter 2, 2.5.1). Precisely because the colon marks 'an elaborative rather than coordinative interpretation' in cases where two units are linked asyndetically (Huddleston & Pullum 2002: 1743, but see also Quirk et al. 1985: 1620; Onrust et al. 1993: 194), this sentence is analysed as having a nucleus-satellite pattern instead of a coordinated nuclei pattern (cf. 7.4 and 7.5 on type of relation introduced by semi-colon and dash respectively).

Not only did the close analysis of punctuation marks in the discourse segmentation process show that punctuation plays an important role in signalling the hierarchical and rhetorical relations between different units, it also created the impression that punctuation is used to different effects in English and Dutch and the four different genres within these languages. This appeared to apply to three punctuation marks in particular: the colon, the semi-colon and the dash. For this

reason, the present chapter will analyse the use and occurrence of these three punctuation marks in more detail. It will focus on the use of these punctuation marks in signalling the hierarchical and rhetorical relations between different discourse units.

The chapter will start with an overview of the frequencies of these three punctuation marks in the two languages and four genres, which will be followed by an analysis of the use and occurrence of each of these three marks in more detail. The chapter will end with a brief discussion of the occurrence of the comma splice, which concerns the use of a comma that is typically classified as stylistically inappropriate.

## 7.2 Colons, semi-colons and dashes

This section provides an overview of the frequencies of the colon, the semi-colon and the dash in English and Dutch in four different genres. As was already explained in 7.1, in those cases where the semi-colon or dash separates the items on a list or introduces a bullet point list have been excluded from the present analysis. A particular use of semi-colons in the academic prose genre, to separate a list of references from each other, has also been excluded from the analysis. Furthermore, the present analysis has been restricted to the occurrence of single dashes between nuclei and appended satellites (cf. Chapter 8 on the use and occurrence of paired dashes around interpolated satellites). As for colons, those that introduce reported speech have also been excluded from the analysis. Table 1 presents an overview of these frequencies of the three punctuation marks under analysis.

**Table 1** Frequencies of colons, semi-colons and dashes

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
colons	31 (44.9%)	16 (17.4%)	34 (22.2%)	54 (51.4%)	135 (32.2%)
semi-colons	16 (23.2%)	12 (13.0%)	71 (46.4%)	10 (9.5%)	109 (26.0%)
dashes	22 (31.9%)	64 (69.6%)	48 (31.4%)	41 (39.0%)	175 (41.8%)
<b>Total</b>	<b>69 (100%)</b>	<b>92 (100%)</b>	<b>153 (100%)</b>	<b>105 (100%)</b>	<b>419 (100%)</b>
<b>Dutch</b>					
colons	76 (76.8%)	41 (70.7%)	51 (58.0%)	99 (87.6%)	267 (74.6%)
semi-colons	16 (16.2%)	14 (24.1%)	25 (28.4%)	11 (9.7%)	66 (18.4%)
dashes	7 (7.1%)	3 (5.2%)	12 (13.6%)	3 (2.7%)	25 (7.0%)
<b>Total</b>	<b>99 (100%)</b>	<b>58 (100%)</b>	<b>88 (100%)</b>	<b>112 (100%)</b>	<b>358 (100%)</b>

Table 1 shows that even though all three punctuation marks occur in both languages, the dash appears to be particularly frequent in English (English: 175 vs. Dutch: 25), whereas the colon appears to be particularly frequent in Dutch (Dutch: 267 vs. English: 135). In relation to the total number of sentences, this means that 3.1% of all Dutch sentences contain a colon (8708), compared to 1.7% of all English sentences (8040), and that 2.2% of all English sentences contain a single dash, compared to 0.3% of all Dutch sentences. Moreover, it should be noted that the dash is particularly frequent in the English newspaper genre, with 3.5% of all sentences containing a dash (1844). Finally, the semi-colon occurs in 109 sentences in English, making up 1.4% of all sentences, compared to 66 sentences in Dutch, making up 0.8% of all sentences.

In testing these perceived differences statistically, the loglinear analysis showed a significant three-way interaction between language, genre and the occurrence of these three types of punctuation marks ( $\chi^2(6) = 20.32$ ,  $p < .01$ ). Subsequent chi-square tests showed a difference in the frequency of the different types of punctuation marks between English and Dutch for all genres (academic prose:  $\chi^2(2) = 22.03$ ,  $p < .001$ , Cramer's  $V = .36$ ; newspaper articles:  $\chi^2(2) = 62.14$ ,  $p < .001$ , Cramer's  $V = .64$ ; short stories:  $\chi^2(2) = 31.83$ ,  $p < .001$ , Cramer's  $V = .36$ , and leaflets:  $\chi^2(2) = 45.90$ ,  $p < .001$ , Cramer's  $V = .46$ ).

In the academic prose genre the main differences between the languages can be found in the frequency of the dash, which is far more frequent in English (22 (31.9%)) than Dutch (7 (7.1%)), and the colon, which is more frequent in Dutch (76 (76.8%)) than in English (31 (44.9%)).

In the newspaper genre the main differences between the languages can be found for both the colon and the dash, but in this genre it is even more pronounced than in the academic prose genre. Specifically, the colon is again much more frequent in Dutch (41 (70.7%)) than in English (16 (17.4%)), whereas the latter is far more frequent in English (64 (69.6%)) than in Dutch (3 (5.2%)).

In the short stories genre it is again the colon that shows particularly high frequencies in Dutch (51 (58.6%)) when compared to English (34 (22.2%)). The dash shows higher in English (48 (31.4%)) than in Dutch (12 (13.6%)), but this difference is not as pronounced as the difference the occurrence of the colon in the two languages.

Finally, in the leaflets genre the main differences between the languages can also be found in the occurrence of the dash and the colon, with differences between the languages being most pronounced for the former punctuation mark. The former is again much more frequent in English (41 (39.0%)) than in Dutch (3



(2.7%) and the latter is more frequent in Dutch (99 (87.6%)) than in English (54 (51.4%)).

The following sections will look into the use of each of these three punctuation marks in more detail.

## 7.3 The use of colons

Section 7.2 above already showed the colon is particularly frequent in Dutch in all genres when compared to its occurrence in English. This section will look in more detail at the use of the colon in both languages. Specifically, it will look at the type and hierarchical status of the discourse units linked by colons. It will then provide an overview of the grammatical realisations of the discourse units occurring before and after the colon.

### 7.3.1 Type and status of units linked by colon

An analysis of the use and occurrence of the colon shows that they typically occur in three positions in a sentence, creating the following three patterns: A:C(X), (X)C:D(X), (X)CD:E(X). Sentences (2) to (4) provide examples of these positions of the colon in sentences, and Table 2 below provides the frequencies of these different positions.

- (2) <A>Wat betreft de verschillen tussen de accenten onderling<A>: <C>deze waren over de hele linie zeer gering<C>. <s6669, academic prose>
- <A>With respect to the differences between the different accents<A>: <C>these were very slight across the board<C>.)
- (3) <Ca>Less can be said about why lower-class London men became less violent<Ca>, <Cb>but they clearly did so<Cb>: <D>only a fraction of the decline in male homicide can be explained by the decline in the number of gentlemen accused<D>. <s4460, academic prose>
- (4) <C>Ik heb daar altijd een opgemaakt bed klaarstaan voor wat ik 'de vrouwelijke persoon' noem<C>, <D>al is het niet veel meer dan een symbolisch leger<D>: <Ea>net als dat van mij een nest van krantenpapier<Ea>, <Eb>maar niet veel groter dan een hondenmand<Eb>. <s14644, short stories>

(<C>I always have a bed ready over there for what I call the 'female person' <C>, <D>though it is not more than a symbolic bed<D>: <Ea>just like mine a nest of newspapers<Ea>, <Eb>but not much bigger than a dog's basket<Eb>.)

**Table 2** Frequencies of three main positions of colons in sentences

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
A : C(X)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
(X)C : D(X)	26 (83.9%)	16 (100%)	30 (88.2%)	53 (98.1%)	125 (92.6%)
(X)CD : E(X)	5 (16.1%)	0 (0.0%)	4 (11.8%)	1 (1.9%)	10 (7.4%)
Total	31 (100%)	16 (100%)	34 (100%)	54 (100%)	135 (100%)
<b>Dutch</b>					
A : C(X)	3 (3.9%)	3 (7.3%)	2 (3.9%)	7 (7.1%)	15 (5.6%)
(X)C : D(X)	67 (88.2%)	38 (92.7%)	42 (82.4%)	92 (92.9%)	239 (89.5%)
(X)CD : E(X)	6 (7.5)	0 (0.0%)	7 (13.7%)	0 (0.0%)	13 (4.9%)
Total	76 (100%)	41 (100%)	51 (100%)	99 (100%)	267 (100%)

Due to the low occurrence of certain patterns, the expected frequencies of a number of cells are too low to be tested statistically. For this reason the differences in frequencies will not be tested, but will only be described. Table 2 shows that in both languages the colon typically occurs between the nucleus and the first appended satellite, the D (English: 125 (92.6%); Dutch: 239 (89.5%)). Both languages also show some instances of the colon occurring between the D and E appended satellites (English: 10 (7.4%); Dutch: 13 (4.9%)), which in both languages mainly occurs in the academic prose genre and the short stories genre. The main difference between the languages can be found in the occurrence of the A:CX pattern, which has no occurrences at all in English and occurs in 15 sentences (5.6%) in Dutch, divided fairly evenly across the different genres.

### 7.3.2 Grammatical realisation of discourse units surrounding colons

This section will look at the grammatical realisation of the discourse units surrounding the colon. It will start with the most frequent pattern, in which the colon occurs between the nucleus and the D satellite. It will then briefly describe the pattern in which the colon occurs between the A-satellite and the nucleus, and will finish with the colon occurring between the D and E appended satellites.

### Grammatical realisation of nucleus and the D satellite in (X)C:D(X)

The various grammatical forms of the nucleus in the (X)C:D(X) pattern have been categorised into three main realisation groups: independent clauses, fragments and fragment\_complements. This latter realisation group mainly occurs in the leaflets genre and concerns those sentences that have, as it were, been cut into two parts by the colon, with the part following the colon forming a complement of the verb of the nucleus that precedes the colon. The part before the colon has received the label fragment\_complement(list) (cf. 3.5.1 and sentence (8) below) and the part following the colon the label that indicates whatever grammatical form the complement takes. It should be noted that this complement typically takes the form of a list (see realisation of D-satellite in Table 4 below). As for the second realisation group, the label 'fragment' has been used to classify various types of phrase or clause fragments (see 3.5.1 for the definition of fragment in this study). Most of these are realised as noun phrases (NPs), especially in the short stories and leaflets genres, with a few taking the form of an adverbial clause in the Dutch academic prose genre (3 of the 4). Sentences (5) to (8) present examples of the different realisation forms of the nucleus in this pattern and Table 3 contains the frequencies of the three different realisation forms.

- (5) <C><indepcl>Goodnow and Collins argue that the extant research suggests that parents distinguish between three kinds of intelligence<indepcl><C>: <D><appos\_NP\_list>knowledge about things, abstract problem-solving ability and social intelligence<appos\_NP\_list><D>. <s5241, academic prose>
- (6) <C><fragment\_NP>Vals alarm<fragment\_NP><C>: <D><indepcl>dat is vervelend<indepcl><D>. <s2278, newspaper articles>
- ( <C><fragment\_NP>False alarm<fragment\_NP><C>: <D><indepcl>that is annoying<indepcl><D>.)
- (7) <C><fragment\_VP\_list>Rondrennen, de aandacht ergens niet bij kunnen houden, niet luisteren<fragment\_VP\_list><C>: <D><indepcl>het hoort bij kinderen<indepcl><D>. <s8447, leaflets>
- ( <C><fragment\_VP\_list>Running around, not staying focused, not listening<fragment\_VP\_list><C>: <D><indepcl>that is what children do<indepcl><D>.)

(8) <C><fragment\_comp\_list>De bekendste zijn<fragment\_comp\_list><C>:  
<D><appos\_NP\_list>nachtmerries, slaapwandelen en praten of  
tandenknarsen in de slaap<appos\_NP\_list><D>. <s8420, leaflets>

(<C><fragment\_comp\_list>The most familiar ones are<fragment\_comp\_list><C>:  
<D><appos\_NP\_list>nightmares, sleepwalking and talking or teeth grinding in  
one's sleep<appos\_NP\_list><D>.)

**Table 3** Grammatical realisation of C in (X)C:D(X)

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Independent clause	21 (80.8%)	15 (93.8%)	23 (76.7%)	19 (35.8%)	78 (62.4%)
Fragment	2 (7.7%)	1 (6.2%)	6 (20.0%)	1 (1.9%)	10 (8.0%)
Fragment_complement	3 (11.5%)	0 (0.0%)	1 (3.3%)	33 (62.3%)	37 (29.6%)
Total	26 (100%)	16 (100%)	30 (100%)	53 (100%)	125 (100%)
Dutch					
Independent clause	61 (91.0%)	33 (86.8%)	31 (73.8%)	55 (59.8%)	180 (75.3%)
Fragment	4 (6.0%)	5 (13.2%)	5 (11.9%)	13 (14.1%)	27 (11.3%)
Fragment_complement	2 (3.0%)	0 (0.0%)	6 (14.3%)	24 (26.1%)	32 (13.4%)
Total	67 (100%)	38 (100%)	42 (100%)	92 (100%)	239 (100%)

The loglinear analysis showed no significant three-way interaction between language, genre and the realisation of the nucleus in the (X)C:D(X) pattern at the predetermined alpha level of .01 ( $\chi^2(6) = 14.10$ ,  $p = .03$ ). The analysis did show a significant two-way interaction between language and the realisation of the nucleus in this pattern ( $\chi^2(2) = 13.44$ ,  $p < .001$ , Cramer's  $V = .20$ ). The main difference between the languages can be found in the high frequency of the fragment\_complement category in English (37 (29.6%)) compared to Dutch (32 (13.4%)). As was already noted above, this grammatical realisation of the nucleus is particularly frequent in the leaflets genre, in which the colon introduces a list of phrases or clauses (English: 33 (62.3%); Dutch: 24 (26.1%)). Moreover, Table 3 also shows that in both languages the vast majority of nuclei are realised as independent clauses (English: 62.4%); Dutch: 180 (75.3%)).

Although not indicated by the loglinear analysis as a significant difference between the languages, there is a use of fragments that is particular to Dutch. In these cases the nucleus takes the form of a fragment, typically realised as an NP, and followed by an appended satellite that takes the form of an independent clause. Sentence (6) above already provided an example of such a fragment and sentences (9) and (10) below present two further examples, taken from the Dutch newspaper genre and leaflets genre respectively.

- (9) <C><fragment\_NP>Het gevolg<fragment\_NP><C>: <D><indepcl>vrijwel alle commercial radio- en tv-stations zijn in meerderheid in handen van buitenlandse uitgevers gekomen<indepcl><D>. <s1990, newspaper articles>
- (<C><fragment\_NP>The effect<fragment\_NP><C>: <D><indepcl>almost all commercial radio- en tv-stations are for the majority in the hands of foreign publishers<indepcl><D>.)
- (10) <C><fragment\_NP\_list>Conflicten op het werk, een hoge werkdruk, zorgen over geld of de opvoeding van kinderen, relatieproblemen<fragment\_NP\_list><C>: <D><indepcl>het zijn allemaal spanningsbronnen waar mensen wakker van kunnen liggen<indepcl><D>. <s8435, leaflets>
- (<C><fragment\_NP\_list>Conflicts at work, heavy workload, worries about money or about child raising, relationship problems<fragment\_NP\_list><C>: <D><indepcl>these are all sources of tension that can keep people awake<indepcl><D>.)

The various grammatical forms of the D-satellite in the (X)C:D(X) pattern have been categorised into four main realisation groups: independent clauses, appositions, fragments and lists. The elements in the list can take various grammatical forms, i.e. they can form a list of phrases that together form an apposition (apposition\_list); they can form a list of phrase or clause fragments, or they can form a series of coordinated phrases or clauses (see examples (11) and (12) below). Sentences (11) to (15) will provide examples of the various realisation forms of the D-satellite, taken from various genres. Table 4 below provides the frequencies of the different realisation forms of the D-satellite.

- (11) <C><indepcl>Goodnow and Collins argue that the extant research suggests that parents distinguish between three kinds of intelligence<indepcl><C>: <D><appos\_NP\_list>knowledge about things, abstract problem-solving ability and social intelligence<appos\_NP\_list><D>. <s5241, academic prose>
- (12) <zz><pp>In the following pages<pp><zz> <C><fragment\_comp\_list>we explain<fragment\_comp\_list><C>:  
 - <Da><coord\_a\_emb\_asyn\_list>what ULDs are<coord\_a\_emb\_asyn\_list><Da>;  
 - <Db><coord\_b\_phr\_asyn\_list>their symptoms<coord\_b\_phr\_asyn\_list><Db>;  
 - <Dc><coord\_c\_emb\_asyn\_list>how you can avoid them<coord\_c\_emb\_asyn\_list><Dc>; <Dd><coord\_d\_emb\_list>and  
 - what you can do to help<coord\_d\_emb\_list><Dd>. <s7551, leaflets>

- (13) <zz><conj>Daarbij<conj><zz> <C><indepcl>zijn de auteurs er niet in geslaagd ondubbelzinnig te bepalen wat het centrale onderwerp van het rapport moest zijn<indepcl><C>: <D><appos\_NP\_contrast>het Bosnische drama of de Nederlandse rol daarin<appos\_NP\_contrast><D>. <s3848, academic prose>

(<zz><conj>In addition to that<conj><zz> <C><indepcl>the authors have not succeeded in determining unambiguously what should have been the main subject of the report <indepcl><C>: <D><appos\_NP\_contrast>the Bosnian drama or the Dutch part therein <appos\_NP\_contrast><D>.)

- (14) <A><advcl\_cond>Als het paard zou zien wat hem te wachten stond<advcl\_cond><A>, <C><indepcl>zou het nog maar een ding willen<indepcl><C>: <D><fragment\_VP>vluchten<fragment\_VP><D>. <s15913, short stories>

(<A><advcl\_cond>If the horse were to see what was in store for him<advcl\_cond><A>, <C><indepcl>he would want only one thing<indepcl><C>: <D><fragment\_VP>bolt<fragment\_VP><D>.)

- (15) <A/zz><disj>Natuurlijk<disj><A/zz> <C><indepcl>heeft De Nederlandsche Bank gelijk<indepcl><C>: <D><indepcl>het aandeel valse biljetten in de totale geldomloop is gering<indepcl><D>. <s2014, newspaper articles>

(<A/zz><disj>Of course<disj><A/zz> <C><indepcl>The Dutch Bank is right<indepcl><C>: <D><indepcl>the share of forged notes in the total circulation of money is limited<indepcl><D>.)

**Table 4 Grammatical realisation of D in (X)C:D(X)**

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Independent clause	10 (38.5%)	8 (50.0%)	7 (23.3%)	5 (9.4%)	30 (24.0%)
Apposition	3 (11.5%)	3 (18.8%)	9 (30.0%)	3 (5.7%)	18 (14.4%)
Fragment	3 (11.5%)	1 (6.2%)	6 (20.0%)	2 (3.8%)	12 (9.6%)
List (apposition, coordination, fragments)	10 (38.5%)	4 (25.0%)	8 (26.7%)	43 (81.1%)	65 (52.0%)
Total	26 (100%)	16 (100%)	30 (100%)	53 (100%)	125 (100%)
<b>Dutch</b>					
Independent clause	32 (47.8%)	20 (52.6%)	18 (42.9%)	28 (30.4%)	98 (41.0%)
Apposition	22 (32.8%)	12 (31.6%)	8 (19.0%)	12 (13.0%)	54 (22.6%)
Fragment	2 (3.0%)	4 (10.5%)	10 (23.8%)	5 (5.4%)	21 (8.8%)
List (apposition, coordination, fragments)	11 (16.4%)	2 (5.3%)	6 (14.3%)	47 (51.1%)	66 (27.6%)
Total	67 (100%)	38 (100%)	42 (100%)	92 (100%)	239 (100%)

The loglinear analysis showed no significant three-way interaction between language, genre and the various realisations of the D-satellite in the (X)C:D(X) pattern ( $\chi^2(9) = 9.10$ ,  $p = .428$ ). The analysis did show a significant two-way interaction between language and the realisation of the D satellite in this pattern ( $\chi^2(3) = 24.20$ ,  $p < .001$ , Cramer's  $V = .25$ ). Table 4 shows that the main differences between the languages can be found in the cases in which the D-satellite takes the form of a list. This occurs significantly more frequently in English than in Dutch across the different genres (English: 65 (52.0%); Dutch: 66 (27.6%)). The realisation category that is underrepresented in English, on the other hand, is the independent clause, which occurs more frequently in Dutch (100 (41.3%)) than in English (30 (24.0%)).

### Grammatical realisation of A and C in the A:C(X) pattern

As was shown in Table 2 above, the A:C(X) pattern has 15 occurrences in Dutch and no occurrences in English. This particular pattern shows overlap with (X)C:D(X) pattern in those cases in which the nucleus is realised as a fragment and the D-satellite is realised as an independent clause, a pattern that is also more common in Dutch than in English (see above). In the A:C(X) pattern, the A-satellite is realised as a conjunct in 8 of the 15 cases, with the nucleus taking the form of an independent clause in the academic prose and newspaper genre. In the leaflets genre this conjunct combines with a nucleus that takes the form of a list in the majority of cases. When the A-satellite does not take the form of a conjunct, it takes the form of an adverbial clause of condition (2) or comparison (2), and it takes the form of a phrase fragment in 3 cases. One of the main reasons that the elements that precede the colon have been classified as A-satellites and not as nuclei is thus mainly influenced by their grammatical form and function, i.e. conjuncts and adverbial clauses. Another reason is that the colon can often be replaced by a comma, as in all three examples below (16, 17 and 18), as opposed to, for instance, the colon in example (9) above.

Sentences (16) to (18) exemplify the use of the colon in the A:C(X) pattern in Dutch.

- (16) <A><conj>Met andere woorden<conj><A>: <C><indepcl>er moeten meer  
 asielzoekers alsnog een verblijfsvergunning krijgen<indepcl><C>. <s2945,  
 newspaper articles>

(<A><conj>In other words<conj><A>: <C><indepcl>more asylum seekers should receive a residence permit after all<indepcl><C>.)

- (17) <A><advcl\_compar>Zoals Johanna ter Meulen constateerde in haar onderzoek van 1903<advcl\_compar><A>: <C><indepcl>mensen met minder kinderen, <1\_prefv><coord\_b\_phr>dus een ruimere beurs<coord\_b\_phr><1\_prefv>, kunnen zich een grotere woning permitteren<indepcl><C>. <s5087, academic prose>

(<A><advcl\_compar>As Johanna ter Meulen noted in her study of 1903<advcl\_compar><A>: <C><indepcl>people with fewer children, <1\_prefv><coord\_b\_phr>and therefore more money <coord\_b\_phr><1\_prefv>, can afford a larger house<indepcl><C>.)

- (18) <A><conj>Immers<conj><A>: <C><fragment><correlative\_a>hoe<correlative\_a> kleiner het lichaam<fragment><C>, <D><fragment><correlative\_b>hoe<correlative\_b> hoger de alcoholconcentratie in het bloed na het drinken van alcohol<fragment><D>. <s8632, leaflets>

(<A><conj>After all<conj><A>: <C><fragment><correlative\_a>the<correlative\_a> smaller the body<fragment><C>, <D><fragment><correlative\_b>the<correlative\_b>higher the alcohol concentration in the blood after the drinking of alcohol<fragment><D>.)

### Grammatical realisation of D and E in the (X)CD:E(X) pattern

As Table 2 above showed, there are only a few instances in which the colon separates two appended discourse units, as the vast majority of colons separate the nucleus and the appended satellite that directly follows the nucleus. In English there are only 10 instances of the (X)CD:E(X) pattern, most of which occur in the academic prose genre (5) and the short stories genre (4), with the leaflets genre only showing one instance. Dutch follows the same pattern, with 6 of the 13 instances occurring in the academic prose genre and 7 in the short stories genre (cf. 5.6 & 5.7, which showed that the sentence pattern in which the nucleus is followed by two or more appended satellites is most frequent in the academic prose genre and short stories genre).

In both languages, no distinct patterns can be found in the grammatical realisation of the units that are separated by means of a colon. The D-satellite is realised as an adverbial clause of various kinds, a non-restrictive relative clause or a PP, and the E-satellite is realised as an apposition, and independent clause or a PP.



Sentences (19) to (28) present some examples of sentences with the (X)CD:E(X) pattern.

(19) <C><indepcl>The model would specify that these three phenomena persist when the addiction stage is reached<indepcl><C>, <D><nonrestr\_relcl>where the stronger core facets of addiction become apparent<nonrestr\_relcl><D>: <E><appos\_NP\_list>withdrawal symptoms, relapse and reinstatement, conflict, and behavioural salience<appos\_NP\_list><E>. <s5897, academic prose>

(20) <C><indepcl>He lost the custom of Tommy Miller<indepcl><C>, <D><nonrestr\_relcl>who took to drinking in Finnegan's way up towards the Heather Hills<nonrestr\_relcl><D>: <E><appos\_NP>the only place, <1><commcl\_fragment>he said<commcl\_fragment><1>, where he didn't get pestered by nutters about the bloody statue and his bloody wife<appos\_NP><E>. <s10486, short stories>

(21) <Ca><coord\_a>Vooral de laatste vraag is niet eenvoudig te beantwoorden<coord\_a><Ca>, <Cb><coord\_b>want er zijn verschillende trends<coord\_b><Cb>, <D><nonrestr\_relcl>die elkaar gedeeltelijk tegenwerken<nonrestr\_relcl><D>: <E><appos\_NP\_list>de gezinscyclus, de toenemende woningnood en de toenemende bestaanszekerheid<appos\_NP\_list><E>. <s5083, academic prose>

(<Ca><coord\_a>It is especially the last question that is not easy to answer<coord\_a><Ca>, <Cb><coord\_b>for there are different trends<coord\_b><Cb>, <D><nonrestr\_relcl>which partly oppose each other<nonrestr\_relcl><D>: <E><appos\_NP\_list>the family cycle, the increasing housing shortage and the increasing social security<appos\_NP\_list><E>.)

(22) <C><indepcl>De banderillero stond stil<indepcl><C> - <1><indepcl>tussen hem en de stier lag nu een meter of twintig<indepcl><1> - <D><indepcl>hij liet zijn armen zakken<indepcl><D>: <E><indepcl>de stier reageerde niet<indepcl><E>. <s15948, short stories>

(<C><indepcl>The banderillero stood still<indepcl><C> - <1><indepcl>between him and the bull was now around 20 meters<indepcl><1> - <D><indepcl>he lowered his arms<indepcl><D>: <E><indepcl>the bull did not respond<indepcl><E>.)

## 7.4 The use of semi-colons

Section 7.2 above already indicated that the main differences between English and Dutch with respect to the frequencies of the punctuation marks under analysis here were found for the colon and the dash and were not as pronounced for the semi-colon. Although the differences in frequencies between the languages for this punctuation mark do appear considerable (110 instances in English vs. 68 in Dutch), a loglinear analysis shows that there is no three-way interaction between language, genre and the occurrence of semi-colons at the predetermined alpha level of .01 ( $\chi^2(3) = 10.20$ ,  $p = .02$ ). Moreover, even though the analysis did show a significant two-way interaction between language and occurrence of semi-colons ( $\chi^2(1) = 14.33$ ,  $p < .001$ , Cramer's  $V = .03$ ), with English containing a slightly higher frequency of semi-colons, the effect size indicates that this involves a very small effect. It should therefore be borne in mind that the main frequency differences between English and Dutch can be found for the colon and the dash, and not for the semi-colon.

The present section will look into the use and occurrence of the semi-colon in more detail. It will first provide an overview of what type of discourse units the semi-colon typically links and will then look into the grammatical realisation of these units.

### 7.4.1 Position of semi-colon between discourse units

Whereas colons and dashes typically link a nucleus and a satellite (cf. 7.3 above and 7.5 below), the semi-colon can be used to link coordinated nuclei or a nucleus and an appended satellite. When considering how semi-colons are typically characterised, it becomes clear that Quirk et al., for instance, associate this punctuation mark especially with formal writing and regard it primarily as the 'coordinating mark of punctuation' (1985: 1622). However, they explain that in certain cases it 'shows affinity of use with the colon', mainly when it precedes a marker of apposition, such as *namely* or *that is* (p. 1623). Similarly, Nunberg et al. explain that, when used to link units *asyndetically*, this punctuation mark mainly occurs in formal writing (2002: 1740), and continue by stating that it allows both 'coordinative and elaborative interpretations' (2002: 1742). They distinguish between these interpretations by testing whether a coordinative conjunction can be inserted, such as *and* or *but*, which would give the relation between the units a coordinative interpretation. Onrust et al. (1993: 192), who describe the use of

commas, semi-colons and colons in Dutch, also explain that the semi-colon is typically used to link independent clauses, but at the same time state explicitly that this does not necessarily have to be the case (*ibid*). Instead of restricting its use to that of being primarily a coordinating device, they explain that the semi-colon can be used to link various types of units. For instance, it can link two units that are in contrast with each other; the second unit can repeat or explain what has been stated in the first unit, or it can provide an additional argument or reason for something that has been stated in the first unit (1993: 192-193).

Precisely because the semi-colon allows for both coordinative and elaborative interpretations, establishing the relationship between the units linked by a semi-colon was not always clear-cut in the annotation process. Although punctuation was not only used to identify discourse units, but also to determine their hierarchical status (cf. 2.5.1), with respect to units linked by semi-colons the other criteria that were established to identify units and determine their status, i.e. syntactic and semantic ones, were given more weight in the sentences linked by semi-colons. Specifically, to distinguish between *asyndetically coordinated units* on the one hand and a *nucleus-satellite relationship* on the other hand, the test of the possibility to insert a coordinative conjunction was applied. In those cases where this could easily be inserted, the units linked by a semi-colon were classified as *asyndetically coordinated nuclei*. In those cases where this could not be inserted, an additional semantic analysis was carried out to establish the relation between the two units, which often led to the assignment of a *nucleus-satellite relationship*.

Sentences (23) to (27) present some of the more complex cases, the analysis of which will be described and motivated for each of the examples.

- (23) <Ca><coord\_a\_asyn>Recognizable subject-matter seemed not to be necessary for the 3- to 4-year-olds<coord\_a\_asyn><Ca>;  
<Cb><coord\_b\_asyn>this finding is consistent with several other observations<coord\_b\_asyn><Cb>. <s5529, academic prose>

Sentence (23) presents an example in which it may be difficult to pinpoint the relation between the independent clauses and to determine the role of the semi-colon in signalling this relation. The question is whether the two clauses should be interpreted as constituting an instance of *asyndetic coordination*, which would mean that a coordinating conjunction can readily be inserted, or whether the relation should be classified as a *nuclear-satellite relation*, in which case the satellite would provide a further specification or reason for what is presented in

the nucleus. Even though this particular example does not easily allow for the insertion of a coordinating conjunction, classifying it as a nucleus-satellite relation, in which the semi-colon would have the function of a colon, also does not seem to yield the proper analysis. Because this constitutes a relation between two independent clauses of which the second cannot be interpreted as being in a satellite relation with the preceding one, it is analysed as an instance of asyndetic coordination of two independent clauses, between which the exact semantic relation is difficult to pinpoint. An elaborate analysis of such an example shows how syntactic, semantic and punctuational criteria interact in determining the discourse and rhetorical relation between various discourse units.

Sentence (24) provides an example of a situation in which the semi-colon has a similar function as a colon, as the relation the units more readily allows for an elaborative than a coordinative interpretation. For this reason, the first unit is analysed as having nuclear status and the second unit as having satellite status.

- (24) <C><indepcl>Nothing happened<indepcl><C>; <D><fragment>no reassuring click<fragment><D>. <s12445,>

Especially in the short stories genre in both languages the semi-colon often performs the function of linking two independent clauses that are asyndetically coordinated with each other, with the second coordinative providing an addition to what is stated in the first coordinative. In some cases this is a logical addition to what has preceded, but in other cases the relation between the units is not always clear. Sentences (25) and (26) are taken from the English and Dutch short stories genre respectively.

- (25) <Ca><coord\_a\_asyn>The eyes never focused on her<coord\_a\_asyn><Ca>; <Cb><coord\_b\_asyn>there was no-one there<coord\_b\_asyn><Cb>. <s12011, short stories>

- (26) <Ca><coord\_a\_asyn>Hij was wat ouder en zwaarder dan de jongens met de capes<coord\_a\_asyn><Ca>; <Cb><coord\_b\_asyn>in zijn hand hield hij een lange lans met een stalen punt<coord\_b\_asyn><Cb>. <s15910, short stories>

(<Ca><coord\_a\_asyn>He was a bit older and heavier than the boys with the capes<coord\_a\_asyn><Ca>; <Cb><coord\_b\_asyn>in his hand he held a long spear with a steel point<coord\_b\_asyn><Cb>.)

In addition to the cases in which it is more difficult to pinpoint the relation between the two units that are linked by a semi-colon, there are also situations in which this

is more clear-cut. For instance, in sentence (27) the semi-colon links two nuclear units that are in contrast with each other.

- (27) (<C>The civil city is always a work in progress<C>. <s4588, academic prose>  
 <Ca><coord\_a\_asyn>It is never assembled<coord\_a\_asyn><Ca>;  
 <Cb><coord\_b\_asyn>it is always being maintained, inspected,  
 improved<coord\_b\_asyn><Cb>. <s4589, academic prose>

An analysis of the sentences containing a semi-colon shows that they typically occur in five positions in a sentence, creating the following patterns: (X)Ca;Cb(X), (X)Ca,Cb;Cc(X)/(X)Ca;Cb,Cc(X), (X)C;D(X), (X)CDa;Db(X) and (X)CD;E(X). As was already explained in Section 7.2 above, all instances in which semi-colons separate the items on a list or are used in references in academic texts have been excluded from the analysis. The use of a single semi-colon in the (X)Ca,Cb;Cc(X) or (X)Ca;Cb,Cc(X) patterns are thus not analysed as separating the items in a list. Instead, they are taken to separate one coordinate from another. Although the distinction between lists and coordinated units is not always clear-cut (cf. 2.4.3), the sentences included in the analysis have all been manually checked and all constitute instances of coordination instead of lists. Sentences (28) to (33) exemplify each of these positions of the semi-colon and Table 5 below presents their frequencies.

- (28) <Ca><coord\_a\_asyn>Haiti has few natural resources<coord\_a\_asyn><Ca>;  
 <Cb><coord\_b\_asyn>its economy is mainly agricultural<coord\_b\_asyn><Cb>.  
 <s1229, newspaper articles>
- (29) <Ca>She called him Monsieur<Ca>, <Cb>and they addressed one another  
 decorously as vous<Cb>; <Cc>he found the tension between this linguistic  
 formality and the assumption of intimacy voluptuous<Cc>. <s10821, short  
 stories>
- (30) <Ca>Mark didn't know the boy's name<Ca>; <Cb>he was taller than the  
 others<Cb>, <Cc>and his white face looked as though it had been chiselled  
 out of marble that was sickening from some pollution in its veins<Cc>.  
 <s11300, short stories>
- (31) <C><indepcl>Maar de manier waarop Ahold en Shell de affaires behandelen,  
 stemt vooralsnog weinig hoopvol over de transparantie en het  
 zelfreinigend vermogen van het bedrijfsleven<indepcl><C>;  
 <D><appos\_NP>noodzakelijke voorwaarden om het door de schandalen  
 aangetaste vertrouwen te herstellen<appos\_NP><D>. <s2753, newspaper articles>

(<C><indepcl>But the way in which Ahold and Shell deal with the affairs, does not yet give much hope with respect to the transparency and self-cleaning capacity of businesses<indepcl><C>; <D><appos\_NP>necessary requirements for repairing the damage that has been caused by the scandals <appos\_NP><D>.)

(32) <C>But there is an important difference between Conservative means tests and Labour's income-related tax credits<C> - <Da>the first were doubled to cut public spending<Da>; <Db>the latter introduced to ensure a major increase in spending was targeted on those most in need<Db>. <s1298, newspaper articles>

(33) <C>Een van de meer kritische spectators, <1>De Denker<1>, nam al in mei 1763 een essay uit een Engelse periodiek in vertaling op<C>, <D>'Over de vryheid van denken en schryven over godsdienstige onderwerpen'<D>; <E>het werd in twee volgende stukken door De Denker zelf voortgezet<E>. <s3709, academic prose>

(<C>One of the more critical spectators, <1>The Thinker<1>, already included a translation of an essay from an English periodical in May 1763<C>, <D>'About the freedom of thinking and writing about religious topics'<D>; <E>it was continued by The Thinker himself in the two following pieces<E>.)

Sentence (29) exemplifies the (X)Ca,Cb;Cc(X) pattern and sentence (30) the (X)Ca;Cb,Cc(X) pattern. Note also that sentence (32) contains both a dash and a semi-colon and has therefore been included in both the analysis of the dash (see 7.5 below) and the present analysis, but has only been counted once in the overview Table 1 in Section 7.2. It should be noted that there are only two examples of sentences that contain both a dash and a semi-colon, of which sentence (32) is one.

**Table 5**                      **Frequencies of five main positions of semi-colons in sentences**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
(X)Ca;Cb(X)	12	8	32	5	57 (51.8%)
(X)Ca;Cb,Cc(X) or (X)Ca,Cb;Cc(X)	2	2	19	1	24 (21.8%)
(X)C;D(X)	0	1	13	4	18 (16.4%)
(X)CDa;Db(X)	0	2	2	1	5 (4.5%)
(X)CD;E(X)	2	0	4	0	6 (5.5%)
<b>Total</b>	<b>16</b>	<b>13</b>	<b>70</b>	<b>11</b>	<b>110 (100%)</b>
<b>Dutch</b>					
(X)Ca;Cb(X)	6	4	12	3	25 (36.8%)
(X)Ca;Cb,Cc(X) or (X)Ca,Cb;Cc(X)	0	2	2	1	5 (7.4%)
(X)C;D(X)	8	8	8	5	29 (42.6%)
(X)CDa;Db(X)	1	0	0	1	2 (2.9%)
(X)CD;E(X)	2	0	3	2	7 (10.3%)
<b>Total</b>	<b>17</b>	<b>14</b>	<b>25</b>	<b>12</b>	<b>68 (100%)</b>

Due to the detailed overview of positions of the semi-colon in the sentence, the expected frequencies of a number of cells are too low to be tested statistically. However, when grouping the patterns in which the semi-colon has a coordinative function, i.e. (X)Ca;Cb(X), (X)Ca;Cb, Cc(X)/(X)Ca, Cb; Cc(X), and (X)CDa;Db(X), and the patterns in which it has an elaborative function, i.e. (X)C;D(X) and (X)CD;E(X), and comparing the frequencies of the resulting two main patterns, the loglinear analysis showed a significant three-way interaction between language, genre and the occurrence of these two main patterns ( $\chi^2(3) = 11.70$ ,  $p < .01$ ). Subsequent chi-square tests showed a difference in the frequency of the two main uses of the semi-colon between English and Dutch for the academic prose genre ( $\chi^2(1) = 16.10$ ,  $p < .001$ , Cramer's  $V = .70$ ) and the short stories genre ( $\chi^2(1) = 27.52$ ,  $p < .001$ , Cramer's  $V = .54$ ). No difference in frequency between the languages for either of these two patterns was found for the newspaper genre ( $\chi^2(1) = 1.00$ ,  $p = .32$ , Cramer's  $V = .19$ ) and the leaflets genre ( $\chi^2(1) = .03$ ,  $p = .85$ , Cramer's  $V = .04$ ).

The English academic prose genre shows a higher frequency of the patterns in which the semi-colon has a coordinative interpretation (14 vs. 2)), whereas the Dutch academic prose genre shows a higher frequency of cases in which this punctuation mark has an elaborative interpretation (10 vs. 7).

The English newspaper genre only contains one instance in which the semi-colon has an elaborative interpretation, whereas Dutch counts 8 instances, but the difference between the languages is not significant.

In the short stories genre the main difference between the languages can be found for the significantly higher frequency of the use of the semi-colon in which it has an elaborative interpretation in Dutch (11 out of 25 (44.0%)) when compared to English (17 out of 70 (24.3%)).

The English leaflets genre shows a few more instances of the semi-colon in which it is used with a coordinative interpretation (7), when compared to the elaborative interpretation (4). In Dutch the situation is reversed, containing somewhat more instances of an elaborative interpretation (7) than a coordinative interpretation (5), but the difference between the languages is not significant.

In addition, Table 5 also shows that in English the most frequent pattern is the pattern in which the semi-colon separates two coordinated nuclei (57 (51.8%)), which is particularly frequent in the academic prose genre (12 out of 16) and the newspaper genre (8 out of 13). The second most frequent pattern in English is the pattern in which the semi-colon is used to separate one of three coordinates, either in the (X)Ca;Cb,Cc(X) pattern or the (X)Ca,Cb;Cc(X) pattern (24 (21.8%)), the vast majority of which occur in the short stories genre (19). In Dutch, on the other hand, the most frequent pattern is the one in which the semi-colon occurs between the nucleus and the D-satellite, the (X)C;D(X) pattern (29 (42.6%)), which is particularly frequent in the Dutch academic prose genre (8 out of 17), the newspaper genre (8 out of 14) and the leaflets genre (5 out of 12). The second most frequent pattern in Dutch is the (X)Ca;Cb(X) pattern, with 25 occurrences (36.9%), 12 of which occur in the short stories genre.

## 7.4.2 Grammatical realisation of discourse units surrounding semi-colons

### Grammatical realisation of coordinated nuclei in (X)Ca;Cb(X)

The nuclei in the (X)Ca;Cb(X) pattern can be realised syntactically either as independent clauses or as non-independent clauses, i.e. as subordinate clauses, phrases or fragments. Note that coordinated subordinate clauses or phrases have only been annotated when each of the coordinates is presented as a separate punctuation unit (cf. 2.4.3 and 3.3.1). In these cases the second coordinate consists of a subordinate clause, a phrase or a fragment that is coordinated with a subordinate clause, a phrase or a fragment of the first coordinate. Note furthermore that the independent clauses or non-independent clauses can either



be coordinated syndetically, i.e. by means of a coordinator, or asyndetically, i.e. without a coordinator present. In both English and Dutch the vast majority of coordinated nuclei that are linked by a means of semi-colon are realised as asyndetically coordinated independent clauses (English: 44 (77.2%); Dutch: 22 (88.0%)). A further analysis of these sentences shows that in both languages in the majority of cases the relation between the asyndetically coordinated clauses constitutes one of addition (English: 33 out of 44; Dutch: 17 out of 22), with the minority being in a relation of contrast (English: 11 out of 44; Dutch: 5 out of 22). The additive relations are particularly frequent in the short stories genres in both languages, of which sentences (25) and (26) above provide examples. Furthermore, English contains a few instances (6 out of 57) of syndetically coordinated independent clauses, with a few instances in each genre. In both languages, the vast majority of coordinated nuclei take the form of independent clauses, with English containing on 7 examples of coordination of non-independent clauses and Dutch only two. Sentences (34) through (37) present examples of the different realisation forms.

- (34) <Ca><coord\_a\_asyn>The Shuttle terminal at Cheriton slipped  
by<coord\_a\_asyn><Ca>; <Cb><coord\_b\_asyn>the train manager announced that  
they were approaching the Channel<coord\_b\_asyn><Cb>. <s10827, short stories>
- (35) <Ca><coord\_a\_asyn>Nederland kan geen grote stromen asielzoekers  
toelaten<coord\_a\_asyn><Ca>; <Cb><coord\_b\_asyn>daar is geen maatschappelijk  
draagvlak voor<coord\_b\_asyn><Cb>. <s2093, newspaper articles>
- (<Ca><coord\_a\_asyn>The Netherlands cannot allow large streams of asylum  
seekers<coord\_a\_asyn><Ca>; <Cb><coord\_b\_asyn>there is no public support for  
that<coord\_b\_asyn><Cb>.)
- (36) <A>At this point<A>, <Ca><coord\_a>momentum is all<coord\_a><Ca>;  
<Cb><coord\_b><coordinator>and<coordinator> Mr Kerry has it<coord\_b><Cb>. <s1111,  
newspaper articles>
- (37) <A/zz><pp>Van cannabis uit de coffeeshop<pp><A/zz>  
<Ca><coord\_a\_phr\_asyn>weet u nooit of het aan deze kwaliteitseisen  
voldoet<coord\_a\_phr\_asyn><Ca>; <Cb><coord\_b\_phr\_asyn>van de medicinale  
cannabis uit de apotheek wel<coord\_b\_phr\_asyn><Cb>. <s8789, leaflets>
- (<A/zz><pp>Of cannabis from the coffeeshop<pp><A/zz>  
<Ca><coord\_a\_phr\_asyn>you never know if it meets the quality  
standards<coord\_a\_phr\_asyn><Ca>; <Cb><coord\_b\_phr\_asyn>of the medicinal  
cannabis from the chemist's you do know<coord\_b\_phr\_asyn><Cb>.)

### Grammatical realisation of C and D in (X)C;D(X) pattern

Table 5 above showed that the (X)C;D(X) pattern has more occurrences in Dutch (29) than in English (18). With respect to the grammatical realisation of the nucleus in this pattern, this takes the form of an independent clause in all cases in both languages, with the exception of one sentence in English. The D-satellite, on the other hand, can be realised grammatically in various ways. In Dutch it is predominantly realised as an independent clause (20 of 29 cases), with the other Ds taking the form of an apposition (4), a phrasal fragment (3) or a subordinate clause (2). In English, 8 of the 18 instances take the form of an independent clause, with the others taking the form of an apposition (4), an appended clause (1), a phrasal fragment (4) and a non-finite clause (1). Despite the fact that English contains a small number of sentences with this particular pattern, the grammatical realisation of the D-satellite is quite diverse.

In both languages the largest realisation groups of the D-satellite are thus formed by independent clauses and appositions. By their very nature, appositions typically have the function of further specifying a piece of information given in the previous discourse unit. An analysis of the Ds that take the form of independent clauses shows that the majority of these also have a specifying function, or allow for an elaborative interpretation, in which they further explain or provide a reason for the information that is presented in the preceding nucleus. Examples (38) to (42) below provide examples of the different grammatical realisations of the D satellite.

(38) <C><indepcl>Don't assume that it will never happen to you<indepcl><C>;  
<D><indepcl>such complacency can put you at risk<indepcl><D>. <s7794, leaflets>

(39) <C><indepcl>There was always something<indepcl><C>; <D><appos\_list>taking  
the car to be serviced, or Melanie to the dentist, or the cat to the vet, or  
Teddy's suits to the dry-cleaner<appos\_list><D>. <s12770, short stories>

(40) <zz><adj>Vaak<adj><zz> <C><indepcl>gaat het dan om mensen die goed  
ingeburgerd lijken<indepcl><C>; <D><indepcl>hun kinderen gaan naar school  
en spreken vaak goed Nederlands<indepcl><D>. <s3081, newspaper articles>

(<zz><adj>Oftentimes<adj><zz> <C><indepcl>it is about people who appear to be  
culturally well assimilated<indepcl><C>; <D><indepcl>their children go to  
school and often speak good Dutch<indepcl><D>.)

- (41) <C><indepcl>Maar er moet wel hard worden opgetreden tegen de kwakzalvers<indepcl><C>; <D><pp>tegen de mensen die zich genezer noemen en patiënten valse hoop geven<pp><D>. <s2159, newspaper articles>
- (<C><indepcl>But firm action needs to be taken against quacks <indepcl><C>; <D><pp>against the people who call themselves healers and give patients false hope<pp><D>.)
- (42) <C><indepcl>Hij was een schat<indepcl><C>; <D><indepcl>hij deed zijn best alles waarvan Vincent zei dat het leuk was, leuk te vinden<indepcl><D>. <s14860, short stories>
- (<C><indepcl>He was a doll<indepcl><C>; <D><indepcl>he tried hard to like everything of which Vincent said was nice<indepcl><D>.)

### Use of semi-colons in other patterns

Table 5 above showed that the (X)Ca;Cb,Cc(X) pattern or the (X)Ca,Cb;Cc(X) pattern is particularly frequent in the English short stories genre (19 out of 24 total occurrences). It should be noted that these sentences constitute some of the more problematic sentences with respect to the distinction between coordinative and elaborative interpretations. With respect to their content, they often contain character descriptions or descriptions of situations, of which sentences (43) and (44) provide examples. In (43) the first two sentences are coordinated by means of the coordinating conjunction *and*, thus yielding a Ca-Cb analysis. The third sentence, linked to Cb by means of a semi-colon, is not interpreted as constituting a satellite of this Cb element. Rather, it constitutes a unit at the same hierarchical level, yielding a Ca,Cb;Cc analysis. In (44) an analysis of the sentences surrounding this particular sentence made clear that the second independent clause in this example could not be interpreted as constituting a further explanation or specification of the unit that precedes it. Because of this clear lack in semantic link, the sentences are considered to be coordinated, instead of being in a nucleus-satellite relationship.

- (43) <Ca><coord\_a>She called him Monsieur<coord\_a><Ca>, <Cb><coord\_b><coordinator>and<coordinator> they addressed one another decorously as vous<coord\_b><Cb>; <Cc><coord\_c\_asyn>he found the tension between this linguistic formality and the assumption of intimacy voluptuous<coord\_c\_asyn></Cc>. <s10821, short stories>

- (44) <Ca><coord\_a\_asyn>Mark didn't know the boy's name<coord\_a\_asyn><Ca>;  
<Cb><coord\_b\_asyn>he was taller than the others<coord\_b\_asyn><Cb>,  
<Cc><coord\_c><coordinator>and<coordinator> his white face looked as though it  
had been chiselled out of marble that was sickening from some pollution  
in its veins<coord\_c><Cc>. <s11300, short stories>

Furthermore, the (X)CD;E(X) pattern also has a few instances in both languages (English: 6; Dutch: 7), most of which occur in the short stories genre (English: 4; Dutch: 3). As the number of occurrences is too low to analyse them in any great detail, sentences (45) and (46) are provided as examples of this pattern.

- (45) <C><indepcl>Dolores had kept the car exactly as it was when he had  
it<indepcl><C>, <D><advcl>except the back seat was piled with her  
laundry<advcl><D>; <E><indepcl>she was always on her way to or from the  
laundrette<indepcl><E>. <s12438, short stories>

- (46) <A/zz><disj>Mogelijk<disj><A/zz><C><indepcl>wordt de stoornis bij meisjes  
minder snel herkend dan bij jongens<indepcl><C>, <D><advcl\_reas>omdat ze  
vaker ADD hebben<advcl\_reas><D>; <Ea><coord\_a>ze zijn minder agressief en  
hyperactief<coord\_a><Ea> <Eb><coord\_b><coordinator>en<coordinator> komen  
daardoor minder snel in de problemen<coord\_b><Eb>. <s8499, leaflets>

(<zz><disj>Possibly<disj><zz><C><indepcl>the disorder is less easily recognised  
with girls than with boys<indepcl><C>, <D><advcl\_reas>because they have ADD  
more often<advcl\_reas><D>; <Ea><coord\_a>they are less aggressive and  
hyperactive<coord\_a><Ea> <Eb><coord\_b><coordinator>and<coordinator> therefore  
do not get as quickly into problems<coord\_b><Eb>.)

Note that the decision to analyse sentence (46) as following an AC,D;EaEb pattern and not, for example, an AC,Da;DbDc pattern is mainly based on syntactic considerations. In Dutch a characteristic of subordinate clauses is that the finite verb occurs in clause-final position, as it does in the D-satellite in (46) ('hebben' – *have*) (cf. 3.3.1 & Haeseryn et al. 1997: 1095). The coordinated E-satellites do not follow this subordinate clause word order, but a main clause word order, in which subject and finite verb occur at the start of the clause. This syntactic difference clearly distinguishes the D and E satellites from each other.

## 7.5 The use of dashes

Section 7.2 above showed that the dash occurs significantly more frequently in all four genres in English than in Dutch. This section will look more closely at the use of the dash in both languages. It will examine what types of discourse units are typically linked by dashes and it will provide insight into the grammatical realisation of the units linked by dashes.

### 7.5.1 Position of dash between discourse units

An analysis of the use and occurrence of the dash shows that they are typically occur in two positions in the sentence, either between the nucleus and the first appended satellite, or between two appended satellites, creating the following two patterns: (X)C-D(X) and (X)CD-E(X). Sentences (47) and (48) below provide examples of these positions and Table 6 below presents their frequencies.

(47) <A>In particular<A>, <C>the great mass of the Catholic Irish could not be readily incorporated in an over-arching Britishness<C> - <D>at least not until progress was made in dismantling the penal laws<D>. <s4195, academic prose>

(48) <A>Hoewel ze al lang uit de mode waren<A> <C>liep Christine in oosterse gewaden door haar huis<C>, <D>waar permanent een broeikasklimaat heerste<D> - <E>goedkoper dan een reis naar India, was haar credo<E>. <s14099, short stories>

(<A>Even though they had been out of fashion for a long time<A> <C>Christine was walking around in oriental robes in her house <C>, <D>which had a constant greenhouse climate <D> - <E>cheaper than a trip to India was here credo<E>.)

**Table 6**                    **Frequencies of two main positions of dashes in sentence patterns**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
(X)C-D(X)	20	64	44	41	169 (94.4%)
(X)CD-E(X)	2	3	5	0	10 (5.6%)
<b>Total</b>	22	67	49	41	179 <sup>45</sup> (100%)
<b>Dutch</b>					
(X)C-D(X)	6	3	9	3	21 (80.8%)
(X)CD-E(X)	1	0	4	0	5 (19.2%)
<b>Total</b>	7	3	13	3	26 (100%)

Due to infrequency of the dash in Dutch, the expected frequencies of a number of cells are too low to test differences in frequencies statistically and will therefore only be described. Even though the numbers are much higher for English, in both languages the dash typically occurs between the nucleus and the D-satellite (English: 169 (94.4%); Dutch: 21 (80.8%)) and not between the appended satellites D and E (English: 10 (5.6%); Dutch: 5 (19.2%)). This applies to all four genres in both languages, with the Dutch short stories genre showing a few instances of the (X)CD-E(X) pattern.

Furthermore, a closer look at the most frequent pattern, the (X)C-D(X) pattern, shows that most of these sentences consist of just one appended satellite, the D, that is not followed by an E-satellite. Specifically, for English, only 18 (10.7%) of all sentences containing a dash between the nucleus and the D-satellite also contain an E-satellite; for Dutch this applies to 6 of the 21 cases, which occur in the academic prose genre (2) and the short stories genre (4).

---

<sup>45</sup> Note that the totals as presented in Table 6 deviate somewhat from the total number of dashes as presented in Table 1. This is because 4 sentences in English and 1 sentence in Dutch contain both a colon and a dash. In the totals overview as presented in Table 1 these sentences have been counted as instances of sentences containing colons and not as sentences containing dashes. In the present analysis of the dash they have, however, been included in the total number of dashes.

## 7.5.2 Grammatical realisation of discourse units surrounding dashes

### Grammatical realisation of D in (X)C-D(X) pattern

In taking a closer look at the grammatical realisation of the discourse units surrounding the dash, the focus will be on the grammatical realisation of the appended satellites following the dash, as the nucleus is typically realised as an independent clause in both languages in the vast majority of cases across the different genres (English: 159 (94.0%); Dutch: 15 (71.4%)). In English the few occurrences of nuclei realised as fragments occur in all genres, whereas in Dutch they occur predominantly in the short stories genre. Sentences (49) and (50) below provide examples of nuclei realised as fragments, taken from the English leaflets genre and the Dutch short stories genre respectively.

(49) <C><fragment\_Adj>Dead</fragment\_Adj><C> - <D><indepcl>what does it mean</indepcl><D>? <s4517, leaflets>

(50) <C><fragment\_VP>Ademen</fragment\_VP><C> - <D><fragment\_Adj>diep</fragment\_Adj><D>, <E><adj>traag</adj><E>. <s16378, short stories>

(<C><fragment\_VP>Breathing</fragment\_VP><C> - <D><fragment\_Adj>deeply</fragment\_Adj><D>, <E><adj>slowly</adj><E>.)

With respect to the grammatical form of the D-satellite in the (X)C-D(X) pattern, this can be realised as a phrase or a clause. Table 7 presents the frequencies of both realisation groups.

**Table 7** Grammatical realisation of D in (X)C-D(X) pattern as phrase or clause

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
D in (X)C-D(X) as phrase	12	26	22	10	70 (41.4%)
D in (X)C-D(X) as clause	8	38	22	31	99 (58.6%)
Total	20	64	44	41	169 (100%)
<b>Dutch</b>					
D in (X)C-D(X) as phrase	1	1	5	0	7 (33.3%)
D in (X)C-D(X) as clause	5	2	4	3	14 (66.7%)
Total	6	3	9	3	21 (100%)

Due to infrequency of the dash in Dutch, the expected frequencies of a number of cells are too low to be tested statistically and will therefore only be described. Table 7 shows that the D-satellite in the (X)C-D(X) pattern can take the form of a phrase or a clause. English shows a fairly even distribution of phrases and clauses, but the group of clausal D-satellites is larger than that of phrases (99 (58.6%)) when looking at the overall frequencies irrespective of genre. When looking at the differences between the genres, phrases are only more frequent in the academic prose genre (12 out of 20), with the newspaper and leaflets genres containing more instances of D-satellites realised as clauses. The short stories genre contains as many Ds that are realised as clauses as Ds that are realised as phrases. In Dutch most D-satellites also take the form of a clause (14 out of 21), which applies to all genres, except for the short stories genre. In this genre 5 of the 9 D-satellites are realised as phrases.

When realised as a phrase, the D-satellite can take the form of 1) an apposition, 2) a PP, 3) a phrasal fragment or 4) a phrasal fragment that is introduced by a coordinator<sup>46</sup>. Whereas the first three of these categories speak for themselves, the last one may need some clarification. Both Chapter 2 and Chapter 3 described how discourse units were defined and identified in the present study, how the hierarchical status of a discourse unit was determined and how the grammatical categorisation was carried out. Section 3.4.1 described the gradient between coordination and subordination and presented how these notions have been distinguished from each other and defined in the present study. It introduced the notions of appended coordination (cf. Quirk et al. 1985: 975) and supplementation (cf. Huddleston & Pullum 2002: 1350) to refer to those situations in which two or more units are linked to each other, often by means of a coordinating conjunction, with the last unit having the status of an afterthought. Precisely because the last unit can be interpreted as having afterthought status, it is analysed as an appended satellite instead of a coordinated nucleus at the level of discourse in the present study (cf. 3.4.1 for a more elaborate discussion on this issue). The particular type of punctuation mark, the dash in this case, is considered to indicate or even reinforce

---

<sup>46</sup> It should be noted that a number of D-satellites that are introduced by coordinating conjunctions were initially annotated as constituting instances of appended clauses. The category of appended clauses has, however, been described as constituting a container category in Chapter 3 (3.4.1), and a closer analysis of a number of sentences containing dashes and coordinating conjunctions has thus lead to a new categorisation.



this satellite status (cf. Nunberg et al. 2002: 1750, who state that dashes are not used to separate coordinated units). The reason for classifying phrases that are introduced by a coordinator as a separate realisation group of the D-satellite in the (X)C-D(X) pattern in this section is to provide insight into the use and frequency of constructions that are on the gradient between coordination and subordination, i.e. units that could be considered to be in a paratactic relation when looked at from a grammatical point of view and in a hypotactic relation when looked at from a discourse perspective. Sentences (51) through (54) provide examples of each of these realisation groups and Table 8 below presents their frequencies.

- (51) <C><indepcl>She also secured us our first ever gig<indepcl><C> -  
<D><appos\_NP>an all-woman's disco at Manchester University<appos\_NP><D>.  
<s11172, short stories>
- (52) <C><indepcl>Probabilistic inference can only be used effectively if it is  
possible to separate knowledge into discrete chunks<indepcl><C> -  
<D><pp>with a relatively sparse network of probabilistic dependencies  
between the chunks<pp><D>. <s5725, academic prose>
- (53) <C><fragment\_NP>Dood hout<fragment\_NP><C> -  
<D><fragment\_VP>knip<fragment\_VP><D>, <E><fragment>schaar erin en weg  
ermee<fragment><E>. <s16308, short stories>
- ( <C><fragment\_NP>Dead wood<fragment\_NP><C> -  
<D><fragment\_VP>cut<fragment\_VP><D>, <E><fragment>use your scissors and get  
rid of it<fragment><E>.)
- (54) <A>Besides<A>, <C><indepcl>leaving young people with a massive debt as  
they begin their working life won't be good for them<indepcl><C> -  
<D><fragment\_NP><coordinator>or<coordinator> the economy<fragment\_NP><D>.  
<s157, newspaper articles>

**Table 8** Grammatical realisation of phrasal D in (X)C-D(X) in 4 realisation groups

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Apposition	6	9	18	8	41 (58.6%)
PP	4	11	2	2	19 (27.1%)
Phrasal fragment	0	0	2	0	2 (2.9%)
Phrasal fragment + coordinator	2	6	0	0	8 (11.4%)
Total	12	26	22	10	70 (100%)
<b>Dutch</b>					
Apposition	1	0	1	0	2 (28.6%)
PP	0	1	0	0	1 (14.3%)
Phrasal fragment	0	0	4	0	4 (57.1%)
Phrasal fragment + coordinator	0	0	0	0	0 (0.0%)
Total	1	1	5	0	7 (100%)

Table 8 shows that the vast majority of phrases in English are realised as appositions (41 (58.6%)) and that the second largest group is formed by PPs (19 (25.7%)). It should be noted that the distinction between PPs on the one hand and non-nominal appositions on the other hand is not always clear-cut. This means that in certain cases the PPs could also have been classified as instances of non-nominal appositions (cf. 3.4.3 for a discussion on the gradient between these categories and for a definition of appositions in the present study). The group of fragments introduced by a coordinator mainly occurs in the newspaper genre, which contains 6 of the total number of 8 occurrences in English. As Dutch contains only 7 instances of phrasal Ds it is hard to find any patterns in grammatical realisation. Most of these phrases (4) are realised as phrasal fragments and occur in the short stories genre.

When realised as a clause, it can take the form of 1) an adverbial clause, 2) an appended clause, 3) a non-restrictive relative clause, 4) two or more coordinated clauses, 5) an independent clause, and 6) an independent clause that is preceded by a coordinator. This last category is similar to the phrasal fragments category introduced by a coordinator as described above, in that the independent clauses introduced by a coordinator can again be seen as being in a paratactic relation with the unit they precede when looking at their grammatical realisation and a hypotactic relation when looking at their discourse status. Sentences (55) to (61) provide examples of the different grammatical realisation groups and Table 9 below presents their frequencies.

- (55) <C><indepcl>Studying mental incapacity in the past may, <1>after all<1>, tell us something about pathologies<indepcl><C> - <D><advcl\_conces>even if that is not this historian's interest here<advcl\_conces><D>. <s4642, academic prose>
- (56) <C><indepcl>The reality is that a new, expensive and dangerous arms and technology race is gathering pace<indepcl><C> - <D><appcl>hardly a good non-proliferation example to set<appcl><D>. <s1355, newspaper articles>
- (57) <C><g\_indepcl>His eyes are firmly fixed on cutting taxes<indepcl><C> - <D><nonrestr\_relcl>which, <1>naturally<1>, will mainly benefit the rich<nonrestr\_relcl><D>. <s287, newspaper articles>
- (58) <C><indepcl>Give your son or daughter the chance to express an opinion<indepcl><C> - <D><indepcl>encourage this then discuss it<indepcl><D>. <s8173, leaflets>
- (59) <C><indepcl>We don't usually need to think very much about our sleep<indepcl><C> - <D><indepcl>it's just a part of life that we take for granted<indepcl><D>. <s7330, leaflets>
- (60) <A/zz><pp>Met een glas wijn in zijn hand<pp><A/zz> <Ca><coord\_a>stond hij tegen de deur geleund<coord\_a><Ca> <Cb><coord\_b>en sloeg de dansende gade<coord\_b><Cb> - <D><indepcl>er ging iets vreeswekkend eenzaam van hem uit<indepcl><D>. <s15295, short stories>
- ( <zz><pp>With a glass of wine in his hand<pp></zz> <Ca><coord\_a>he was leaning against the door<coord\_a><Ca> <Cb><coord\_b>and looking at the dancing crowd<coord\_b><Cb> - <D><indepcl>he appeared frighteningly lonely<indepcl><D>.)
- (61) <C><indepcl>Tony Blair insists that taxes must not go up to pay for higher education<indepcl><C> - <D><indepcl><coordinator>and<coordinator> a great many taxpayers agree with him<indepcl><D>. <s153, newspaper articles>

**Table 9** Grammatical realisation of clausal D in (X)C-D(X) in 6 realisation groups

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Adverbial clause	4	11	5	4	24 (24.3%)
Appended clause	0	6	0	1	7 (7.0%)
Non-restrictive relative clause	0	2	0	0	2 (2.0%)
Coordinated Da/b clauses	0	2	0	3	5 (5.0%)
Independent clause	2	9	13	22	46 (46.5%)
Independent clause + coordinator	2	8	4	1	15 (15.2%)
Total	8	38	22	31	99 (100%)
<b>Dutch</b>					
Adverbial clause	3	0	0	0	3 (21.4%)
Appended clause	1	1	0	0	2 (14.3%)
Non-restrictive relative clause	0	0	0	0	0 (0.0%)
Coordinated Da/b clauses	1	1	1	0	3 (21.4%)
Independent clause	0	0	3	3	6 (42.9%)
Independent clause + coordinator	0	0	0	0	0 (0.0%)
Total	5	2	4	3	14 (100%)

Even though the clausal D-satellite in the (X)C-D(X) pattern can take on a wide variety of grammatical forms, in both languages it is most frequently realised as an independent clause (English: 46+15 (61.6%); Dutch: 6 out of 14). This realisation group is particularly dominant in the English short stories genre (13+4 out of 22) and the English leaflets genre (22+1 out of 31), with the academic prose genre and newspaper genre showing some more variation in the grammatical realisation of D. The English newspaper genre is the genre that shows most variation in the grammatical form of D, but it should be noted that this is also the genre that shows the highest frequency of dashes when compared to the other genres (see Table 1 above). The particular type of independent clause that is preceded by a coordinator occurs predominantly in the English newspaper genre (8 of the 15 instances), similar to the phrase introduced by a coordinator (see Table 8 above).

As the use of the dash is rather infrequent in Dutch it is difficult to describe its main uses. In English, on the other hand, the use of the single dash can have various functions. One of these functions is similar to that of the colon, namely the specifying function (cf. Siepmann et al. 2008: 211), when the D-satellite can, for instance, take the form of an apposition or an independent clause that has a specifying function (see example (59) above). Another frequent use of the single dash is to introduce a comment that the writer makes, in which he reflects on the information contained in the unit that precedes the dash. Siepmann et al.

characterise the dash as '[a] mark of a writer who is very much present in the text, closely monitoring it, aware of the reader's needs, and ready to jump in with a comment, a correction, or a reformulation' (2008: 211). This particular use is illustrated by examples (55), (56), (57) and (60) above.

### Grammatical realisation of E in (X)CD-E(X) pattern

Table 6 above already showed that there are only a few instances in both languages in which the dash occurs between the D and E appended satellites. This occurs in 10 sentences in English and 5 sentences in Dutch. As this is a very infrequent use of the dash, this pattern will not be described in great detail, but will only be exemplified by the following sentences.

- (62) <A>If, <1>meanwhile<1>, there is another exodus of boatpeople<A>, <C><indepcl>the US has a plan<indepcl><C>: <D><indepcl>it is going to intern them<indepcl><D> - <E><pp>in Guantanamo Bay<pp><E>. <s1251, newspaper articles >
- (63) <C><indepcl>The Bath Philosophical Society (<1>1779<1>), <2>for instance<2>, <3>not untypically<3>, held that 'any subject within the circle of the arts and sciences, natural history, the history of nations, or any branch of polite Literature' was open for discussion<indepcl><C>, <D><advcl\_nonfin>with only law, medicine, theology and politics being excluded from meetings<advcl\_nonfin><D> - <E><appos\_NP>subjects that might inflame the passions and split the company<appos\_NP><E>. <s4735, academic prose>
- (64) <A/zz><pp>Als kind<pp><A/zz> <C><indepcl>had ik veel aan mijn lichaam lelijk gevonden<indepcl><C>: <D><appos\_NP\_list>aderen, vlekjes, nagelriemen, neus, huidcellen en haartjes<appos\_NP\_list><D> - <E><indepcl>er moest een tijd, misschien zelfs maar een moment zijn geweest waarop ik die afkeer kwijt was geraakt<indepcl><E>. <s16450, short stories>
- (zz><pp>As a child<pp><zz> <C><indepcl>I had disliked many things about my body<indepcl><C>: <D><appos\_NP\_list>veins, spots, cuticles, nose, skin cells and little hairs<appos\_NP\_list><D> - <E><indepcl>there needs to have been a time, perhaps only a moment at which I lost that dislike<indepcl><E>.)

Example (62) shows that the specifying function and 'commenting' function of the dash may not always be two clearly distinct functions, as the unit that follows the dash in this example could be argued to perform both functions at the same time.

Specifically, the sentence final PP could, on the one hand, be seen as a further specification of the information contained in the D-satellite that precedes it. On the other hand, the choice to present this information in two discourse units, of which the last is preceded by a dash, has the effect of putting much emphasis on the final unit, which could be interpreted as a means of the writer to draw attention to this particular piece of information. It is also noteworthy that this sentence is the last sentence of the whole text, a newspaper editorial, which reinforces the interpretation that the writer wants to make a final statement.

## 7.6 Comma splice

Although the focus in the present chapter has been on the use and occurrence of the colon, the semi-colon and the dash, there is a particular use of the comma that will also be looked at in more detail. This is when a comma is used to link two independent clauses. Because the comma is typically considered too weak a punctuation mark to link two independent clauses, when used in this way this is referred to as a *comma splice* and is usually advised against (cf. Quirk et al. 1985: 1615; Onrust et al. 1993: 190; Nunberg et al. 2002: 1742). The situations in which the comma can be used to link two independent clauses is, for instance, when the two clauses linked are short and parallel in structure (cf. Nunberg et al. 2002: 1742), or when the use of lexical markers, such as *some...others*, *on the one hand... on the other hand*, can contribute to making explicit the relation between the two sentences (cf. Onrust et al. 1993: 190).

For the present analysis of sentences that are classified as having a comma splice, a distinction is made between sentences that are parallel in structure or use a lexical marker, and sentences that do not contain these devices. It is the latter group that is seen as constituting the 'unacceptable' comma splices in the present study (see example (65) below). It should also be noted that the analysis of this use of the comma has been restricted to the academic prose genre, the newspaper genre and the leaflets genre. The reason that the short stories genre has been excluded from the analysis is that this genre contains considerable sections of simulated dialogue. Not only is the use of punctuation in dialogue different at times (cf. 2.5.2 on the segmentation of texts containing dialogues into discourse units using punctuational criteria), but as the comma is a very frequent punctuation

mark, restricting the analysis to the former three genres was mainly done for practical purposes.

**Table 10** Frequency of two types of comma splices in English and Dutch in three genres

English	Academic prose	Newspaper articles	Leaflets	Total
IC , IC	0	1	3	4 (40.0%)
IC , IC (parallel structure / lexical marker)	4	1	1	6 (60.0%)
Total	4	2	4	10 (100%)
<b>Dutch</b>				
IC , IC	5	2	7	14 (34.2%)
IC , IC (parallel structure / lexical marker)	9	12	6	27 (65.8%)
Total	20	16	13	41 (100%)

As the expected frequencies of a number of cells are too low to be tested statistically, differences in frequencies will only be described. Table 10 shows that the use of a comma to link independent clauses is rather infrequent in either language. English contains only 10 sentences in the three genres under consideration, which means that only 0.2% of the sentences included (4871) contain this use of the comma. For Dutch, this applies to 41 sentences, which is 0.7% of the sentences included in the analysis (5554). In both languages, the ‘unacceptable’ comma splices are particularly infrequent (English: 4; Dutch 14), with the majority of the sentences that are linked by a comma forming either a parallel construction or making use of a lexical marker to make the relation between the sentences explicit. The Dutch leaflets genre contains the highest frequency of ‘unacceptable’ comma splices (7 out of 13).

Sentences (65) to (69) will exemplify both ‘unacceptable’ comma splices and comma splices that could be considered acceptable. The first example (65) is generally considered unacceptable.

- (65) <A><adj>In a circular argument<adj><A>, <Ca><coord\_a\_asyn>Yannick Ripa claims that if women tried to escape from social conventions they would be labeled mad<coord\_a\_asyn><Ca>, <Cb><coord\_b\_asyn><conj>therefore<conj>madness is itself nothing more or less than an escape attempt<coord\_b\_asyn><Cb>. <s4635, academic prose>

In this particular example the relation between the two sentences is made explicit by the use of the word therefore. This is, however, generally regarded as unacceptable

(cf. Nunberg et al. 2002: 1742), because therefore is a conjunct that signals a meaning relation and not a grammatical relation (cf. Siepmann et al. 2008: 213). In English this only occurred twice, with both sentences occurring in the academic prose genre. Dutch contained no examples of this particular use of the comma in combination with a conjunct at the start of the second independent clause. Now consider example (66).

- (66) <Ca><coord\_a\_asyn>Some are embittered former ministers<coord\_a\_asyn><Ca>,  
<Cb><coord\_b\_asyn>others simply unreconstructed old-Labour  
losers<coord\_b\_asyn><Cb>. <s432, newspaper articles>

Sentence (66) constitutes an example of what is here considered an acceptable use of the comma to link the two independent clauses, as the clauses are parallel in structure and the contrast relation between them is made explicit by the words some... others. It should be noted that this only occurs 4 times in English and Section 7.4 above showed that English typically uses the semi-colon to link such sentences.

- (67) <C><indepcl>Beide pogingen mislukten<indepcl><C>, <D><indepcl>de NU  
bleef<indepcl><D>. <s3880, academic prose>

(<C><indepcl>Both attempts failed<indepcl><C>, <D><indepcl>the NU  
stayed<indepcl><D>.)

Sentence (67) presents an example of a case in which another punctuation mark, namely the colon, would have been more appropriate, as the second independent clause elaborates or further specifies what is stated in the first clause. In the present analysis this relation between the sentences is captured by the discourse labels nucleus and satellite respectively. This is an instance of an 'unacceptable' comma splice.

- (68) <Ca><coord\_a\_asyn>Te hopen is dat deze maatregelen voldoende  
zijn<coord\_a\_asyn><Ca>, <Cb><coord\_b\_asyn>anders zijn meer acties om  
ouderen aan het werk te krijgen nodig<coord\_b\_asyn><Cb>. <s2151, newspaper  
articles>

(<Ca><coord\_a\_asyn>It is hoped that these measures will be  
sufficient<coord\_a\_asyn><Ca>, <Cb><coord\_b\_asyn>otherwise more work needs  
to be put into getting elderly people to work<coord\_b\_asyn><Cb>.)



Sentence (68) presents an example of how the comma is typically used in Dutch when it links two independent clauses. In this example, the sentences linked are rather short and the lexical marker *anders* (otherwise) marks the contrast relation between the sentences. English typically uses a semi-colon in such cases.

(69) <du\_Ca><coord\_a\_asyn>Pijn kan variëren tussen licht tintelend gevoel of brandende pijn<coord\_a\_asyn><Ca>, <Cb><coord\_b\_asyn>soms is er sprake van krachtverlies<coord\_b\_asyn><Cb>. <s9814, leaflets>

(<du\_Ca><coord\_a\_asyn>Pain can vary from a light tingling feeling to a burning pain<coord\_a\_asyn><Ca>, <Cb><coord\_b\_asyn>sometimes there is loss of strength<coord\_b\_asyn><Cb>.)

Finally, sentence (69) presents an example of what is here considered an ‘unacceptable’ comma splice. In this case the sentences linked are not parallel in structure and there is no lexical marker that makes the relation between the two sentences explicit. Though rare, the Dutch leaflets genre contained most instances, i.e. 7 in total.

## 7.7 Conclusion

This chapter looked at the use and occurrence of three punctuation marks in more detail: the colon, the semi-colon and the dash. These were the marks that not only presented the main annotation difficulties in the discourse segmentation and grammatical categorisation process, but also appeared to be used to different effects in the two languages and the four genres within these languages.

An analysis of these three marks did indeed show that significant frequency differences can be found between the languages and genres and that each of the marks is associated with particular uses. A number of these uses provide insight into differences in style between the two languages, where discourse structure, grammatical realisation and punctuation interact in such a way to achieve a particular rhetorical effect.

First, an analysis of the use and occurrence of the colon showed that this punctuation mark has significantly higher frequencies in Dutch than in English in all four genres. A closer inspection of the position of this mark in the sentence showed that in both languages it mainly occurs between the nucleus and the first appended satellite, the D-satellite. In both languages the nucleus takes the form of an

independent clause in the vast majority of cases, although there is some variation between the different genres. Specifically, the leaflets genre shows a relatively high frequency of a particular type of fragment that introduces a list. In addition, Dutch shows a particular use of the colon in which the prepended satellite and the nucleus are separated by a colon and another particular use in which a nucleus that is realised as a phrase or fragment is separated from the following appended satellite, often realised as an independent clause, by means of a colon. Although by its very nature the colon already serves as a focusing device, the situation in which the nucleus takes the form of a phrasal fragment adds to this effect. Besides a difference in the grammatical form of the nucleus, more pronounced differences between the languages can be found in the grammatical realisation of the D-satellite that follows the colon. Whereas the colon in English is very frequently used to introduce a list, Dutch shows particularly high frequencies of cases in which the colon introduces an independent clause. This clause often provides a further specification or elaboration of the information contained in the nucleus. In other words, whereas the dominant use of the colon in English appears to be practical, in introducing lists, the dominant use in Dutch appears to be a focusing device, especially in those cases where the nucleus takes the form of a non-independent clause.

Second, an analysis of the use and occurrence of semi-colons showed that they are somewhat more frequent in English than in Dutch, although this does not constitute a large effect. In English the semi-colon is mainly used to separate two or three asyndetically coordinated nuclei, whereas in Dutch the most frequent pattern is formed by cases in which the semi-colon links a nucleus and an appended satellite. In English this latter use is mainly restricted to the short stories genre and shows some instances in the leaflets genre, although it should also be noted that especially the sentences in the short stories genre presented some difficulties with respect to determining the hierarchical status of the units linked by semi-colons. Furthermore, a closer analysis of the units linked by semi-colons showed that these are typically both realised as independent clauses, where the second clause is in an addition relation with the first clause. There are also examples of cases where the two clauses are in a contrast relation, but these are less frequent in both languages. Although the semi-colon is typically used to link two units of equal status asyndetically, because it can also be used to link a nucleus and a satellite, it was at times difficult to distinguish between a coordinative and an elaborative interpretation. Although punctuation hardly ever functions as the sole criterion in determining the hierarchical status of a unit, it was particularly in the case of semi-colons that syntactic and semantic criteria had to be used to determine this status.

Third, an analysis of the use and occurrence of dashes showed that this is significantly more frequent in English in all genres, and that it shows a particularly high frequency in the English newspaper genre. Despite the fact that Dutch shows a very low frequency of this punctuation mark, in both languages it predominantly occurs in between a nucleus and a D-satellite. This nucleus typically takes the form of an independent clause, with the D-satellite showing much more variation in its grammatical realisation. English shows a fairly even distribution across clausal and phrasal Ds, with the former showing somewhat higher frequencies. When realised as a phrase, this mainly takes the form of an apposition or PP, and when realised as a clause, the largest group is formed by independent clauses in both languages. English contains a particular use of the dash that appears to have a clear rhetorical function. This is presented by those cases in which the clause-final appended satellite, typically a D-satellite, is not only introduced by a dash, but also by a coordinator. It is precisely the use of both a coordinator on the one hand, which signals a paratactic relation, and the use of the dash on the other hand, which signals a hypotactic relation, which achieves a particular rhetorical effect. More specifically, it is the tension between the paratactic and hypotactic signals that creates this effect.

In addition to an analysis of these three punctuation marks, the chapter also presented the results of an analysis of the occurrence of comma splices in English and Dutch in the academic prose, newspaper and leaflets genre. This analysis showed that comma splices are rare in both languages, with Dutch showing somewhat higher frequencies, as it appears to use the comma in those situations where English would use a semi-colon. Furthermore, both languages contained very few instances of 'unacceptable' comma splices, i.e. sentences that are linked by a comma, but which are not parallel in structure or between which the relation is not made explicit by means of a lexical marker. The Dutch leaflets genre contained 7 sentences, which was the highest frequency.

In short, an analysis of the interaction between discourse status, grammatical realisation and punctuation marks shows that the latter, especially colons, semi-colons and dashes, are used differently in the two languages and the four genres within these languages to achieve different rhetorical effects. This confirms the assumption that a thorough sentencing analysis should include all these aspects of sentence composition and the interaction between them.

## 8. Interruptions

### 8.1 Introduction

In addition to the sentence patterns that can be created by various combinations of prepended and appended satellites and nuclei, another subset of sentence patterns can be identified if interpolated satellites are taken into account as well. Despite the fact that interpolated satellites show overlap with what in the literature are often referred to as parentheticals (cf. Burton-Roberts 2005 for an overview of various definitions and approaches), the present approach to these units directly follows from the discourse segmentation system developed for this study. This means that the discourse units that are here classified as interpolated satellites concern those units that interrupt or occur in between other discourse units, both nuclei and prepended or appended satellites (cf. 2.5.3) and can be identified by the fact that they are always presented as separate punctuation units, surrounded by two punctuation marks. In this study, all interruptions<sup>47</sup> have received a numerical label, starting with <1> for the first interruption in a sentence, continuing with <2> for the second interruption in a sentence, and so on. In addition to this numerical label, the position of the interruption with respect to the finite verb of the unit that it interrupts has also been indicated by adding the labels ‘prefinite’ or ‘postfinite’ to the numerical codes. Sentences (1) and (2) present examples of sentences that contain interruptions. The first sentence contains one interruption that follows the finite verb and the second sentence contains two interruptions, one of which precedes and one of which follows the finite verb.

- (1) <C><zz>In veel gevallen<zz> is een stevige bestraffing, <1\_postfv>dat wil zeggen celstraf<1\_postfv>, op zijn plaats<C>. <s2085, newspaper articles>
- (<C><zz>In many cases<zz> is a severe punishment, <1\_postfv>meaning confinement<1\_postfv>, in place<C>.)

---

<sup>47</sup> Note that the terms ‘interpolated satellites’ and ‘interruptions’ will be used interchangeably in this chapter.

(<C><zz>In many cases<zz> a severe punishment, <1\_postfv>meaning solitary confinement<1\_postfv>, is in place<C>.)

- (2) <C>The Tories, <1\_prefv>as they explained yesterday<1\_prefv>, will destroy all that by their tax cuts which, <2\_postfv>if they ever happened<2\_postfv>, would only significantly benefit the rich<C>. <s271, newspaper articles>

The fact that they create a range of subpatterns constitutes an important reason in itself for taking a closer look at their frequency and behaviour, as the focus of the present study as a whole is to identify the main sentencing patterns in English and Dutch. However, the main motivation for analysing them in more detail actually arose during the annotation process, as this led to the impression that they not only showed different frequencies in the two languages under consideration, but also that their use and occurrence was dependent on the genre within the language in which they occurred. Specifically, English appeared to show a higher frequency of interruptions, with the academic prose genre in particular showing high frequencies. The analysis of interpolated satellites as presented in this chapter will not only identify the main sentence patterns formed by interruptions, but also analyse their frequencies to determine whether the hunch that had arisen during the annotation process can be confirmed.

The chapter will first present the frequencies of all sentences that contain one interruption and all sentences that contain two interruptions. It will then provide an overview of the various subpatterns that can be created by interruptions in each of the four main sentence patterns, C, XC, CX and XCX, focussing on the most frequent subpatterns. The infrequent subpatterns will only be illustrated by means of examples from both languages.

## 8.2 Overall frequencies of sentences with one and two interruptions

To give an indication of the occurrence of interruptions, Table 1 provides the frequencies of all sentences in English and Dutch across the different genres that contain one interruption. These totals are based on a count of the interruptions that occur either in the nucleus or in the appended satellite. Interruptions that occur in the prepended satellites have been discussed in detail in Chapter 6 (see 6.3.1 & 6.4.1 on interpolated satellites in the A1C(X) and A1AC(X) subpatterns respectively). In

addition, another use of the interpolated satellite in the short stories genre has also been excluded from the present analysis. This concerns those interruptions that occur between the nucleus and the appended satellite and typically take the form of reporting clauses, interrupting reported speech. These special types of interruptions will be described separately in the sections that deal with the interruptions in the CX pattern (8.5) and the XCX pattern (8.6). Sentence (3) below contains an example of one such interruption.

- (3) <C>En dit is Philip<C>, <1>zegt Julia<1>, <D>u weet wel<D>, <E>de kunstenaar van de familie<E>. <s13935, short stories>

(<C>And this is Philip<C>, <1>says Julia<1>, <D>you know<D>, <E>the artist of the family<E>.)

**Table 1** Frequencies of all sentences with no interruptions, 1 interruption and 2 interruptions

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Sentences with 0 interruption	1131 (78.7%)	1602 (86.9%)	2971 (93.8%)	1439 (90.6%)	7143 (88.2%)
Sentences with 1 interruption	260 (18.1%)	227 (12.3%)	179 (5.6%)	135 (8.5%)	801 (10.0%)
Sentences with 2 interruptions	47 (3.3%)	15 (0.8%)	19 (0.6%)	15 (0.9%)	96 (1.2%)
Total	1438 (100%)	1844 (100%)	3169 (100%)	1589 (100%)	8040 (100%)
<b>Dutch</b>					
Sentences with 0 interruption	1473 (84.7%)	1586 (90.4%)	3026 (95.9%)	1933 (93.8%)	8018 (92.1%)
Sentences with 1 interruption	243 (14.0%)	160 (9.1%)	115 (3.6%)	115 (5.6%)	633 (7.3%)
Sentences with 2 interruptions	24 (1.4%)	8 (0.5%)	13 (0.4%)	12 (0.6%)	57 (0.7%)
Total	1740 (100%)	1754 (100%)	3154 (100%)	2060 (100%)	8708 (100%)

The loglinear analysis<sup>48</sup> showed no significant three-way interaction between language, genre and sentences with 0, 1 or 2 interruptions ( $\chi^2(6) = 3.06$ ,  $p = .80$ ). It did, however, show a significant two-way interaction between language and occurrence of interruptions ( $\chi^2(2) = 62.30$ ,  $p < .001$ , Cramer's  $V = .06$ ). The main difference between

<sup>48</sup> Similar to the procedure followed in the main results chapter, Chapter 5, all tables that satisfy the criteria for the loglinear analysis will be tested statistically and those that do not satisfy the criteria will only be described.

English and Dutch is that English contains significantly more sentences that have one interruption and significantly more sentences with two interruptions. Specifically, 801 (10%) of all English sentences contain one interruption, compared to 633 (7.3%) Dutch sentences, and 96 (1.2%) of all English sentences contain two interruptions, compared to 57 (0.7%) sentences in Dutch.

Moreover, the analysis also showed a significant two-way interaction between genre and occurrence of interruptions ( $\chi^2(6) = 436.65$ ,  $p < .001$ , Cramer's  $V = .12$ ) that is also noteworthy. Although differences between genres irrespective of language are not the main focus of this study, it is interesting to see that in both languages the frequency of interruptions is to a large extent dependent on genre. Specifically, when looking at the sentences with one interruption, in both languages these are most frequent in the academic prose genre (English: 260 (18.1%); Dutch: 243 (14.0%)), followed by the newspaper genre (English: 227 (12.3%); Dutch: 160 (9.1%)), followed by the leaflets genre (English: 135 (8.5%); Dutch: 115 (5.6%)), with the short stories genre containing the lowest frequency of interruptions (English: 179 (5.6%); Dutch: 115 (3.6%)). Furthermore, when looking at the sentences with two interruptions, it is again the academic prose genre that shows the highest frequency (English: 47 (3.3%); Dutch: 24 (1.4%)).

Table 2 provides the distribution of sentences with one interruption across the main sentence patterns, C, XC, CX and XCX. Note that for the sentences patterns that contain prepended satellites, the XC and XCX patterns, the focus will be on the interruptions that occur in the nucleus of the subpatterns in which the X-element consists of one element, the A/zz-satellite (AC subpattern). The patterns in which the X-element consists of two or more satellites will be described separately at the end of 8.4. This applies to only 21 sentences in total (14 English, 7 Dutch). The loglinear analysis showed no significant three-way interaction between language, genre and distribution of sentences with one interruption across the different sentence patterns ( $\chi^2(9) = 11.16$ ,  $p = .26$ ). Nor did it show a two-way interaction between language and distribution across the sentence patterns ( $\chi^2(3) = 7.05$ ,  $p = .07$ , Cramer's  $V = .07$ ). Table 2 shows that in both languages the vast majority of sentences with one interruption belong to the C-pattern across the different genres (English: 477 (59.6%); Dutch: 362 (57.2%)). In both languages the XCX pattern shows the lowest frequencies of sentences that contain one interruption (English: 39 (4.9%); Dutch: 42 (6.6%)).

**Table 2** Distribution of sentences with one interruption across C, AC, CX and ACX pattern

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
C-pattern: 1 interruption	139 (53.5%)	150 (66.1%)	107 (59.8%)	81 (60.0%)	477 (59.6%)
AC pattern: 1 interruption	52 (20.0%)	31 (13.7%)	22 (12.3%)	14 (10.4%)	119 (14.9%)
CX pattern: 1 interruption	52 (20.0%)	43 (18.9%)	40 (22.3%)	31 (23.0%)	166 (20.7%)
ACX pattern: 1 interruption	17 (6.5%)	3 (1.3%)	10 (5.6%)	9 (6.7%)	39 (4.9%)
Total	260 (100%)	227 (100%)	179 (100%)	135 (100%)	801 (100%)
<b>Dutch</b>					
C-pattern: 1 interruption	132 (54.3%)	111 (69.4%)	55 (47.8%)	64 (55.7%)	362 (57.2%)
AC pattern: 1 interruption	55 (22.6%)	24 (15.0%)	17 (14.8%)	24 (20.9%)	120 (19.0%)
CX pattern: 1 interruption	39 (16.0%)	20 (12.5%)	28 (24.3%)	22 (19.1%)	109 (17.2%)
ACX pattern: 1 interruption	17 (7.0%)	5 (3.1%)	15 (13.0%)	5 (4.3%)	42 (6.6%)
Total	243 (100%)	160 (100%)	115 (100%)	115 (100%)	633 (100%)

Table 3 gives the frequencies of sentences with two interruptions for each of the main sentence patterns.

**Table 3** Distribution of sentences with two interruptions across four main sentence patterns

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
C-pattern: two interruptions	21	11	13	6	51 (53.1%)
AC pattern: two interruptions	7	2	0	0	9 (9.4%)
CX pattern: two interruptions	12	2	3	6	23 (24.0%)
ACX pattern: two interruptions	7	0	3	3	13 (13.5%)
Total	47	15	19	15	96 (100%)
<b>Dutch</b>					
C-pattern: two interruptions	9	6	4	5	24 (42.1%)
AC pattern: two interruptions	4	2	1	4	11 (19.3%)
CX pattern: two interruptions	8	0	5	3	16 (28.1%)
ACX pattern: two interruptions	3	0	3	0	6 (10.5%)
Total	24	8	13	12	57 (100%)

Due to the rare occurrence of sentences with two interruptions, the expected frequencies of a number of cells are too low to compare differences in frequencies statistically. What Table 3 shows is that in both languages the vast majority of sentences with two interruptions belong to the C-pattern (English: 51 (53.1%); Dutch: 24 (42.1%)). This is followed by sentences with the CX pattern in both languages (English: 23 (24.0%); Dutch: 16 (28.1%)). In English the XC pattern



contains the lowest number of sentences with two interruptions (9 (9.4%)), whereas in Dutch this is the XCX pattern (6 (10.5%)).

Although the present focus is on presenting the absolute frequencies of interruptions, it is also interesting to look at their relative frequencies. These show that despite the fact that most sentences with an interruption belong to the C-pattern, it is in fact the XCX pattern that shows the highest relative frequency when looking at the frequency of interruptions with respect to the total number of sentences that belong to this sentence pattern. Specifically, 14.2% (52) of the sentences that occur in the English XCX pattern (366) and 10.6% (48) of the sentences that occur in the Dutch XCX pattern (453) contain one or more interruptions. When comparing this to the C-pattern, 10.8% (528) of the sentences that occur in the English C-pattern (4885) contain one or more interruptions and 8.1% (386) of the Dutch sentences (4759) that occur in this pattern. In the XC pattern, in which the initial X consists of one element, 10.0% (119) of the English sentences (1185) contain interruptions and 6.6% of the Dutch sentences (1935). Last, in the CX pattern, 13.4% of all the English sentences in this pattern (1410) contain interruptions and 8.8% of all Dutch sentences in this pattern (1419). The fact that sentences that contain more discourse units also show a higher relative frequency of interruptions is, however, in itself not surprising, as these thus automatically contain more possible positions for interruptions to occur.

### **8.3 Interruptions in the C-pattern**

Tables 2 and 3 above showed that most sentences with an interruption belong to the C-pattern. This is to be expected, as the vast majority of sentences belong to this main pattern. This section will present an overview of a number of subpatterns that are formed by interruptions in the C-pattern, containing one, two or even more interruptions. It will also look at the position of these interruptions with respect to the finite verb of the unit they interrupt and their grammatical realisation.

The C-pattern can contain different numbers and various combinations of interruptions. There are sentences that contain just one interruption (C-1-C); sentences with two interruptions (C-12-C), and sentences with three interruptions (C-123-C). In addition to these, there can also be sentences that contain two

interruptions, in which the second interruption is hierarchically dependent on the first interruption (C-1i-C) (cf. 2.5.3). Finally, another subpattern is formed by interruptions that occur between two coordinated nuclei (Ca – 1 – Cb). Sentences (4) to (8) provide examples of each of these subpatterns.

- (4) <C>Discriminatie, <1>op welke grond dan ook<1>, is verboden<C>. <s9106, leaflets>  
(<C>Discrimination, <1>on whatever ground<1>, is prohibited<C>.)
- (5) <C>The active sporting group, <1>however<1>, may have managed to engage in sports in spite of movement difficulties and, <2>by using effort to replace ability<2>, achieved good levels of participation<C>. <s5485, academic prose>
- (6) <Ca>The concept of cognitive salience (<1>the activity dominating a person's mental life<1>) and the definitions of euphoria (<2>the gaining of a 'buzz' or a 'high' from the activity<2>) and tolerance (<3>the need to engage in the activity to a progressively greater extent to acquire the same 'buzz'<3>) do not involve negative consequences for the individual, <Cb>and the occurrence of harmful consequences is central to the labelling of appetitive behaviours is excessive (Orford, 1985)<sup>49</sup><Cb>. <s5905, academic prose>
- (7) <C>It is also unhealthy because it is encouraging the subsidised output of a product that the World Health Organisation, <1>courageously - <i><1> - says we should be cutting back on<C>. <s1256, newspaper articles>
- (8) <Ca>Niet echt een groot wetenschappelijk talent<Ca>, <1>had Johannsen geantwoord<1>, <Cb>geen sterk onderzoekster<Cb>, <Cc>maar zeer serieus<Cc>. <s15096, short stories>  
(<Ca>Not really a great academic talent<Ca>, <1>Johannsen had answered<1>, <Cb>not a great researcher<Cb>, <Cc>but very serious<Cc>.)

Table 4 gives an overview of the frequencies of the various subpatterns that can be formed by interruptions in the C pattern.

---

<sup>49</sup> References, which occur mainly in the academic prose genre, have not been considered interpolated satellites in this study.

**Table 4**                      **Frequencies of subpatterns formed by interruptions in C-pattern**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
C – 1 – C	139 (84.8%)	150 (88.8%)	107 (80.5%)	81 (91.0%)	477 (85.9%)
C – 1-2- C	19 (11.6%)	10 (5.9%)	10 (7.5%)	6 (6.7%)	45 (8.1%)
C - 1-2-3 – C	2 (1.2%)	1 (0.6%)	3 (2.3%)	0 (0.0%)	6 (1.1%)
C - 1-1 – C	4 (2.4%)	8 (4.7%)	2 (1.5%)	2 (2.2%)	16 (2.9%)
Ca – 1 – Cb	0 (0.0%)	0 (0.0%)	11 (8.3%)	0 (0.0%)	11 (2.0%)
<b>Total</b>	<b>164 (100%)</b>	<b>169 (100%)</b>	<b>133 (100%)</b>	<b>89 (100%)</b>	<b>555 (100%)</b>
<b>Dutch</b>					
C – 1 – C	132 (91.0%)	111 (92.5%)	55 (85.9%)	64 (88.9%)	362 (90.3%)
C – 1-2- C	7 (4.8%)	6 (5.0%)	4 (6.2%)	4 (5.6%)	21 (5.2%)
C - 1-2-3 – C	2 (1.4%)	0 (0.0%)	0 (0.0%)	1 (1.4%)	3 (0.7%)
C - 1-1 – C	4 (2.8%)	3 (2.5%)	0 (0.0%)	3 (4.2%)	10 (2.5%)
Ca – 1 – Cb	0 (0.0%)	0 (0.0%)	5 (7.8%)	0 (0.0%)	5 (1.2%)
<b>Total</b>	<b>145 (100%)</b>	<b>120 (100%)</b>	<b>64 (100%)</b>	<b>72 (100%)</b>	<b>401 (100%)</b>

The pattern in which the C is interrupted by only one interpolated satellite is by far the most frequent sentence pattern in both languages across all genres (English: 477 (85.9%); Dutch: 362 (90.3%)). Of the remaining subpatterns, the pattern in which the C is interrupted by two interpolated satellites is the most frequent one (English: 45 (8.1%); Dutch: 21 (5.2%)). The remaining subpatterns show very low frequencies in both languages.

As the C1C and the C12C subpatterns are the most frequent ones in both languages, these will be described in more detail in the sections below.

### **C1C subpattern**

This section will first provide information about the position of the interruptions in the C1C subpattern with respect to the finite verb of the nucleus that they interrupt, which will be followed by an analysis of their grammatical realisation.

In those cases in which the C is realised as an independent clause that contains a finite verb, the interruption can occur either before or after the finite verb, indicated by the labels (1\_pre) or (1\_post) respectively. Sentences (9) and (10) provide examples of both positions of interruptions. Sentence (11) gives an example of a sentence that does not contain a finite verb, in which no specification with respect to the position of the interruption is added to the label. As these cases are very rare, they will not be considered any further.

- (9) <C>But Michael Mates, <1\_prefv>the Tory MP appointed to the inquiry<1\_prefv>, has refused to pull out<C>. <s591, newspaper articles>
- (10) <C>De meeste aandacht in de historiografie is - <1\_postfv>al dan niet terecht<1\_postfv> - uitgegaan naar de meer spectaculaire migratiestromen in de Vroegmoderne Tijd en de twintigste eeuw<C>. <s5114, academic prose>
- ( <C>Most attention in historiography has been- <1\_postfv>rightly so or not<1\_postfv> - paid to the more spectacular migration streams in the Early Modern Age and the twentieth century<C>.)
- (11) <C>Soft drinks (<1>not diet drinks<1>), sweets, jam and sugar, as well as foods such as cakes, puddings, biscuits, pastries and ice-cream<C>. <s6866, leaflets>

Table 5 provides the frequencies of the various positions interruptions can take with respect to the finite verb of the nucleus.

**Table 5**                      **Frequencies of different positions of interruptions with respect to finite verb in C1C pattern**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
C – 1_prefin – C	66 (47.5%)	90 (60.0%)	41 (38.3%)	37 (45.7%)	234 (49.1%)
C – 1_postfin – C	72 (51.8%)	60 (40.0%)	62 (57.9%)	43 (53.1%)	237 (49.7%)
C – 1_no fin verb – C	1 (0.7%)	0 (0.0%)	4 (3.7%)	1 (1.2%)	6 (1.3%)
<b>Total</b>	<b>139 (5.0%)</b>	<b>150 (100%)</b>	<b>107 (100%)</b>	<b>81 (100%)</b>	<b>477 (100%)</b>
<b>Dutch</b>					
C – 1_prefin – C	60 (45.5%)	51 (45.9%)	21 (38.2%)	31 (48.4%)	163 (45.0%)
C – 1_postfin – C	72 (54.5%)	58 (52.3%)	34 (61.8%)	31 (48.4%)	195 (53.9%)
C – 1_no fin verb – C	0 (0.0%)	2 (1.8%)	0 (0.0%)	2 (3.1%)	4 (1.1%)
<b>Total</b>	<b>132 (100%)</b>	<b>111 (100%)</b>	<b>55 (100%)</b>	<b>64 (100%)</b>	<b>362 (100%)</b>

When comparing the frequencies of interruptions occurring before or after the verb, excluding the sentences that contain no verb from the analysis, the loglinear analysis showed no significant three-way interaction ( $\chi^2(3) = 3.12, p = .37$ ), nor a significant two-way interaction between language and position of interruption ( $\chi^2(1) = 1.70, p = .19$ , Cramer's  $V = .04$ ). What Table 5 shows is that the frequencies of interruptions that occur before or after the finite verb are similar in Dutch and in English. In English 234 (49.1%) sentences contain an interruption that precedes the finite verb and 237 (49.7%) sentences one that follows the finite verb. In Dutch 163 (45.0%) interruptions precede the verb and 195 (53.9%) follow the finite verb.

### Grammatical realisation of 1\_prefin in C1C subpattern

The prefinite interruption can be grammatically realised as a phrase or a clause. Table 6 provides the frequencies of these two main realisation groups.

**Table 6** Grammatical realisation of 1\_prefin in C1C subpattern as phrase or clause

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
1_prefin – phrase	50 (75.8%)	58 (64.4%)	23 (56.1%)	28 (75.7%)	159 (67.9%)
1_prefin – clause	16 (24.2%)	32 (35.6%)	18 (43.9%)	9 (24.3%)	75 (32.1%)
Total	66 (100%)	90 (100%)	41 (100%)	37 (100%)	234 (100%)
Dutch					
1_prefin – phrase	39 (65.0%)	33 (64.7%)	10 (47.6%)	26 (83.9%)	108 (66.3%)
1_prefin – clause	21 (35.0%)	18 (35.3%)	11 (52.4%)	5 (16.1%)	55 (33.7%)
Total	60 (100%)	51 (100%)	21 (100%)	31 (100%)	163 (100%)

The loglinear analysis showed no significant three-way interaction between language, genre and grammatical form of the prefinite interruption ( $\chi^2(3) = 2.36$ ,  $p = .49$ ), nor did it show a significant two-way interaction between language and grammatical form ( $\chi^2(1) = .47$ ,  $p = .49$ , Cramer's  $V = .02$ ). Table 6 shows that the frequencies of interruptions that take the form of a clause or a phrase are similar for English and Dutch, with the majority taking the form of a phrase in both languages (English: 159 (67.9%); Dutch: 108 (66.3%)).

The phrases can be further subcategorised into seven realisation groups: 1) appositions, 2) adjunct/PPs, 3) conjuncts, 4) disjuncts, 5) premodifiers, 6) the second coordinate of coordinated phrases and 7) subjuncts or discourse markers, which each only contain one example. Sentences (12) to (17) provide examples of each type of phrase and Table 7 provides the frequencies of these different realisation groups.

(12) <C>The education passport - <1\_prefv><appos\_NP>the old voucher policy with a new name<appos\_NP><1\_prefv> - has also been trimmed<C>. <s1302, newspaper articles>

(13) <C>Het zal niemand verbazen dat dit, <1\_prefv><pp>ten tijde van de Koude Oorlog in Nederland<pp><1\_prefv>, tegen de klippen op werken was<C>. <s3874, academic prose>

<C>It won't surprise anyone that this, <1\_prefv><pp>during the Cold War in the Netherlands<pp><1\_prefv>, meant working overtime<C>.)

- (14) <C>This type of male violence, <1\_pfv><conj>however<conj><1\_pfv>, was primarily carried out behind closed doors<C>. <s4471, academic prose>
- (15) <C>This, <1\_pfv><disj>of course<disj><1\_pfv>, may change over time as a function of the child's age<C>. <s5276, academic prose>
- (16) <C>Onderhandelingen over (<1\_pfv><premod>het traject naar<premod><1\_pfv>) het lidmaatschap van de Europese Unie lopen al vele jaren<C>. <s3107, newspaper articles>
- (<C>Negotiations about (<1\_pfv><premod>the road to<premod><1\_pfv>) membership of the European Union have been going on for many years<C>.)
- (17) <C>The decision to release five of the nine - <1\_pfv><coord\_b\_phr>and the admission by Home Secretary David Blunkett that they pose no threat<coord\_b\_phr><1\_pfv> - is shameful recognition they were wrongly detained<C>. <s339, newspaper articles>
- (18) <Ca>Het was een benedenwoning met een smalle gang van namaakmarmen<Ca> <Cb>en, <1\_pfv><dm>allemensen<dm><1\_pfv>, wat een lucht hing er in de kamer die hij aan het eind daarvan ontsloot<Cb>! <s14936, short stories>
- (<Ca>It was a ground-floor flat with a narrow hallway with imitation marble<Ca> <Cb>and, <1\_pfv><dm>oh boy<dm><1\_pfv>, what a smell was there in the room that he opened on the end of it<Cb>!)

Note that sentence (17) provides an example of an interruption that, from a grammatical viewpoint, is in a paratactic relation with the unit it precedes, as it is the second coordinate of two coordinated noun phrases. However, from a discourse perspective, it could be argued that it is hypotactically related to the surrounding sentence (cf. 3.3.1 and 9.3.2 on *Interruptions*). A further example of this ‘mismatch’ between discourse and grammar is expressed in the singular form of the verb *is*. If the two units were paratactically related to each other at both the levels of grammar and discourse, the form of the verb would be plural instead of singular, i.e. *are*. The fact that it is singular underlines that the hypotactic status of the unit between paired dashes in this sentence. This ‘mismatch’ between discourse and grammar is captured by the respective labels in the annotation system.

**Table 7** Grammatical realisation of phrasal 1\_prefin in C1C subpattern in 7 realisation groups

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Apposition	19 (38.0%)	45 (77.6%)	11 (47.8%)	21 (75.0%)	96 (60.4%)
Adjunct/PP	13 (26.0%)	7 (12.1%)	8 (34.8%)	1 (3.6%)	29 (18.2%)
Conjunct	13 (26.0%)	0 (0.0%)	1 (4.3%)	2 (7.1%)	16 (10.1%)
Disjunct	1 (2.0%)	0 (0.0%)	3 (13.0%)	1 (3.6%)	5 (3.1%)
Premodifier	1 (2.0%)	0 (0.0%)	0 (0.0%)	1 (3.6%)	2 (1.3%)
Coordinate_b_phrase	3 (6.0%)	5 (8.6%)	0 (0.0%)	2 (7.1%)	10 (6.3%)
Subjunct/discourse marker	0 (0.0%)	1 (1.7%)	0 (0.0%)	0 (0.0%)	1 (0.6%)
Total	50 (100%)	58 (100%)	23 (100%)	28 (100%)	159 (100%)
<b>Dutch</b>					
Apposition	23 (59.0%)	28 (84.8%)	4 (40.0%)	20 (76.9%)	75 (69.4%)
Adjunct/PP	11 (28.2%)	0 (0.0%)	5 (50.0%)	2 (7.7%)	18 (16.7%)
Conjunct	1 (2.6%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.9%)
Disjunct	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Premodifier	3 (7.7%)	4 (12.1%)	0 (0.0%)	3 (11.5%)	10 (9.3%)
Coordinate_b_phrase	1 (2.6%)	1 (3.0%)	0 (0.0%)	1 (3.8%)	3 (2.8%)
Subjunct/discourse marker	0 (0.0%)	0 (0.0%)	1 (10.0%)	0 (0.0%)	1 (0.9%)
Total	39 (100%)	33 (100%)	10 (100%)	26 (100%)	108 (100%)

In both languages the vast majority of prefinite interruptions take the form of an apposition (English: 75 (60.4%); Dutch: 75 (69.4%)). This realisation group is particularly large in the newspaper genre (English: 45 (77.6%); Dutch: 28 (84.8%)) and the leaflets genre (English: 21 (75.0%); Dutch: 20 (76.9%)).

In the English academic prose genre the group of appositions is again the largest (19 (38.0%)), but other large groups are formed by adjuncts/PPs (13 (26.0%)) and conjuncts (13 (26.0%)). Examples of conjuncts are *however*, which forms the largest group, but also *for example/for instance*, *on the other hand* and *in turn*. In Dutch the group of appositions is also the largest (23 (59.0%)), followed by adjunct/PPs (28.2%). It also has a few instances of premodifiers (see sentence (16) above), which is a particular function of the interruption in Dutch.

In the English newspaper genre, besides the large group of appositions, a few interruptions take the form of adjunct/PPs (7 (12.1%)), whereas Dutch shows no examples of this group, but instead again contains a few instances of the premodifiers (4 (12.1%)).

In the English short stories genre, the largest realisation group is again formed by appositions (11 (47.8%)) and the second largest group by adjunct/PPs (8

(34.8%)), whereas in Dutch the situation is reversed, containing 5 adjunct/PPs (50.0%) and 4 appositions (40.0%).

In the English leaflets genre, the few interruptions that do not take the form of an apposition are distributed fairly evenly across the other realisation groups, and the same applies to Dutch, which again contains a few instances of the premodifiers (3 (11.5%)).

As Table 6 above showed, the prefinite interruption can also take the form of a clause. This clause can be further subcategorised into the following realisation groups: 1) finite adverbial clauses, 2) non-finite adverbial clauses, 3) non-restrictive relative clauses, 4) independent clauses, 5) reporting clauses and 6) comment clauses. Sentences (19) to (24) provide examples of each of these types and Table 8 will present their frequencies.

(19) <C>Leading parties and politicians, <1\_prefv><subcl\_advcl\_cond>even if they have not been disqualified<subcl\_advcl\_cond><1\_prefv>, are boycotting the polls<C>. <s1336, newspaper articles>

(20) <C>The use of oaths as bonds of allegiance, <1\_prefv><subcl\_advcl\_nonfin>binding the populace to serve their king<subcl\_advcl\_nonfin><1\_prefv>, dated from Anglo-Saxon times<C>. <s4094, academic prose>

(21) <C>De tabaksindustrie, <1\_prefv><nonrestr\_relcl>die lang de negatieve cijfers over roken ontkende<nonrestr\_relcl><1\_prefv>, heeft dat in 1997 toegegeven<C>. <s9662, leaflets>

(<C>The tobacco industry, <1\_prefv><nonrestr\_relcl>who long denied the negative figures about smoking<nonrestr\_relcl><1\_prefv>, admitted that in 1997<C>.)

(22) <Ca>De tweede relatie die voorkomt uit de hypothese, <1\_prefv><indepcl>hypervigilantie voor somatische sensaties en/of pijn leidt tot verhoogde pijnperceptie<indepcl><1\_prefv>, kreeg tot nu toe minder aandacht<Ca> <Cb>en werd voornamelijk onderzocht met zelfrapportagemethoden<Cb>. <s6467, academic prose>

(<Ca>The second relation that follows from the hypothesis, <1\_prefv><indepcl>hypervigilance for somatic sensations and/or pain leads to increased perception of pain<indepcl><1\_prefv>, received less attention up to now<Ca> <Cb>and was mainly being investigated with self-reporting methods<Cb>.)



- (23) <C>The IRA, <1\_prefv><reportingcl>he said<reportingcl><1\_prefv>, acted in good faith on both occasions<C>. <s511, newspaper articles>
- (24) <C>That, <1\_prefv><commcl\_fragment>they believe<commcl\_fragment><1\_prefv>, is the only way to get at the truth<C>. <s10, newspaper articles>

**Table 8 Grammatical realisation of clausal 1\_prefin in the C1C subpattern in 6 realisation groups**

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Adverbial clause finite	1	3	4	1	9 (12.0%)
Adverbial clause non-finite	5	12	7	6	30 (40.0%)
Non-restrictive relative clause	8	15	4	2	29 (38.7%)
Independent clause	1	0	0	0	1 (1.3%)
Reporting clause	1	1	1	0	3 (4.0%)
Comment clause	0	1	2	0	3 (4.0%)
Total	16	32	18	9	75 (100%)
<b>Dutch</b>					
Adverbial clause finite	3	1	3	0	7 (12.7%)
Adverbial clause non-finite	6	4	4	0	14 (25.6%)
Non-restrictive relative clause	8	12	3	4	27 (49.1%)
Independent clause	3	1	0	1	5 (9.0%)
Reporting clause	0	0	1	0	1 (1.8%)
Comment clause	1	0	0	0	1 (1.8%)
Total	21	18	11	5	55 (100%)

Table 6 above already showed that clausal interruptions occur much less frequently than phrasal interruptions. Table 8 shows that when it takes the form of a clause, this is most frequently realised as a non-restrictive relative clause or non-finite adverbial clause in both languages. The frequencies of the former group are higher in Dutch than English (Dutch: 27 (49.1%); English: 29 (38.7%)), whereas the frequencies of the latter group are higher for English than for Dutch (English: 30 (40.0%); Dutch: 14 (25.6%)). The occurrence of finite adverbial clauses is similar in both languages, with English containing 9 instances (12.0%) and Dutch 7 instances (12.7%).

The relatively high frequency of non-finite clauses in English is not surprising, as Chapter 5 has already indicated that their occurrence is significantly higher in English when compared to Dutch. What is rather surprising is the relatively high frequency of non-finite clauses in Dutch. In English 16 of the 30 non-finite clauses have a past participle and 11 have a present participle. In Dutch

almost all non-finite clauses contain a past participle (12 out of 14). Sentence (24) is an example of a Dutch non-finite clause, taken from the academic prose genre.<sup>50</sup>

(25) <C>De SEO, <1\_prefv><subcl\_advcl\_nonfin>gespecialiseerd in econometrisch onderzoek<subcl\_advcl\_nonfin><1\_prefv>, weet weinig van integratie<C>. <s3163, academic prose>

<C>The SEO, <1\_prefv><subcl\_advcl\_nonfin>specialised in econometric research<subcl\_advcl\_nonfin><1\_prefv>, knows little about integration<C>.)

As for the few finite adverbial clauses, in English these occurrences take the form of adverbial clauses introduced by *as* (3), condition (2), time (3) and concession (1). In Dutch they take the form of adverbial clauses of place (3) or time (2), concession (1) and one introduced by *zoals* (*as*).

### Grammatical realisation of 1\_postfin in C1C subpattern

The postfinite interruption in the C1C subpattern can also take the form of a phrase or a clause, the frequencies of which are presented in Table 9.

**Table 9** Grammatical realisation of 1\_postfin in C1C subpattern as phrase or clause

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
1_postfin – phrase	55 (76.4%)	43 (71.7%)	41 (66.1%)	37 (86.0%)	176 (74.3%)
1_postfin – clause	17 (23.6%)	17 (28.3%)	21 (33.9%)	6 (14.0%)	61 (25.7%)
Total	72 (100%)	60 (100%)	62 (100%)	43 (100%)	237 (100%)
Dutch					
1_postfin – phrase	56 (77.8%)	46 (79.3%)	18 (52.9%)	28 (90.3%)	148 (75.9%)
1_postfin – clause	16 (22.2%)	12 (20.7%)	16 (47.1%)	3 (9.7%)	47 (24.1%)
Total	72 (100%)	58 (100%)	34 (100%)	31 (100%)	195 (100%)

<sup>50</sup> Similar to the remark made in Footnote 6 of Chapter 3, it should be noted that the grammatical realisation of the interruption in example (25) can be classified as a non-finite clause, but actually not as an adverbial clause. It was explained in Footnote 6 of Chapter 3 that no distinction has been made between different types of non-finite clause in the present study (such as between participial clauses and non-finite adverbial clauses), mainly with the aim of achieving consistent annotation by keeping a limit to the number of grammatical categories. In other words, this decision was thus mainly driven by practical considerations.

The loglinear analysis showed no significant three-way interaction between language, genre and grammatical form of the postfinite interruption ( $\chi^2(3) = 2.87$ ,  $p = .41$ ), nor did it show a significant two-way interaction between language and grammatical form ( $\chi^2(1) = .02$ ,  $p = .88$ , Cramer's  $V = .02$ ). Table 9 shows that, similar to the grammatical form of the prefinite interruptions, the frequencies of postfinite interruptions that take the form of a clause or a phrase are similar for English and Dutch, with the vast majority taking the form of a phrase in both languages (English: 176 (74.3%); Dutch: 148 (75.9%)).

The phrases can again be further subcategorised into seven realisation groups, similar to the realisation groups for the prefinite interruptions.

**Table 10** Grammatical realisation of phrasal 1\_postfin in C1C subpattern in 7 realisation groups

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Apposition	17 (30.9%)	21 (48.8%)	16 (39.0%)	20 (54.1%)	74 (42.0%)
Adjunct/PP	19 (34.5%)	13 (30.2%)	13 (31.7%)	8 (21.6%)	53 (30.1%)
Conjunct	10 (18.2%)	3 (7.0%)	4 (9.8%)	5 (13.5%)	22 (12.5%)
Disjunct	3 (5.5%)	1 (2.3%)	4 (9.8%)	1 (2.7%)	9 (5.1%)
Premodifier	1 (1.8%)	0 (0.0%)	0 (0.0%)	1 (2.7%)	2 (1.1%)
Coordinate_b_phrase	5 (9.1%)	3 (7.0%)	1 (2.4%)	2 (5.4%)	11 (6.2%)
Subjunct/discourse marker	0 (0.0%)	2 (4.7%)	3 (7.3%)	0 (0.0%)	5 (2.8%)
Total	55 (100%)	43 (100%)	41 (100%)	37 (100%)	176 (100%)
<b>Dutch</b>					
Apposition	27 (48.2%)	29 (63.0%)	3 (16.7%)	10 (35.7%)	69 (46.6%)
Adjunct/PP	14 (25.0%)	8 (17.4%)	13 (72.2%)	6 (21.4%)	41 (27.7%)
Conjunct	1 (1.8%)	1 (2.2%)	1 (5.6%)	1 (3.6%)	4 (2.7%)
Disjunct	1 (1.8%)	1 (2.2%)	0 (0.0%)	0 (0.0%)	2 (1.4%)
Premodifier	11 (19.6%)	6 (13.0%)	0 (0.0%)	9 (32.1%)	26 (17.6%)
Coordinate_b_phrase	2 (3.6%)	1 (2.2%)	1 (5.6%)	2 (7.1%)	6 (4.1%)
Subjunct/discourse marker	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Total	56 (100%)	46 (100%)	18 (100%)	28 (100%)	148 (100%)

Similar to the prefinite interruption, the postfinite interruption takes the form of an apposition in both languages in the majority of cases (English: 74 (42.0%); Dutch: 69 (46.6%)), although it should be noted that the frequencies for appositions are higher with prefinite interruptions (see Table 7 above). With the postfinite interruptions, a second large group is formed by adjuncts/PPs (English: 53 (30.1%);

Dutch: 41 (27.7%)), and in English a third group by conjuncts (22 (12.5%)) and in Dutch by premodifiers (26 (17.6%)).

When comparing the English and Dutch academic prose genre, it is clear that the most obvious difference is the relatively high frequency of conjuncts in English (10 (18.2%)), compared to Dutch (1 (1.8%)). Appositions are somewhat more frequent in Dutch (27 (48.2%) vs. 17 (30.9%)), whereas adjuncts/PPs are somewhat more frequent in English (19 (34.5%) vs. 14 (25.0%)). The group of premodifiers again predominantly occurs in Dutch (11 (19.6%)) and hardly in English (1 (1.8%)).

In the newspaper genre the appositions form the largest group, particularly in Dutch (29 (63.0%) vs. 21 (48.8%)). The adjunct/PPs are somewhat more frequent in English (13 (30.2%)) and the premodifiers again in Dutch (6 (13.0%)).

Of the few phrasal interruptions that occur in the Dutch short stories genre, hardly any take the form of an apposition (3 (16.7%)) and most take the form of an adjunct/PP (13 (72.2%)). In English, the appositions form the largest group (16 (39.0%)), followed by adjunct/PPs (8 (21.6%)).

In the leaflets genre, the appositions again form the largest group, but this applies more to English (20 (54.1%)) than Dutch (10 (35.7%)), as the premodifiers also form a particularly large group in Dutch (9 (32.1%)).

The clausal postfinite interruptions can also be further subcategorised into six realisation groups, similar to the clausal prefinite interruptions. Table 11 provides the frequencies of the different types of clauses.

**Table 11** Grammatical realisation of clausal 1\_postfinite in C1C subpattern in 6 realisation groups

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Adverbial clause finite	3	4	6	3	16 (26.2%)
Adverbial clause non-finite	8	3	5	1	17 (29.8%)
Non-restrictive relative clause	4	2	2	0	8 (13.1%)
Independent clause	1	6	3	2	12 (19.6%)
Reporting clause	0	1	1	0	2 (3.3%)
Comment clause	1	1	4	0	6 (9.8%)
Total	17	17	21	6	61 (100%)
<b>Dutch</b>					
Adverbial clause finite	10	2	1	1	14 (29.8%)
Adverbial clause non-finite	2	3	3	2	10 (21.3%)
Non-restrictive relative clause	2	5	2	0	9 (19.1%)
Independent clause	0	1	3	0	4 (8.5%)
Reporting clause	0	1	6	0	7 (14.9%)
Comment clause	2	0	1	0	3 (6.4%)
Total	16	12	16	3	47 (100%)

The languages show similarities in the largest realisation groups, those of finite and non-finite adverbial clauses, and differences in the smaller realisation groups.

In the academic prose genre, the non-finite clauses form the largest group in English (10 of 17), followed by non-restrictive relative clauses (4) and finite adverbial clauses (3). In Dutch, the largest group is formed by finite adverbial clauses (10 of 16).

In the English newspaper genre, the largest group is formed by independent clauses (6), followed by finite adverbial clauses (4) and non-finite adverbial clauses (3). In Dutch the non-restrictive relative clauses form the largest group (5), followed by non-finite clauses (3).

In the English short stories genre, the finite adverbial clauses form the largest group (6), followed by non-finite adverbial clauses (5) and comment clauses (4). In Dutch the different categories all contain low frequencies, with the reporting clauses forming the largest group (6).

Finally, the leaflets genre in both languages contains hardly any postfinite interruptions realised as clauses. In English the few instances take the form of a finite adverbial clause (3) and an independent clause (2) and in Dutch they take the form of a non-finite clause in two of the three cases.

Sentences (26) and (27) contain an interruption realised as an independent clause and a non-restrictive relative clause, taken from the English and Dutch newspaper genre respectively.

(26) <C>Mr Duncan Smith was so incensed by the charges that his wife did not earn her secretarial pay (<1\_postfv><indepcl>they helped end his leadership<indepcl><1\_postfv>) that he made what people familiar with the case call "quite heavy-handed use of lawyers" to challenge critics<C>. <s1639, newspaper articles>

(27) <C>Het ziet er naar uit dat het kabinet voornemens is de missie van de Nederlandse militairen in Irak, <1\_postfv><nonrestr\_relcl>die in juli afloopt<nonrestr\_relcl><1\_postfv>, te verlengen<C>. <s2489, newspaper articles>

(<C>It looks like the government is intending the mission of the Dutch troops in Iraq, <1\_postfv><nonrestr\_relcl>which ends in July<nonrestr\_relcl><1\_postfv>, to extend<C>.)

When realised as a non-finite clause, most clausal interruptions in English contain a present participle (10 out of 17), and seven sentences a past participle. In Dutch, most of these sentences are verbless (5 out of 10) and three sentences have a past participle. Sentence (28) provides an example of a Dutch postfinite interruption that is realised as a verbless clause.

(28) <C>De opgedane kennis en ervaring zorgt, <1\_postfv><subcl\_advcl\_verbless>waar nodig<subcl\_advcl\_verbless><1\_postfv>, voor verbetering van de vaccins, het vaccinatieprogramma en de voorlichting over de vaccinaties<C>. <s10108, leaflets>

(<C>The acquired knowledge and experience leads, <1\_postfv><subcl\_advcl\_verbless>wherever necessary<subcl\_advcl\_verbless><1\_postfv>, to improvement of the vaccines, the vaccination programme and the information about vaccinations <C>.)

When realised as a finite adverbial clause, in English these take the form of an adverbial clause introduced by *as* in four cases (out of 16), an adverbial clause of condition in four cases, time in four cases and concession in three cases. In Dutch, there are four adverbial clauses of time and four introduced by *zoals* (*as*) (out of 14), the remaining clauses are realised as result, condition and place clauses.

### **The C12C subpattern**

Table 4 above already showed that the sentence pattern in which the nucleus contains two interruptions occurs much less frequently than the pattern in which it contains only one interruption. In fact, only 45 English sentences that belong to the C-pattern contain two interruptions, 19 of which occur in the academic prose genre, and only 21 Dutch sentences belong to this pattern. This section will briefly present the position and grammatical realisation of both interruptions.

#### **<1> in the C12C subpattern**

In English, of the 45 interruptions in this pattern, 20 occur before the finite verb and 23 after the finite verb. Two sentences contain no finite verb. In Dutch, seven of the 21 interruptions occur before the finite verb and 12 after the finite verb, with two sentences containing no finite verb.

In English, 16 of the 20 prefinite interruptions take the form of a phrase and four take the form of a clause. Most phrases are realised as appositions (11). Of the postfinite interruptions, 14 are realised as a phrase, mainly appositions and adjuncts/PPs, and nine as clauses of various types. In Dutch, of the seven prefinite interruptions, five are realised as appositions and two as non-finite clauses. Of the postfinite interruptions, seven are realised as appositions, three as adjuncts/PPs and two as finite subordinate clauses.

#### **<2> in the C12C subpattern**

In English, the majority of second interruptions in the C12C subpattern occur after the finite verb (35 out of 45), with only ten occurring before the finite verb. In Dutch, the group of interruptions that occur after the finite verb is also larger, 17 out of 21.

In English the few interruptions that occur before the finite verb are realised as appositions in six sentences and as clauses of various types in four sentences. The postfinite interruptions predominantly take the form of an apposition (10 sentences), an adjunct/PP (9) or a subordinate clause. In Dutch the few prefinite interruptions take the form of an apposition, which also applies to the postfinite interruptions, although these are also realised as adjuncts/PPs and subordinate clauses.

Sentences (29) and (30) provide examples of sentences with the C12C subpattern. The first sentence is taken from the English academic prose genre, which shows the

highest frequency of this pattern. Note that in English the distribution of this pattern across the two subgenres is similar, with nine sentences belonging to the psychology subgenre and ten to the history subgenre. The second sentence is taken from the Dutch newspaper genre. Although there are only few occurrences, the Dutch *Telegraaf* and *Volkscrant* each contain the same number of occurrences (3-3), whereas in English the Guardian contains more occurrences when compared to the Daily Mirror (9-1).

- (29) <C>The emergence of forms of mutual hostility,  
<1\_postfv><appos\_AdjP\_list>political, cultural, and  
religious<appos\_AdjP\_list><1\_postfv>, between England and Spain in the  
sixteenth century is not, <2\_postfv><disj>of course<disj><2\_postfv>, new  
topic<C>. <s4077, academic prose>
- (30) <C>De VVD is zo verstandig geweest de vrijheid van godsdienst  
(<1\_postfv><appos\_NP>het dragen van hoofddoekjes in de openbare  
ruimte<appos\_NP><1\_postfv>) en de vrijheid van onderwijs  
(<2\_postfv><appos\_NP>islamitische scholen<appos\_NP><2\_postfv>) ongemoeid  
te laten<C>. <s3055, newspaper articles>
- (<C>The Dutch Liberal Party has been sensible enough to leave the  
freedom of religion (<1\_postfv><appos\_NP>the wearing of a veil in a public  
place<appos\_NP><1\_postfv>) and the freedom of education  
(<2\_postfv><appos\_NP>Islamic schools<appos\_NP><2\_postfv>) undisturbed<C>.)

## 8.4 Interruptions in the XC pattern

Similar to the interruptions in the C-pattern, the interruptions in the XC pattern can also create various subpatterns. As was already explained above, we will focus on the interruptions that occur in the nucleus of the subpatterns in which the X-element consists of one element, the A/zz-satellite (AC subpattern), as there are only 21 sentences in total (14 English, 7 Dutch) that contain interruptions when the X-elements consists of two or more satellites. Note that the interruptions that occur in the X of the XC pattern were already described in Chapter 6.

In the AC subpattern, the nucleus can be interrupted by one interpolated satellite, creating the AC1C subpattern; it can be interrupted by two satellites, creating the AC12C subpattern, and it can be interrupted by two satellites that are hierarchically related to each other, creating the AC1iC subpattern. Sentences (31)



to (33) give examples of each of these subpatterns, with Table 12 below presenting their frequencies.

- (31) <A>Kortom<A>, <C>redenen genoeg voor alle betrokkenen (<1>NS, overheid en consumentenorganisaties<1>) om gezamenlijk te bezien of het treinkaartje wel extra duur gemaakt moet worden<C>. <s2007, newspaper articles>
- (<A>In short<A>, <C>reasons enough for all those involved (<1>National Rail, government and consumers' organisations<1>) to determine together whether the train ticket should be made extra expensive<C>.)
- (32) <A>In fact<A>, <Ca>Elizabeth's religious beliefs appear to have been, <1\_postfv>as Suzan Doran has recently observed<1\_postfv>, remarkably consistent<Ca> <Cb>and an emphasis on the centrality of the Bible together with a tendency (<2\_prefv>in contradistinction to many Protestants<2\_prefv>) to view the Mass as less than an absolute evil were characteristic throughout her life<Cb>. <s4233, academic prose>
- (33) <A>Second<A>, <C>the new man in charge of the Home Office's police standards unit, <1\_prefv>Paul Evans, <i>the former Boston police chief<i><1\_prefv>, will be there to stop British commentators overromanticising the FBI<C>. <s1196, newspaper articles>

**Table 12** Frequencies of subpatterns formed by interruptions in AC subpattern

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
AC – 1 – C	48	28	18	13	107 (89.9%)
AC – 1 2 – C	6	1	0	0	7 (5.9%)
AC – 1   – C	1	3	0	1	5 (4.2%)
Total	55	32	18	14	119 (100%)
<b>Dutch</b>					
AC – 1 – C	51	23	17	22	113 (88.9%)
AC – 1 2 – C	4	2	1	4	11 (8.7%)
AC – 1   – C	1	1	0	1	3 (2.4%)
Total	56	26	18	27	127 (100%)

The AC1C subpattern is by far the most frequent subpattern, as 107 (89.9%) of the English sentences belong to this pattern and 113 (88.9%) of the Dutch sentences. The AC12C subpattern only has a few instances. In English these predominantly occur in the academic prose genre (6) and one example in the newspaper genre, with the short stories and leaflets genre containing no examples. In Dutch, on the

other hand, occurrences are also low, but they do occur in the short stories genre (1) and leaflets genre (4). Finally, the occurrences for the AC1iC subpattern are also very low in both languages (English: 5 (4.2%); Dutch 3 (2.4%)). In English three of the five sentences occur in the newspaper genre and in Dutch each genre contains one example, except for the short stories genre.

As the AC1C subpattern is the most frequently occurring subpattern of the AC pattern that contains interruptions, this is the only pattern that will be described in more detail.

### AC1C subpattern

Similar to the interruptions that occur in the C-pattern, the interruptions in the AC subpattern can occur before or after the finite verb of the nucleus. Table 13 provides the frequencies of these two positions.

**Table 13**                      **Frequencies of different positions of interruptions with respect to finite verb in AC1C subpattern**

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
AC – 1_prefin – C	20	18	7	7	52 (48.6%)
AC – 1_postfin – C	28	10	11	6	55 (51.4%)
Total	48	28	18	13	107 (100%)
<b>Dutch</b>					
AC – 1_prefin – C	3	0	0	0	3 (2.6%)
AC – 1_postfin – C	48	23	17	22	110 (97.4%)
Total	51	23	17	22	113 (100%)

Table 13 shows that in English the distribution of interruptions that occur either before or after the finite verb is quite even (prefin: 52 (48.6%); postfin: 55 (51.4%)), whereas in Dutch there are practically no instances of interruptions that occur before the finite verb (3 (2.6%)), with almost all interruptions occurring after the finite verb (110 (97.4%)).

English shows some variation between the different genres, with the newspaper genre, containing more prefinite interruptions (18 out of 28) and the short stories genre showing the reverse pattern (11 postfin out of 18). In Dutch the few instances of prefinite interruptions occur in the academic prose genre (3). Sentence (34) contains an example of the rare occurrence of a prefinite interruption in Dutch and sentence (35) contains the more typical situation in which the interruption follows the finite verb. The rare occurrence of prefinite

interruptions in Dutch can be explained by the fact that Dutch is a verb second language, which means that the finite verb is typically placed in second position and no more than one element can occur in sentence-initial position (cf. Chapter 6, 6.1, see also Haeseryn et al. 1997: 1261; Smits 2002: 22). The effect of this verb-second principle becomes clear when translating an English sentence that belongs to the AC subpattern and contains a prefinite interruption in C into Dutch. First consider the original English sentence (34a), then the literal Dutch translation (34b), and then the grammatically appropriate Dutch translation (34c).

- (34a) <A>Last week<A>, <C>Conservative leader, <1\_pfv>Michael Howard<1\_pfv>, **went** to Burnley to denounce the BNP as "a bunch of thugs dressed up as a political party" <C>. <s1809, newspaper articles>
- (34b) <A>Vorige week<A>, <C>Conservatieven leider, <1\_pfv>Michael Howard<1\_pfv>, **ging** naar Burnley om te betichten de BNP van "een stelletje schurken zich voordoend als politieke partij" <C>. <s1809, newspaper articles>
- (34c) <A>Vorige week<A>, <C>**ging** de leider van de Conservatieven, <1\_postfv>Michael Howard<1\_postfv>, naar Burnley om de BNP te betichten van "een stelletje schurken dat zich voordoet als politieke partij" <C>. <s1809, newspaper articles>

Both in the literal (34b) and the final translation (34c), the finite verb (*ging*) is marked in bold to show how this shifts position and does indeed immediately follow the sentence-initial adjunct *vorige week* (*last week*) in the sentence that provides the grammatically correct translation (34c). The interruption in the original sentence in (34a) modifies the subject 'conservative leader', and does so as well in the correct translation (34c), changing the position of the interruption from occurring before the finite verb to occurring after the finite verb, a change that is also reflected in the labels. Furthermore, in all three Dutch sentences with a prefinite interruption the sentence starts with an element that has the function of a conjunct, as in example (35) below.

- (35) <A>Of anders gezegd<A>, <C>de buitenlandse politiek - <1\_pfv>en de Vietnamoorlog in het bijzonder<1\_pfv> - bleek uitermate geschikt om van 'een progressieve grondhouding' te getuigen<C>. <s3786, academic prose>
- ( <A>Or put differently<A>, <C>foreign politics - <1\_pfv>and the Vietnam war in particular<1\_pfv> - proved to be particularly suitable for demonstrating 'progressive principles'<C>.)

- (36) <C><A/zz>Volgens de minister<A/zz> is de huidige geweldsinstructie, <1\_postfv>die in grote trekken overgenomen werd van de Britten<1\_postfv>, 'helder en werkbaar'<C>. <s3216, newspaper articles>

(<C><zz>According to the minister<zz> is the current instruction on the use of force, <1\_postfv>which to a large extent was adopted from the British<1\_postfv>, 'clear and practicable'<C>.)

(<C><zz>According to the minister<zz> the current instruction on the use of force, <1\_postfv>which to a large extent was adopted from the British<1\_postfv>, is 'clear and practicable'<C>.)

### Grammatical realisation of 1\_prefin in AC1C subpattern

The prefinite interruption can be grammatically realised as a phrase or a clause. Table 14 provides the frequencies of these two main realisation groups.

**Table 14** Grammatical realisation of 1\_prefin in AC1C subpattern as phrase or clause

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
1_prefin – phrase	11	12	3	6	32 (61.5%)
1_prefin – clause	9	6	4	1	20 (38.5%)
Total	20	18	7	7	52 (100%)
<b>Dutch</b>					
1_prefin – phrase	1	0	0	0	1 (33.3%)
1_prefin – clause	2	0	0	0	2 (66.7%)
Total	3	0	0	0	3 (100%)

In English the majority of the prefinite interruptions are grammatically realised as a phrase (32 (61.5%)), and this applies especially to the newspaper genre (12 out of 18) and the leaflets genre (6 out of 7). In Dutch two of the three prefinite interruptions are realised as clauses.

Most phrases are realised as appositions in English (20 out of 32), especially in the newspaper genre (11 out of 12). The few conjuncts and disjuncts, three in total, only occur in the academic prose genre and each of the four genres contains a few adjuncts/PPs. The only phrasal prefinite interruption in the Dutch academic prose genre is realised as an apposition.

When looking at a further subcategorisation of the clauses, there is no one subcategory that dominates in English. The largest group is formed by non-finite clauses (7 out of 20), which occur mainly in the academic prose genre (3) and short

stories genre (3). The non-finite verb takes the form of a past participle in four sentences and of a present participle in three sentences. This realisation group is followed by finite adverbial clauses (4), which all occur in the academic prose genre, and by non-restrictive relative clauses (4), which all occur in the newspaper genre. The finite adverbial clauses have the semantic roles of time, reason and concession. The remaining clausal interruptions are realised as an independent clause, reporting clause and coordinated subordinate clause. In Dutch the two clausal prefinite interruptions are realised as a non-restrictive relative clause and coordinated subordinate clause. Sentence (37) contains a clausal prefinite interruption that is realised as a non-finite clause, taken from the English academic prose genre:

- (37) <A>Finally<A>, <C>historical demography and epidemiology, <1\_prefv><subcl\_advcl\_nonfin>simultaneously located within and reacting against a dominant McKeownite paradigm<subcl\_advcl\_nonfin><1\_prefv>, have undergone conceptual and methodological transformation<C>. <s4691, academic prose>

#### Grammatical realisation of 1\_postfin in AC1C subpattern

The postfinite interruption can be grammatically realised as a phrase or a clause. Table 15 provides the frequencies of these two main realisation groups.

**Table 15** Grammatical realisation of 1\_postfin in AC1C subpattern as phrase or clause

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
1_postfin – phrase	24	6	8	4	42 (76.4%)
1_postfin – clause	4	4	3	2	13 (23.6%)
Total	28	10	11	6	55 (100%)
Dutch					
1_postfin – phrase	33	16	9	19	77 (70.0%)
1_postfin – clause	15	7	8	3	33 (30.0%)
Total	48	23	17	22	110 (100%)

The loglinear analysis showed no significant three-way interaction between language, genre and grammatical realisation of the postfinite interruption in the AC1C subpattern ( $\chi^2(3) = 4.77$ ,  $p = .19$ ), nor did it show a significant two-way interaction between language and grammatical realisation ( $\chi^2(1) = 1.01$ ,  $p = .35$ , Cramer's  $V = .07$ ). In both English and Dutch the majority of postfinite interruptions are realised as phrases across all genres (English: 42 (76.4%); Dutch: 77 (70.0%)).

For English this is particularly the case in the academic prose genre, with 24 out of 28 postfinite interruptions taking the form of a phrase. In the Dutch short stories genre, on the other hand, the distribution between phrases and clauses is quite even (9-8).

The phrasal postfinite interruptions can be further subcategorised into five realisation groups: 1) appositions, 2) adjuncts/PPs, 3) conjuncts/disjuncts, 4) premodifiers, 5) the second coordinate of coordinated phrases. Table 16 presents the frequencies of these five realisation groups.

**Table 16** Grammatical realisation of phrasal 1\_postfin in AC1C subpattern in 5 realisation groups

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
Apposition	10	6	5	3	24 (57.1%)
Adjunct/PP	6	0	0	1	7 (16.7%)
Conjunct/disjunct	3	0	3	0	6 (14.3%)
Premodifier	3	0	0	0	3 (7.2%)
Coordinate_b_phrase	2	0	0	0	2 (4.7%)
Total	24	6	8	4	42 (100%)
Dutch					
Apposition	16	9	1	8	34 (44.2%)
Adjunct/PP	10	4	7	5	26 (33.8%)
Conjunct/disjunct	0	1	1	0	2 (2.6%)
Premodifier	4	2	0	6	12 (15.6%)
Coordinate_b_phrase	3	0	0	0	3 (3.8%)
Total	33	16	9	19	77 (100%)

In English the majority of phrases are realised as appositions, with the academic prose genre showing some more variation in the grammatical realisation of phrases, as this genre also contains instances of other grammatical categories, particularly adjuncts/PPs (6). In Dutch appositions also form the largest group (34 (44.2%)), in all genres except the short stories genre. The second largest realisation group in Dutch is formed by adjuncts/PPs (26 (33.8%)), especially in the short stories genre (7 out of 9). A relatively large group is again formed by premodifiers, especially in the leaflets genre (6 out of 19). Sentences (38) and (39) provide examples of an English and a Dutch premodifier respectively. Note that this is one of the very few examples in English; interruptions realised as premodifiers are particularly common in Dutch.

- (38) <A>Equally<A>, <C>Furnham and Gasson (1998) found that parents thought that their (<1\_postfv><premod>male<premod><1\_postfv>) children were about 3 IQ points higher than their own overall IQ<C>. <s5266, academic prose>
- (39) <A>Als vrouwen rond hun achttiende jaar met borstzelfonderzoek beginnen<A>, <C>wordt zo'n (<1\_postfv><premod>maandelijks<premod><1\_postfv>) onderzoek routine<C>. <s8664, leaflets>
- (<A>If women start with breast examination around the age of eighteen<A>, <C>such a (<1\_postfv><premod>monthly<premod><1\_postfv>) examination becomes routine<C>.)

The clausal postfinite interruptions are only few, especially in English (13 in total). A further subcategorisation shows that in English most clauses are realised as adverbial clauses, both finite (7 out of 13) and non-finite (5). The finite clauses take on various semantic roles, such as concession (2), result (1), condition (1), while three clauses are introduced by *as*. The non-finite clause has a present participle in three out of five cases. Dutch shows quite a high number of non-restrictive relative clauses (12 out of 33), followed by non-finite clauses (9) and finite adverbial clauses (8). Almost all non-finite clauses have a past participle and the finite clauses take on the semantic role of time/place in three cases, of condition once and are introduced by *zoals* in four cases. Sentence (40) gives an example of an interruption that is realised as a non-finite clause, taken from the Dutch academic prose genre.

- (40) <A/zz>Daarnaast<A/zz> <C>is een preventieproject, <1\_postfv><subcl\_advcl\_nonfin>gericht op jonge kinderen<subcl\_advcl\_nonfin><1\_postfv>, in een eerste stadium van uitvoering<C> (De Bok, 1998; Slot, Duivenvoorde, Orobio de Castro, Afkirin, & Speekenbrink, 2000). <s6283, academic prose>
- (<A/zz>In addition to that<A/zz> <C>is a prevention project, <1\_postfv><subcl\_advcl\_nonfin>aimed at young children<subcl\_advcl\_nonfin><1\_postfv>, in a first stage of execution<C> (De Bok, 1998; Slot, Duivenvoorde, Orobio de Castro, Afkirin, & Speekenbrink, 2000).)

### Interruptions in the XC pattern: X = two elements

The introduction of this section (8.4) already mentioned that there are only a few examples of sentences that contain an interruption if the sentence starts with two

or more sentence-initial elements (14 in English; 7 in Dutch). In English the distribution across the different genres is as follows: five in the academic prose genre, four in the newspaper genre, four in the short stories genre and one in the leaflets genre. In Dutch four of the seven instances occur in the academic prose genre, one in the newspaper genre and two in the leaflets genre. Consider the following examples, which are taken from the English academic prose genre (41), the English newspaper genre (42) and the Dutch academic prose genre (43) and contain one or more interruptions.

(41) <A>While it is clearly suggested that autobiographical memories of childhood are, <1>on the whole<1>, subject to some form of contamination across the life-span<A>, <C>Ross and Conway (1986) and Baddeley (1990), <2\_prefv>although sceptical about the efficacy of retrospective studies<2\_prefv>, have conceded that, <3\_postfv>in their view<3\_postfv>, most people's recall of past events remains relatively accurate across time<C>. <s5423, academic prose>

(42) <A>Yet<A> <B>less than two years ago<B>, <C>a major report from the Joseph Rowntree Foundation showed house building in Britain - <1\_postfv>both private and public<1\_postfv> - had fallen to its lowest rate since 1924<C>. <s1161, newspaper articles>

(43) <A>Al werd de volkskunde, <1>zowel in Vlaanderen als Nederland<1>, beheerst door een krachtig etnonationaal idioom<A>, <C>de keuzes en opties van iedere onderzoeker, <1\_prefv>binnen of buiten de universiteit<1\_prefv>, liepen telkens weer uiteen<C>. <s5216, academic prose>

(<A>Even though the ethnology was, <1>both in Flanders and the Netherlands<1>, dominated by a powerful ethno-national idiom<A>, <C>the choices and options of every researcher, <1\_prefv>inside or outside university<1\_prefv>, kept diverging<C>.)

## 8.5 Interruptions in the CX pattern

The interruptions in the CX pattern create a total of seven subpatterns. The most frequently occurring subpattern is formed by an interruption that occurs in the nucleus, creating C1CX. Another subpattern is formed by two interruptions, one of which occurs in the nucleus and one in the satellite, creating C1CX1X. A third pattern is formed by one interruption that occurs in the satellite: CX1X. A fourth



pattern is formed by two interruptions that occur in the nucleus: C12CX. A fifth pattern contains two interruptions that occur in the satellite: CX12X. The sixth pattern forms a different type of subpattern, as the interruption actually occurs between the nucleus and the satellite (C1X), similar to the A1C subpattern, as described in Chapter 6. Sentence (49) below provides an example of this pattern, which will be discussed briefly in the text following the examples. Finally, similar to the sixth subpattern, the seventh subpattern is formed by an interruption that also occurs in between two units, in this case two coordinates, which is why this is also classified as an *intrapolated* satellite instead of an *interpolated* satellite.

In this section the focus will again be on the most frequently occurring subpatterns, which are the C1CX and the CX1X subpatterns. The frequencies of all seven patterns are provided in Table 17 below and sentences (44) to (50) give examples of each type of subpattern.

**Table 17**                      **Frequencies of subpatterns formed by interruptions in CX pattern**

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
C – 1 – CX	26 (40.0%)	35 (72.9%)	23 (37.7%)	19 (51.4%)	103 (48.8%)
C – 1 – CX – 1 – X	3 (4.6%)	0 (0.0%)	2 (3.3%)	0 (0.0%)	5 (2.4%)
CX – 1 – X	26 (40.0%)	8 (16.7%)	17 (27.9%)	12 (32.4%)	63 (29.9%)
C – 1 2 – CX	5 (7.7%)	2 (4.2%)	1 (1.6%)	1 (2.7%)	9 (4.3%)
CX – 1 2 – X	4 (6.2%)	0 (0.0%)	0 (0.0%)	5 (13.5%)	9 (4.3%)
C – 1 – X	1 (1.5%)	2 (4.2%)	7 (11.5%)	0 (0.0%)	10 (4.7%)
Ca – 1 – Cb X	0 (0.0%)	1 (2.1%)	11 (18.0%)	0 (0.0%)	12 (5.7%)
Total	65 (100%)	48 (100%)	61 (100%)	37 (100%)	211 (100%)
<b>Dutch</b>					
C – 1 – CX	25 (52.1%)	10 (47.6%)	14 (18.7%)	10 (40.0%)	59 (34.9%)
C – 1 – CX – 1 – X	2 (4.2%)	0 (0.0%)	2 (2.7%)	1 (4.0%)	5 (3.0%)
CX – 1 – X	14 (29.2%)	10 (47.6%)	14 (18.7%)	12 (48.0%)	50 (29.6%)
C – 1 2 – CX	3 (6.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	3 (1.8%)
CX – 1 2 – X	3 (6.3%)	0 (0.0%)	3 (4.0%)	2 (8.0%)	8 (4.7%)
C – 1 – X	0 (0.0%)	0 (0.0%)	42 (56.0%)	0 (0.0%)	42 (24.9%)
Ca – 1 – Cb X	1 (2.1%)	1 (4.8%)	0 (0.0%)	0 (0.0%)	2 (1.2%)
Total	48 (100%)	21 (100%)	75 (100%)	25 (100%)	169 (100%)

In both languages the subpatterns in which there is only one interruption in the sentence are the most frequently occurring ones, namely C1CX (English: 103 (48.8%); Dutch: 59 (34.9%)) and CX1X (English: 63 (29.9%); Dutch: 50 (29.6%)). In Dutch the C1X subpattern is also particularly frequent, but only occurs in the short

stories genre (42 (24.9%)). As has been explained above, because this is considered a different type of interruption that has a clear function, it will not be considered any further.

As for the distribution of the most frequent subpatterns within the different genres, in the English academic prose genre the C1CX and CX1X are also the most frequent subpatterns, both occurring 26 times (40.0%). In the Dutch academic prose genre, the C1CX pattern is more frequent (25 (52.1%)) than the CX1X subpattern (14 (29.2%)).

In the newspaper genre the situation is reversed for English and Dutch, as in this genre both patterns have a similar frequency in Dutch (both 10 sentences, (47.6%)), whereas in English the C1CX pattern occurs more frequently (35 (72.9%)) than the CX1X subpattern (8 (16.7%)).

In the English short stories genre, the C1CX and CX1X patterns are again the most frequently occurring subpatterns (23 (37.7%) and 17 (27.9%)), but the C1X subpattern (7 (11.5%)) and Ca-1-Cb subpattern (11 (18.0%)) also occur quite frequently. In Dutch the frequencies have a different distribution, with the C1X subpattern being the most frequent one (42 (56.0%)), followed by the C1CX and CX1X subpatterns, which have similar distribution (14 (18.7%)).

Finally, in the leaflets genre in both languages the C1CX subpattern (English: 19 (51.4%); Dutch: 10 (40.0%)) and the CX1X subpattern (English: 12 (32.4%); Dutch: 12 (48.0%)) are again the most frequent ones.

(44) <C>The alternative, <1\_pfv>as Mr Hoon admitted<1\_pfv>, is for the present situation to go on and on<C> - <D>with British troops bogged down<D>, <E>facing an increasingly hostile population<E>. <s26, newspaper articles>

(45) <C>Het systeem van het jaarlijkse generaal kapittel en de (<1\_pfv>althans in theorie<1\_pfv>) jaarlijkse visitatie van alle orde kloosters door vader-abten maakte uniformering en controle mogelijk<C>, <D>met de schriftelijke optekening (<1>van de kapittelbesluiten, van de geldende wetgeving, van de visitaties<1>) als machtig instrument<D>. <s5045, academic prose>

(<C>The system of the annual general chapter and the (<1\_pfv>at least in theory<1\_pfv>) annual visitation of all monastic orders by father abbots made uniformization and inspection possible<C>, <D>with the written recording (<1>of the chapter decisions, of the applicable legislation, of the visitations<1>) as a powerful device<D>.)

- (46) <C>Children who have impairments of coordination and are slow to learn movement skills may also have difficulties with social adjustment and educational success<C>, <D>based, <1>at least in part<1>, on their movement difficulties<D> (Cantell, Smyth, & Ahonen, 1994; Losse et al., 1991). <s5478, academic prose>
- (47) <C>His companion, <1\_prefv>Pedro Ladron<1\_prefv>, was, <2\_postfv>however<2\_postfv>, a very different proposition<C>, <D>a young man of unshakeable resolve<D>, <E>who signed himself 'servus servorum Dei'<E>. <s4068, academic prose>
- (48) <C>The cultivation of science was a particularly important part of knowledge-based societies<C>, <D>although it should not be forgotten, <1\_postfv>as Roy Porter argued<1\_postfv>, that - <2\_postfv>as the phrase "literary and scientific" suggests<2\_postfv> - this remained part of a broad polite civic culture<D>. <s4734, academic prose>
- (49) <C>Waar kan ik heen<C>, <1>dacht ze<1>, <D>ik kan nergens heen<D>. <s14084, short stories>
- ( <C>Where can I go<C>, <1>she thought<1>, <D>I can't go anywhere<D>.)
- (50) <Ca>It was late<Ca>, <1\_postfv>after eleven<1\_postfv>, <Cb>and there was a sense of purpose in the way that the lads were striding ahead<Cb>, <D>which was seldom apparent in their usual ambling procession<D>. <s11283, short stories>

Sentence (49) was identified above as representing a special type of interruption, as it does not actually interrupt another discourse unit, but occurs between two discourse units. Table 17 above shows that this subpattern almost exclusively occurs in the short stories genre in both languages, with Dutch showing a particularly high frequency (42). As this discourse unit has one particular function in this genre in both languages, namely that of a reporting clause that 'interrupts' the reported speech, it will not be further analysed. In addition, the interruption in sentence (50) also deserves a special remark, as the status of this discourse unit is ambiguous and could be analysed in various ways. In the present study, it is analysed as an interruption that occurs in between two coordinated nuclei. However, it could also be analysed as a <D> satellite that follows the first coordinate and precedes the second one. When analysed in this latter way, the sentence would still belong to the same main sentence pattern, namely the CX pattern. In other words, either analysis does not affect the count of the number of

sentences that belong to the main sentence patterns. Moreover, this sentence pattern only has few occurrences, as Table 17 above shows, and will therefore not be analysed in more detail.

### The C1CX subpattern

Similar to the interruptions in the other subpatterns discussed so far, the interruptions in the C1CX subpattern can occur before or after the finite verb of the nucleus. The frequencies of both positions are presented in Table 18.

**Table 18** Frequencies of different positions of interruptions with respect to finite verb in C1CX subpattern

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
C – 1_prefin – CX	14	22	15	7	58 (56.3%)
C – 1_postfin – CX	12	13	8	12	45 (43.7%)
Total	26	35	23	19	103 (100%)
<b>Dutch</b>					
C – 1_prefin – CX	9	4	6	4	23 (39.0%)
C – 1_postfin – CX	16	6	8	6	36 (61.0%)
Total	25	10	14	10	59 (100%)

The loglinear analysis showed no significant three-way interaction between language, genre and position of the interruption in the C1CX subpattern ( $\chi^2(3)=1.28$ ,  $p=.73$ ), nor did it show a significant two-way interaction between language and position of interruption at the predetermined alpha level of .01 ( $\chi^2(1)=3.85$ ,  $p=.05$ , Cramer's  $V=.17$ ). Table 18 shows that although it does not concern a significant difference, the distribution of interruptions that occur before and after the finite verb of the nucleus is not similar for English and Dutch. Specifically, Dutch contains a higher frequency of postfinite interruptions (36 (61.0%)), whereas English shows a more even distribution (58 (56.3%) prefinite vs. 45 (43.7% postfinite)

### The grammatical realisation of 1\_prefinite in the C1CX subpattern

The prefinite interruption in the C1CX subpattern can be realised as a phrase or a clause. In both English and Dutch, the majority of interruptions take the form of a phrase (English: 35 out of 58 (60.3%); Dutch (14 out of 23 (60.9%)). For English, this applies to all genres, with the newspaper genre containing most phrases (16 out of

22). For Dutch, there is a little more variation between the genres, although it should be noted that the total number of occurrences of this pattern in Dutch is rather low on the whole. The phrase is found in the academic prose genre (7 out of 9) and the leaflets genre (1 out of 3), whereas the clause is a little more frequent in the short stories genre (4 out of 6).

When realised as a phrase, in both languages, but especially in English, this is typically realised as an apposition (English: 25 out of 35; Dutch: 8 out of 14). In both languages, the second largest group is formed by adjuncts/PPs (English: 7; Dutch: 5) and the third by conjuncts in English (3) and a premodifiers (1) in Dutch.

When realised as a clause, the interruption most frequently takes the form of a non-finite clause in English (6 out of 23), especially in the newspaper genre (4), followed by finite clauses (5), which mainly occur in the short stories genre (4). The 6 sentences in the academic prose genre are realised as reporting clauses, coordinated subordinate clauses and one independent clause. In Dutch, there are even fewer occurrences than English and these take the form of a finite adverbial clause (3), a non-finite adverbial clause (2), non-restrictive clause (3) and an independent clause (1). Sentence (51) is taken from the English newspaper genre and contains an interruption that is realised as a non-finite clause with a present participle.

- (51) <C>The change to China's constitution, <1\_prefv><subcl\_advcl\_nonfin>adding a clause guaranteeing private property rights<subcl\_advcl\_nonfin><1\_prefv>, may be an important step in protecting individual landowners<C>, <D>if enforced<D>. <s1150, newspaper articles>

### **The grammatical realisation of 1\_postfinite in the C1CX subpattern**

The postfinite interruption in the C1CX subpattern can also be realised as a phrase or a clause. In both languages this is typically realised as a phrase (English: 36 (80.0%); Dutch: 30 (83.3%)). The newspaper genre contains only phrases in Dutch (6 sentences) and in English 11 out of 13 take the form of a phrase.

Similar to the grammatical realisation of prefinite interruptions, most phrasal interruptions are realised as an apposition (English: 15 of 30 sentences; Dutch: 22 of 29 sentences). In both languages, the apposition is particularly common in the newspaper genre (English: 7 of 10; Dutch: 5 of 6). In Dutch, this also applies to the academic prose genre (9 of 13), whereas in English the few occurrences of phrases take the form of adjuncts/PPs (3) and conjuncts (4). Although the number is low, English contains six interruptions that are realised as

conjuncts, whereas Dutch contains none and the same applies to coordinated phrases, which occur four times in English and only once in Dutch.

As for the clausal interruptions, although very infrequent in both languages, these are grammatically realised in various ways, i.e. as finite adverbial clauses, mainly of time and condition (English: 4 of 9; Dutch: 2 of 6), as non-finite clauses (English: 2, Dutch: 1), non-restrictive clauses (English: 1, Dutch: 1) and independent clauses (English: 1, Dutch: 2).

### The CX1X subpattern

The interruptions that occur in the CX1X subpattern can occur before the finite verb of the X element, i.e. typically a D or E satellite, or it can occur after the finite verb of one of these satellites. In addition to this, there are a number of cases in which the D or E satellites contain no finite verb, but for example a non-finite verb. Sentence (52) provides an example of an interruption that occurs in the D-satellite that contains a non-finite verb and Table 19 presents the frequencies of the different positions of the interruptions in the CX1X subpattern.

- (52) <C>Professor Larry Witherell has succeeded in reopening the debate into one of the most intriguing developments of modern British political history<C> - <D>the precise role of Parliament, <1>or at least of parliamentarians<1>, in bringing down a government in possession of one of the largest House of Commons majorities of the twentieth century<D>. <s4329, academic prose>

**Table 19**                      **Frequencies of different positions of interruptions with respect to finite verb in CX1X subpattern**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
CX – 1_prefin – X	7	1	2	2	12 (19.0%)
CX – 1_postfin – X	11	7	7	4	29 (46.0%)
CX – 1_no fin verb – X	8	0	8	6	22 (35.0%)
<b>Total</b>	<b>26</b>	<b>8</b>	<b>17</b>	<b>12</b>	<b>63 (100%)</b>
<b>Dutch</b>					
CX – 1_prefin – X	4	5	4	3	16 (32.0%)
C – 1_postfin – CX	7	4	4	4	19 (38.0%)
CX – 1_no fin verb – X	3	1	6	5	15 (30.0%)
<b>Total</b>	<b>14</b>	<b>10</b>	<b>14</b>	<b>12</b>	<b>50 (100%)</b>

In both languages, the majority of interruptions in the CX1X subpattern occur after the finite verb of the D or E satellite, if this has a finite verb (English: 29 (46.0%); Dutch: 19 (38.0%)). The interruptions that occur in the D or E satellites that contain no finite verb predominantly occur in the short stories genre (English: 8 of 17; Dutch: 6 of 14) and the leaflets genre (English: 6 of 12; Dutch: 5 of 12) in both languages. The Dutch newspaper genre is the only genre in which most interruptions occur before the finite verb of the D or E satellite (5 of 10).

### **Grammatical realisation of 1 in CX1X subpattern**

The interruptions that occur before the finite verb of the X element are typically realised as phrases in both English and Dutch (English: 7 of 12; Dutch: 11 of 16). In both languages these phrases mainly take the form of appositions (3 in English; 5 in Dutch) or adjuncts/PPs (3 in English; 4 in Dutch). The few clausal interruptions are realised as non-restrictive (2) and non-finite clauses in English (2) and as finite adverbial clauses (3) in Dutch.

The interruptions that occur after the finite verb again predominantly take the form of a phrase in both languages (English: 21 of 29; Dutch: 16 of 19). In both languages most phrasal interruptions are realised as appositions (English: 7 of 21; Dutch: 8 of 16), followed by adjuncts/PPs (English: 7; Dutch: 4). In English a third group is again formed by conjuncts and disjuncts (5), whereas in Dutch this is formed by premodifiers (2). The few clausal interruptions in English take the form of a non-finite clause (4 of 8) and a few other categories, such as reporting and comment clauses. In Dutch the three clausal interruptions are realised as finite adverbial clauses.

Finally, the third group of interruptions that occur in the satellites that contain no finite verb are again mainly realised as phrases (English: 18 of 22; Dutch: 13 of 15). Most of these phrases take the form of appositions (English: 11; Dutch: 6), followed by adjuncts/PPs (English: 4; Dutch: 6), and again a few conjuncts in English (2). The very few clausal interruptions are independent clauses (2) and a finite adverbial clause in English and an adverbial clause and reporting clause in Dutch.

Sentence (53) provides an example of the most frequently occurring type of interruption in this pattern, i.e. one that takes the form of an apposition.

(53) <C>Het Nederlands tekort over 2003 is uitgekomen op 3,2 procent<C>,  
<D>zo heeft het Centraal Bureau voor de Statistiek  
(<1\_postfv><appos\_NP>CBS<appos\_NP><1\_postfv>) woensdag  
bekendgemaakt<D>. <s3150, newspaper articles>

(<C>The Dutch deficit over 2003 came out at 3.2 percent<C>, <D>as the  
Central Bureau of Statistics (<1\_postfv><appos\_NP>CBS<appos\_NP><1\_postfv>)  
announced Wednesday<D>.)

## 8.6 Interruptions in XCX pattern

Even though the interruptions in the XCX pattern appear low, when looking at the number of sentences that contain an interruption with respect to the total number of sentences with this pattern, the percentage of interruptions is highest in the XCX pattern when compared to the other main sentence patterns. That is, 14.2% of all English sentences that belong to the XCX pattern have an interruption and 10.6% of all Dutch sentences.

Similar to the interruptions in the other main sentence patterns, the interruptions in the XCX pattern create a number of subpatterns: XC1CX, XC1CX1X, XCX1X, XC12CX, XCX12X, XC1X. Sentences (54) to (58) provide examples of the different subpatterns, excluding the XC1X subpattern, as this again concerns a reporting clause that occurs between two discourse units that take the grammatical form of reported speech (see discussion of C1X above). The frequencies of the different subpatterns are provided in Table 20 below.



- (54) <A>Applying this ease of accessibility perspective to our Calidornian friends<A>, <C>Niall, <1\_prefv>despite having recalled more positive information<1\_prefv>, might have been less likely to have supported Connery<C> - <D>if he had a difficult time retrieving the seven positive attributes<D>. <s5845, academic prose>
- (55) <A>Nevertheless<A>, <C>the records of the proceedings of the Genoese councils reveal an acceptance of the realities of disagreement in public life, of conflicts of interest, of the need to find, <1\_postfv>if nothing else could be managed<1\_postfv>, the least unpopular solution to the problem<C>, <D>that showed as acute a political sense, <2\_postfv>in its way<2\_postfv>, as that of the renowned political elites of Florence and Venice<D>. <s4162, academic prose>
- (56) <A/zz>Zeker ten aanzien van de inhoudelijke culturele vormgeving van de identiteiten<A/zz> <C>is er sprake van een sterke verschuiving<C>: <D>in de vroege Nieuwe Tijd komt, <1\_postfv>zowel in de Nederlanden, als in Engeland en Frankrijk<1\_postfv> de nadruk veel meer te liggen op oudtestamentische en klassieke symboliek<D>. <s4836, academic prose>
- <C><zz>Especially with respect to the cultural design of the identities<zz> there is a strong shift<C>: <D>in the early Modern Time much emphasis is, <1\_postfv>both in the Netherlands, and England and France<1\_postfv> on old testament and classical symbolism<D>.)
- (57) <A/zz>However<A/zz> <C>the proportion of 2-year-olds in the present study who passed the test (<1\_prefv>23%<1\_prefv>) is higher than that reported approximately two decades ago (<2\_postfv>11%<2\_postfv>, Weinraub et al., 1984; 2% Blakemore et al., 1979)<C>, <D>suggesting that the passage of time has not altered gender stereotypes<D> (see also Lueptow, Garovich, & Lueptow, 1995). <s5801, academic prose>
- (58) <A/zz>Na een nog vrij ontspannen vrijdag en zaterdag<A/zz> <C>had gisteren de wroeging dan toch de overhand gekregen<C>: <D>door zich, <1\_prefv>na de diefstal<1\_prefv>, <2\_prefv>op het politiebureau<2\_prefv>, als de buurvrouw van 154 voor te doen<D>, <3>zij het ook tijdens een door angst getroubleerd ogenblik<3>, <E>had zij deze vrouw feitelijk vals beschuldigd<E> ... <F>terwijl zij zelf schuldig was<F>! <s15464, short stories>
- <C><zz>After a rather relaxed Friday and Saturday<zz> remorse yesterday gained the upper hand after all<C>: <D>by, <1\_prefv>after the robbery<1\_prefv>, <2\_prefv>at the police station<2\_prefv>, pretending to be the neighbour of 154<D>, <3>albeit at a moment troubled by fear, <E>she had in fact falsely accused this woman<E> ... <F>while she was the one who was guilty<F>!)

**Table 20**      **Frequencies of subpatterns formed by interruptions in CX pattern**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
XC – 1 – CX	7	2	7	5	21 (39.6%)
XC – 1 – CX – 1 – X	3	0	0	0	3 (5.7%)
XCX – 1 – X	10	1	3	4	18 (34.0%)
XC – 1 2 – CX	3	0	3	0	6 (11.3%)
XCX – 1 2 – X	1	0	0	3	4 (7.5%)
XC – 1 – X	0	1	0	0	1 (1.9%)
<b>Total</b>	<b>24</b>	<b>4</b>	<b>13</b>	<b>12</b>	<b>53 (100%)</b>
<b>Dutch</b>					
XC – 1 – CX	9	3	9	1	22 (43.2%)
XC – 1 – CX – 1 – X	0	0	0	0	0 (0.0%)
XCX – 1 – X	8	2	6	4	20 (39.2%)
XC – 1 2 – CX	1	0	1	0	2 (3.9%)
XCX – 1 2 – X	2	0	2	0	4 (7.8%)
XC – 1 – X	0	0	3	0	3 (5.9%)
<b>Total</b>	<b>20</b>	<b>5</b>	<b>21</b>	<b>5</b>	<b>51 (100%)</b>

In both English and Dutch, the subpatterns in which either the nucleus contains one interruption, XC1CX, or the appended satellite contains one interruption, XCX1X, are the most frequent ones. The former occurs in 21 (39.6%) sentences in English and 22 (43.2%) sentences in Dutch, and in both languages this pattern occurs predominantly in the academic prose genre and the leaflets genre. The latter pattern has 18 (34.0%) occurrences in English and 20 (39.2%) in Dutch. In English, this pattern predominantly occurs in the academic prose genre (10), whereas in Dutch, in addition to the academic prose genre (8), it also has a relatively high frequency in the leaflets genre (6). The frequency of the other subpatterns is markedly lower. The XC1CX1X subpattern only occurs in the English academic prose genre (3). Similarly, the subpattern in which the nucleus contains two interruptions, XC12CX, again has more occurrences in English (6) than in Dutch (2), although frequencies for both languages are low. The subpattern in which the appended satellite contains two interruptions, XCX12X has four occurrences in both languages. Finally, the XC1X subpattern only has one occurrence in English and three in Dutch.

Due to the low frequency of four of the six patterns, we will only focus on the most frequent patterns in this section, i.e. XC1CX and XCX1X.

### The XC1CX subpattern

The XC1CX subpattern occurs in 21 sentences in English and 22 sentences in Dutch. As Table 20 above showed, in English the frequencies of this pattern are similar in the academic prose genre (7) and short stories genre (7), followed by the leaflets genre (5) and the newspaper genre (2). In Dutch, frequencies are also similar for the academic prose genre (9) and short stories genre (9), followed by the newspaper genre (3) and the leaflets genre (1).

As for the position of the interruptions in this subpattern with respect to the finite verb in the nucleus, in English 6 sentences contain a prefinite interruption and 15 a postfinite interruption. In Dutch there are no sentences with a prefinite interruption, which means that all interruptions occur in postfinite position. This is in line with the results for the position of interruptions in the XC pattern.

As for the grammatical realisation of the prefinite interruptions in English, these take the form of an apposition (3), a non-restrictive relative clause and an adjunct/PP (2). These frequencies show that the majority of interruptions in this pattern take the form of a phrase instead of a clause. The postfinite interruptions are also predominantly realised as phrases in English, with 8 taking the form of an apposition, 5 that of an adjunct/PP and only 2 of non-finite clauses. None of the genres shows a particular preference for one of the realisation forms. In Dutch the majority of interruptions are also realised as phrases, with 6 taking the form of an apposition, 6 that of an adjunct/PP, 2 of a premodifier and 1 of a discourse marker. The 7 clauses are realised as non-restrictive clauses (2), one finite and one non-finite adverbial clause, two independent clauses and one reporting clause. Sentence (59) presents an example of the XC1CX pattern, taken from the Dutch newspaper genre.

(59) <A/zz>Gisteren<A/zz> <C>bleek in een debat met minister Zalm  
(<1\_postfv>Financien<1\_postfv>) dat de coalitiepartners CDA, VVD en D66 het kabinet nog steeds steunen in haar voornemen het begrotingstekort dit jaar terug te dringen tot 2,9 procent<C>, <D>wat betekent dat minstens 2 miljard euro extra bezuinigd zal moeten worden<D>. <s2556, newspaper articles>

(<A/zz>Yesterday<A/zz> <C>it appeared in a debate with Minister Zalm  
(<1\_postfv>Finances<1\_postfv>) that the coalition partners CDA, VVD en D66 still support the cabinet in its plan to reduce the budget deficit to 2.9 percent<C>, <D>which means an additional cut back of at least 2 billion euros<D>.)

### The XCX1X subpattern

The XCX1X subpattern occurs in 18 sentences in English and 20 sentences in Dutch. In English, most sentences occur in the academic prose genre (10), 4 sentences occur in the leaflets genre, 3 in the short stories genre and one in the newspaper genre. In Dutch, most sentences also occur in the academic prose genre (8), 4 sentences occur in the leaflets genre, 6 in the short stories genre and only 2 in the newspaper genre.

With respect to the position of the interruptions, in English 6 interruptions occur before the finite verb, 5 after the finite verb and in 7 sentences there is no finite verb. In Dutch, only 3 interruptions occur before the finite verb, 10 occur after the finite verb and 7 sentences contain no finite verb.

In English, of the prefinite interruptions 3 sentences are realised as phrases of different grammatical categories (apposition, conjunct, adjunct/PP), and three sentences as clauses, two non-finite clauses and one independent clause. The postfinite interruptions take the form of phrases in 4 sentences (apposition and adjunct/PP) and in 2 as verbless clauses. The interruptions that occur in sentences that contain no finite verb take the form of an apposition in 3 cases, an adjunct in 1 case, and an independent clause in two cases. In Dutch the prefinite interruptions are realised as phrases (2 apposition, 1 premodifier). The postfinite interruptions also all take the form of a phrase, except for one that is realised as an adverbial clause. The phrases consist of 2 appositions, 5 adjuncts/PPs and 2 premodifiers. The final group of interruptions mainly consist of appositions (4), 1 premodifier, disjunct and one fragment. Sentence (60) presents an example of the XCX1X pattern, taken from the English academic prose genre.

- (60) <A>Second<A>, <C>these were being pursued by different interests through the two agencies of crown and parliament<C>, <D>which despite the apparent equipoise of the Restoration Settlement were endemically, <1\_postfv>and specifically in 1673<1\_postfv>, in conflict<D>. <s4678, academic prose>

## 8.7 Interruptions and punctuation

Although the findings are not based on a detailed analysis, a quick analysis of the types of punctuation marks used to separate interpolated satellites from the rest of the sentence reveals a number of differences between the languages. Specifically,

when looking at the type of punctuation marks used around interruptions in the most frequently occurring subpatterns, such as the C1C and AC1C subpatterns, it becomes clear that English shows a much higher frequency of paired dashes around the interruptions, whereas Dutch appears to contain a particularly high frequency of brackets. Furthermore, a closer analysis of the type of punctuation mark in combination with the grammatical realisation and the semantic content of the interruption appears to show that particular punctuation marks are reserved for interruptions with a particular function. Specifically, whereas the parenthesis is often used in combination with the interruptions that are realised as premodifiers or appositions in especially the Dutch newspaper and leaflets genre, giving these units a clear backgrounding function, paired dashes are used rather frequently in English, especially in the newspaper genre, to introduce interruptions that contain some sort of comment of the writer with respect to the information that surrounds it (cf. Chapter 9, 9.3.2 on closer analysis of function of interruptions in English newspaper genre). It should be noted, however, that the vast majority of interruptions in both languages are typically surrounded by commas. Examples of the typical use of brackets around premodifiers in Dutch and the use of dashes in English were already presented in sentences such as (16) and (17) above, which are repeated as (61) and (62) below.

- (61) <C>Onderhandelingen over (<1\_prefv><premod>het traject  
naar<premod><1\_prefv>) het lidmaatschap van de Europese Unie lopen al  
vele jaren<C>. <s3107, newspaper articles>
- (<C>Negotiations about (<1\_prefv><premod>the road to<premod><1\_prefv>)  
membership of the European Union have been going on for many  
years<C>.)
- (62) <C>The decision to release five of the nine - <1\_prefv><coord\_b\_phr>and the  
admission by Home Secretary David Blunkett that they pose no  
threat<coord\_b\_phr><1\_prefv> - is shameful recognition they were wrongly  
detained<C>. <s339, newspaper articles>

## 8.8 A characterisation of the 6 most frequent subpatterns

This section will present the main findings of the chapter in a condensed format in the form of an overview, focussing on the most frequent subpatterns created by interruptions. An analysis of the different types of interruptions that can occur in sentences has shown that, on the whole, they occur more frequently in English than in Dutch. Specifically, English contains more sentences than Dutch that contain one interruption and it contains more sentences that contain two interruptions. Moreover, the various sections in this chapter have shown that interruptions can create a wide range of sentence patterns. Within each of the four main sentence patterns, the interruptions create an additional number of subpatterns. The most frequent of these subpatterns are similar for both languages. The most frequently occurring subpatterns with the interruption occurring in the nucleus are the following ones: C1C, AC1C, C1CX, XC1CX. The following two subpatterns in which the interruption occurs in the appended satellite are the following ones: CX1X XCX1X. Similar for both languages is also that most interruptions occur in the academic prose genre, followed by the newspaper genre, followed by the leaflets genre and last by the short stories genre. With respect to the latter genre, Dutch does, however, show a higher frequency of a special type of sentence pattern, the C1X pattern, in which the interpolated satellite does not interrupt another discourse unit, but occurs in between two discourse units (cf. 9.3.3 on *Interruptions* and *End of sentence* on position of reporting clauses in English and Dutch). This pattern predominantly occurs in the short stories genre and the interruption is typically realised as a reporting clause. Table 21 presents the frequencies of the most frequent subpatterns created by sentences that contain one interruption. It also shows how the total number of these subpatterns combined relates to the total number of sentences per genre per language.

**Table 21**                      **Frequencies of most frequent subpatterns formed by interruptions in main sentence patterns**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
C1C	139 (54.3%)	150 (67.0%)	107 (61.1%)	81 (60.5%)	477 (60.5%)
AC1C	48 (18.8%)	28 (12.5%)	18 (10.3%)	13 (9.7%)	107 (13.6%)
C1CX	26 (10.2%)	35 (15.6%)	23 (13.1%)	19 (14.2%)	103 (13.0%)
CX1X	26 (10.2%)	8 (3.6%)	17 (9.7%)	12 (8.9%)	63 (8.0%)
XC1CX	7 (2.7%)	2 (0.9%)	7 (4.0%)	5 (3.7%)	21 (2.6%)
XCX1X	10 (3.9%)	1 (0.5%)	3 (1.7%)	4 (3.0%)	18 (2.3%)
Total	256 (100%)	224 (100%)	175 (100%)	134 (100%)	789 (100%)
Percentage of total no of sentences	17.8% (1438)	12.1% (1844)	5.5% (3169)	8.4% (1598)	9.8% (8040)
<b>Dutch</b>					
C1C	132 (55.2%)	111 (69.8%)	55 (47.8%)	64 (56.6%)	362 (57.8%)
AC1C	51 (21.4%)	23 (14.5%)	17 (14.8%)	22 (19.5%)	113 (18.0%)
C1CX	25 (10.5%)	10 (6.3%)	14 (12.2%)	10 (8.8%)	59 (9.4%)
CX1X	14 (5.9%)	10 (6.3%)	14 (12.2%)	12 (10.6%)	50 (8.0%)
XC1CX	9 (3.8%)	3 (1.9%)	9 (7.8%)	1 (0.8%)	22 (3.5%)
XCX1X	8 (3.3%)	2 (1.3%)	6 (5.2%)	4 (3.5%)	20 (3.2%)
Total	239 (100%)	159 (100%)	115 (100%)	113 (100%)	626 (100%)
Percentage of total no of sentences	13.7% (1740)	9.0% (1754)	3.7% (3154)	5.5% (2060)	7.1% (8708)

Table 21 shows that the C1C subpattern is the most frequently occurring subpattern in both languages (English: 477 (60.5%); 362 (57.8%)). This is followed by the AC1C subpattern (English: 107 (13.6%); Dutch: 113 (18.0%)), the C1CX subpattern (English: 103 (13.0%); Dutch: 59 (9.4%)), the CX1X subpattern (English: 63 (8.0%); Dutch: 50 (8.0%)), the XC1CX subpattern (English: 21 (2.6%); Dutch: 22 (3.5%)) and finally by the XCX1X subpattern (English: 18 (2.3%); Dutch: 20 (3.2%)). Across all four genres, English shows a higher frequency of sentences with interruptions than Dutch, with the academic prose genre in both languages containing most interruptions (English: 256, 17.8% of all sentences in genre; Dutch: 239, 13.7% of all sentences in this genre), followed by the newspaper genre (English: 224, 12.1% of all; Dutch: 159, 9.0% of all), the leaflets genre (English: 134, 8.4% of all; Dutch: 113, 5.5% of all) and finally short stories (English: 175, 5.5% of all; Dutch: 115, 3.7% of all).

### Position of interruptions

Interpolated satellites can occur at various positions in the sentence. One way of categorising them is by positioning them with respect to the finite verb of the discourse unit in which they occur. Table 22 presents the frequencies of the interruptions that either occur before or after the finite verb. The analysis of the interruptions in the C1C, the C1CX and the CX1X subpatterns in the sections above showed that there are a small number of interruptions that occur in a discourse unit that does not contain a finite verb. As these concern very few cases, we will only present the frequencies of the interruptions that occur before or after the finite verb.

**Table 22**                      **Frequencies of prefinite and postfinite interruptions in 6 subpatterns formed by interruptions**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Prefinite interruptions in 6 subpat	115 (47.1%)	131 (58.5%)	69 (42.9%)	55 (44.0%)	370 (49.1%)
Postfinite interruptions in 6 subpat	129 (52.9%)	93 (41.5%)	92 (57.1%)	70 (56.0%)	384 (50.9%)
<b>Total</b>	<b>244 (100%)</b>	<b>224 (100%)</b>	<b>161 (100%)</b>	<b>125 (100%)</b>	<b>754 (100%)</b>
<b>Dutch</b>					
Prefinite interruptions in 6 subpat	81 (34.5%)	61 (39.1%)	31 (29.0%)	38 (37.3%)	211 (35.2%)
Postfinite interruptions in 6 subpat	154 (65.5%)	95 (60.9%)	76 (71.0%)	64 (62.7%)	389 (64.8%)
<b>Total</b>	<b>235 (100%)</b>	<b>156 (100%)</b>	<b>107 (100%)</b>	<b>102 (100%)</b>	<b>600 (100%)</b>

In English the number of interruptions that occur before the finite verb is rather similar to the number of interruptions that follow the finite verb, whereas in Dutch the number of interruptions that follow the finite verb is considerably higher than the ones that precede it. Sections 8.4 and 8.6 above already showed that especially in sentence patterns in which the nucleus is preceded by a prepended satellite, Dutch shows a preference for interruptions that occur after the finite verb, and suggested that this can be explained the verb-second principle in Dutch.

The distribution of prefinite and postfinite interruptions across the different genres in English rather similar, with newspaper articles showing a slightly higher frequency of prefinite interruptions (131 (58.5%)) and the short stories (92 (57.1%)) and leaflets (70 (56.0%)) a higher frequency of postfinite interruptions.

In Dutch the frequency of postfinite interruptions is higher across the genres, with the short stories genre showing the highest frequency (76 (71.0%)).



### Grammatical realisation: phrases vs. clauses

The prefinite and postfinite interruptions can be roughly categorised into either phrases or clauses. Table 23 presents the frequencies of both groups.

**Table 23**                    **Frequencies of interruptions in 6 subpatterns realised as phrases or clauses**

English	Academic prose	Newspaper articles	Short stories	Leaflets	Total
1 – phrase in 6 subpatterns	179 (73.4%)	155 (69.2%)	100 (62.1%)	97 (77.6%)	531 (70.4%)
1 – clause in 6 subpatterns	65 (26.6%)	69 (30.8%)	61 (37.9%)	28 (22.4%)	223 (29.6%)
Total	244 (100%)	224 (100%)	161 (100%)	125 (100%)	754 (100%)
Dutch					
1 – phrase in 6 subpatterns	169 (71.9%)	115 (73.7%)	57 (53.3%)	89 (87.3%)	430 (71.7%)
1 – clause in 6 subpatterns	66 (28.1%)	41 (26.3%)	50 (46.7%)	13 (12.7%)	170 (28.3%)
Total	235 (100%)	156 (100%)	107 (100%)	102 (100%)	600 (100%)

In both languages the vast majority of interruptions take the form of a phrase across the different genres (English: 531 (70.4%); Dutch: 430 (71.7%)). In English, the frequencies of phrases are particularly high in the leaflets genre (97 (77.6%)) and the academic prose genre (179 (73.4%)), with the short stories genre containing the highest frequency of clauses when compared to the other genres (61 (37.9%)). Dutch shows a similar pattern, with the leaflets genre also showing the highest frequency of phrases (89 (87.3%)) and the short stories genre the lowest frequency (57 (53.3%)).

### Most frequent grammatical forms of phrasal interruptions

Tables 24 and 25 present the most frequently occurring grammatical realisations of the prefinite and postfinite phrasal interruptions respectively in the C1C, AC1C, C1CX, CX1X, XC1CX and XCX1X subpatterns. As most phrasal interruptions in both languages typically take one of three grammatical forms, the focus will be on presenting an overview of these grammatical categories, with other grammatical realisation being grouped as ‘other’ (for exact make-up of this group, see analysis of individual subpatterns in respective sections in this chapter). For English, the most frequent forms are appositions, adjuncts/PPs and conjuncts, showing only two instances of premodifiers. For Dutch the most frequent forms are appositions, adjuncts/PPs and premodifiers, showing only one instance of a conjunct. Note that although the frequencies of the ‘other’ group may appear high in certain cases, this

group consists of various grammatical forms and no one form in particular. The main grammatical forms that belong to this group are disjuncts and the second coordinators of coordinated phrases.

**Table 24**                    **Frequencies of most frequent phrasal realisations of prefinite interruptions in C1C, AC1C, C1CX, CX1X, XC1CX, XCX1X subpatterns**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Apposition	31 (38.8%)	71 (81.6%)	17 (45.9%)	30 (75.0%)	149 (61.1%)
Adjunct/PP	24 (30.0%)	10 (11.5%)	15 (40.5%)	3 (7.5%)	52 (21.3%)
Conjunct	19 (23.8%)	0 (0.0%)	2 (5.4%)	3 (7.5%)	24 (9.8%)
Premodifier	1 (1.2%)	0 (0.0%)	0 (0.0%)	1 (2.5%)	2 (0.8%)
Other phrases	5 (6.2%)	6 (6.9%)	3 (8.1%)	3 (7.5%)	17 (7.0%)
<b>Total</b>	<b>80 (100%)</b>	<b>87 (100%)</b>	<b>37 (100%)</b>	<b>40 (100%)</b>	<b>244 (100%)</b>
<b>Dutch</b>					
Apposition	32 (59.3%)	32 (82.0%)	4 (28.6%)	24 (75.0%)	92 (66.2%)
Adjunct/PP	15 (27.8%)	0 (0.0%)	9 (64.3%)	3 (9.4%)	27 (19.4%)
Conjunct	1 (1.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.7%)
Premodifier	4 (7.4%)	6 (15.4%)	0 (0.0%)	4 (12.5%)	14 (10.1%)
Other phrases	2 (3.7%)	1 (2.6%)	1 (7.1%)	1 (3.1%)	4 (3.6%)
<b>Total</b>	<b>54 (100%)</b>	<b>39 (100%)</b>	<b>14 (100%)</b>	<b>32 (100%)</b>	<b>139 (100%)</b>

In both languages, most prefinite phrasal interruptions take the form of appositions (English: 149 (61.1%); Dutch: 92 (66.2%)). The second largest group is formed by adjuncts/PPs (English: 52 (21.3%); Dutch: 27 (19.4%)). The languages show differences in the frequencies of the third largest realisation group: in English this is formed by conjuncts (24 (9.8%)), whereas in Dutch this is formed by premodifiers (14 (10.1%)).

Differences between the languages can also be found at the level of genre. For instance, in the English academic prose genre appositions form the largest group (31 (38.8%)), but these are closely followed by adjuncts/PPs (24 (30.0%)), and finally by conjuncts (19 (23.8%)). In Dutch, the vast majority of prefinite interruptions take the form of an apposition (32 (59.3%)), followed by adjuncts/PPs (15 (27.8%)) and premodifiers (4 (7.4%)).

In the newspaper genre in both languages appositions form by far the largest realisation group (English: 71 (81.6%); Dutch (32 (82.0%)). In English, a small group is realised as adjuncts/PPs (10 (11.5%)), whereas in Dutch this is again the group of premodifiers (6 (15.4%)).

In the short stories genre in both languages a large group of interruptions takes the form of adjuncts/PPs (English: 15 (40.5%); Dutch: 9 (64.3%)), with appositions forming a large group in English (17 (45.9%)), but a smaller group in Dutch (4 (28.6%)).

Finally, in the leaflets genre most interruptions in both languages are realised as appositions (English: 30 (75.0%); Dutch: 24 (75.0%)), with both languages showing some instances of adjunct/PPs, English of conjuncts (3) and Dutch of premodifiers (4).

**Table 25**                    **Frequencies of most frequent phrasal realisations of postfinite interruptions in C1C, AC1C, C1CX, CX1X, XC1CX, XCX1X subpatterns**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Apposition	31 (31.0%)	37 (54.4%)	25 (39.7%)	31 (54.4%)	124 (43.1%)
Adjunct/PP	33 (33.0%)	16 (23.5%)	18 (28.6%)	13 (22.8%)	80 (27.8%)
Conjunct	19 (19.0%)	4 (5.9%)	8 (12.7%)	5 (8.8%)	36 (12.5%)
Premodifier	4 (4.1%)	1 (1.5%)	0 (0.0%)	1 (1.8%)	6 (2.1%)
Other phrases	13 (13.0%)	10 (14.7%)	12 (19.0%)	7 (12.3%)	42 (14.6%)
<b>Total</b>	<b>100 (100%)</b>	<b>69 (100%)</b>	<b>63 (100%)</b>	<b>57 (100%)</b>	<b>288 (100%)</b>
<b>Dutch</b>					
Apposition	58 (50.0%)	48 (63.2%)	10 (23.3%)	23 (40.4%)	139 (47.6%)
Adjunct/PP	30 (25.9%)	14 (18.4%)	30 (69.8%)	14 (24.6%)	88 (30.1%)
Conjunct	1 (0.9%)	2 (2.6%)	2 (4.7%)	1 (1.8%)	6 (2.1%)
Premodifier	20 (17.2%)	9 (11.8%)	0 (0.0%)	16 (28.1%)	45 (15.4%)
Other phrases	7 (6.0%)	3 (3.9%)	1 (2.3%)	3 (5.3%)	14 (4.8%)
<b>Total</b>	<b>116 (100%)</b>	<b>76 (100%)</b>	<b>43 (100%)</b>	<b>57 (100%)</b>	<b>292 (100%)</b>

In both languages, the largest group of postfinite interruptions takes the form of appositions (English: 124 (43.1%); Dutch: 139 (47.6%)), followed by adjuncts/PPs (English: 80 (27.8%); Dutch: 88 (30.1%)).

In the English academic prose genre, most interruptions take the form of adjuncts/PPs (31 (31.0%)) and appositions (33 (33.0%)), followed by conjuncts (19 (19.0%)). In Dutch, half of the interruptions take the form of an apposition (58 (50.0%)), followed by adjuncts/PPs (30 (25.9%)), and the third large group is formed by premodifiers (20 (17.2%)).

In the English newspaper genre, over half of the interruptions take the form of an apposition (37 (54.4%)), followed by adjuncts/PPs (16 (23.5%)). Conjuncts only form a small group (4 (5.9%)). In Dutch, the group of appositions is

even larger (48 (63.2%)), followed by adjuncts/PPs (14 (18.4%)) and premodifiers forming again a substantial group (9 (11.8%)).

In the English short stories genre, the appositions form the largest group (25 (39.7%)), followed by adjuncts/PPs (18 (28.6%)), with conjuncts forming a third large group (8 (12.7%)). In Dutch the adjuncts/PPs form by far the largest group (30 (69.8%)), followed by appositions (10 (23.3%)).

Finally, in the English leaflets genre over half of the interruptions take the form of an apposition (31 (54.4%)), followed by adjuncts/PPs (13 (22.8%)) and with conjuncts again forming a third group (5 (8.8%)). In Dutch appositions form a large group (23 (40.4%)), followed by premodifiers (16 (28.1%)) and thirdly by adjuncts/PPs (14 (24.6%)).

Although prefinite and postfinite phrasal interruptions are typically realised as appositions, adjuncts/PPs and conjuncts in English and as appositions, adjuncts/PPs and premodifiers in Dutch, there is a difference in distribution of these categories, in that the number of interruptions that take the form of an apposition is higher for prefinite interruptions than postfinite interruptions. This applies particularly to the newspaper genre and leaflets genre in both languages. With the postfinite interruptions, the group of adjuncts/PPs is larger when compared to the prefinite phrasal interruptions.

### **Most frequent grammatical forms of clausal interruptions**

In both languages, the largest clausal realisation groups are formed by finite adverbial clauses, non-finite adverbial clauses and non-restrictive relative clauses. A fourth large group is formed by 'other' clauses, but as these do not consist of one particular type, the focus will be on the three most frequent realisation groups (see sections above for more refined analysis of clausal interruptions that have here been categorised as 'other'). It should be noted that in English this fourth group consists to a considerable extent of independent clauses. Table 26 presents the frequencies of the various types of clausal realisation of prefinite interruptions and Table 27 presents the frequencies of various types of clausal realisation of postfinite interruptions.

**Table 26**                    **Frequencies of most frequent clausal realisations of prefinite interruptions in C1C, AC1C, C1CX, CX1X, XC1CX, XCX1X subpatterns**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Adverbial clause finite	5 (13.9%)	4 (9.1%)	8 (25.0%)	1 (6.7%)	18 (14.2%)
Adverbial clause non-finite	11 (30.6%)	17 (38.6%)	10 (31.2%)	8 (53.3%)	46 (36.2%)
Non-restrictive relative clause	10 (27.8%)	20 (45.5%)	6 (18.8%)	2 (13.3%)	38 (29.9%)
Other clauses	10 (27.8%)	3 (6.8%)	8 (25.0%)	4 (26.7%)	25 (19.7%)
<b>Total</b>	<b>36 (100%)</b>	<b>44 (100%)</b>	<b>32 (100%)</b>	<b>15 (100%)</b>	<b>127 (100%)</b>
<b>Dutch</b>					
Adverbial clause finite	5 (17.9%)	3 (13.6%)	6 (35.3%)	0 (0.0%)	14 (19.2%)
Adverbial clause non-finite	6 (21.4%)	4 (18.2%)	5 (29.4%)	0 (0.0%)	15 (20.5%)
Non-restrictive relative clause	10 (35.7%)	13 (59.1%)	4 (23.5%)	5 (83.3%)	32 (43.8%)
Other clauses	7 (25.0%)	2 (9.1%)	2 (11.8%)	1 (16.7%)	12 (16.4%)
<b>Total</b>	<b>28 (100%)</b>	<b>22 (100%)</b>	<b>17 (100%)</b>	<b>6 (100%)</b>	<b>73 (100%)</b>

In English, the largest group of prefinite clausal interruptions take the form of non-finite adverbial clauses (46 (36.2%)), followed by non-restrictive relative clauses (38 (29.9%)) and last by finite adverbial clauses (18 (14.2%)). In Dutch the largest group is formed by non-restrictive relative clauses (32 (43.8%)), followed by non-finite clauses (15 (20.5%)) and last by finite adverbial clauses (14 (19.2%)).

In the English academic prose genre, the non-finite adverbial clauses form the largest group (11 of 36), followed by non-restrictive relative clauses (10) and finite-adverbial clauses (5). In Dutch the largest group is formed by non-restrictive relative clauses (10 of 28), followed by non-finite clauses (6) and finite adverbial clauses (5).

In both the English and Dutch newspaper genre, most prefinite clausal interruptions take the form of non-restrictive relative clauses (English: 20 of 44; Dutch: 13 of 21). In both languages the second largest group is formed by non-finite adverbial clauses (English: 17; Dutch: 4), and the third by finite adverbial clauses (English: 4; Dutch: 3).

In the English short stories genre the largest realisation group is formed by non-finite adverbial clauses (English: 10 of 32), closely followed by finite adverbial clauses (8) and clauses classified as other (8). In Dutch, the largest group is formed by finite adverbial clauses (6), followed by non-finite ones (5).

Finally, the leaflets genres in both languages contains few occurrences, most of which are realised as non-finite adverbial clauses in English (8 of 15). In Dutch 5 of 6 clauses take the form of a non-restrictive relative clause.

**Table 27**                    **Frequencies of most frequent clausal realisations of postfinite interruptions in C1C, AC1C, C1CX, CX1X, XC1CX, XCX1X subpatterns**

<b>English</b>	<b>Academic prose</b>	<b>Newspaper articles</b>	<b>Short stories</b>	<b>Leaflets</b>	<b>Total</b>
Adverbial clause finite	7 (24.1%)	7 (28.0%)	7 (24.1%)	6 (46.2%)	27 (28.1%)
Adverbial clause non-finite	12 (41.4%)	4 (16.0%)	9 (31.0%)	3 (23.1%)	28 (29.2%)
Non-restrictive relative clause	5 (17.2%)	3 (12.0%)	2 (6.9%)	0 (0.0%)	10 (10.4%)
Other clauses	5 (17.2%)	11 (44.0%)	11 (37.9%)	4 (30.8%)	31 (32.3%)
<b>Total</b>	<b>29 (100%)</b>	<b>25 (100%)</b>	<b>29 (100%)</b>	<b>13 (100%)</b>	<b>95 (100%)</b>
<b>Dutch</b>					
Adverbial clause finite	15 (39.5%)	5 (26.3%)	7 (21.2%)	1 (14.3%)	28 (28.9%)
Adverbial clause non-finite	11 (28.9%)	4 (21.1%)	3 (9.1%)	3 (42.9%)	21 (21.6%)
Non-restrictive relative clause	8 (21.1%)	8 (42.1%)	7 (21.2%)	1 (14.3%)	24 (24.7%)
Other clauses	4 (10.5%)	2 (10.5%)	16 (48.5%)	2 (28.6%)	24 (24.7%)
<b>Total</b>	<b>38 (100%)</b>	<b>19 (100%)</b>	<b>33 (100%)</b>	<b>7 (100%)</b>	<b>97 (100%)</b>

Of the three most frequent clausal forms, the non-finite clauses are most frequent in English (28 (29.2%)), closely followed by finite adverbial clauses (27 (28.1%)) and lastly by non-restrictive clauses (10 (10.4%)). In Dutch, the largest group is formed by finite adverbial clauses (28 (28.9%)), followed by non-restrictive clauses (24 (24.7%)) and non-finite adverbial clauses (21 (21.6%)).

In the English academic prose genre the largest group is formed by non-finite clauses (12 of 29), followed by finite adverbial clauses (7) and non-restrictive clauses (5). In Dutch the largest group is formed by finite adverbial clauses (15 of 38), followed by non-finite clauses (11) and non-restrictive clauses (8).

In the English newspaper genre the largest group is formed by finite adverbial clauses (7 of 24), followed by non-finite (4) and non-restrictive (3). In Dutch the largest group is formed by non-restrictive clauses (8 of 19), followed by finite adverbial clauses (5) and non-finite (4).

In the English short stories genre the non-finite clauses form the largest group (9 of 29), followed by finite adverbial clauses (7). In Dutch the largest groups are formed by non-restrictive clauses and finite adverbial clauses (both 7 of 33).

Finally, the leaflets genres in both languages have very few occurrences, with the finite adverbial clauses forming the largest group in English (6 of 13) and non-finite adverbial clauses in Dutch (3 of 7).

As for the semantic roles of the finite adverbial clauses, in both languages the largest group is formed by clauses introduced by *as* and *zoals* respectively (English: 15 of 45; Dutch: 17 of 39). Also in both languages, the second largest group is formed by adverbial clauses of time (including a few instances of place) (English: 12; Dutch: 13). Condition forms the third largest group, also in both languages (English: 8; Dutch: 6). English also contains 7 clausal interruptions that take the form of an adverbial clause of concession.

The non-finite clauses can contain a verb that takes the form of a present participle, a past participle, an infinitive or it can be a verbless clause. In English, 32 (43.2%) of the 74 non-finite clauses contain a present participle; 37 (50%) contain a past participle; 2 sentences (2.7%) an infinitive and 3 (4.0%) are verbless. In Dutch, on the other hand, only 5 of 36 (13.9%) sentences contain a present participle and the vast majority contain a past participle (22 (61.1%). Only 1 sentence contains an infinitive (2.8%) and 8 sentences are verbless (22.2%).

## 8.9 Conclusion

An analysis of the interpolated satellites in the various genres of English and Dutch shows a number of striking similarities and a number of notable differences between the languages. When looking at the frequencies of the interruptions, English contains significantly more interruptions than Dutch across all genres. This applies to both sentences that contain one interruption and sentences that contain two or more interruptions. What is similar in both languages is that the academic prose genre shows the highest frequency of sentences that contain interruptions and the short stories genre the genre that shows the lowest frequency of sentences with interruptions.

The occurrence of interruptions also varies per sentence pattern, with most sentences in both languages belonging to the C-pattern. Furthermore, even though there is a wide range of patterns that can be created by various combinations of interruptions, the most frequently occurring subpatterns are the same six patterns in both languages. These are all patterns in which the sentence as a whole contains one interruption.

The languages show differences in the frequencies of the two main positions of the interruptions in the sentence: those occurring before the finite

verb of the discourse unit it interrupts and those occurring after the finite verb. English shows a higher frequency of interruptions that precede the finite verb, especially in the XC pattern and XCX pattern, whereas Dutch shows a higher frequency of postfinite interruptions, particularly in these patterns. In the other main sentence patterns, the C-pattern and the CX pattern, the frequencies are more evenly distributed for English and Dutch. The exceptional occurrence of prefinite interruptions in the XC and XCX pattern in Dutch can be explained by the fact that Dutch is a verb-second language.

With respect to the grammatical realisation of the interruptions, the languages show similarities in that both prefinite and postfinite interruptions are more often realised as phrases than clauses. In both languages the percentage of phrases is higher with postfinite interruptions than prefinite interruptions, and the phrasal interruptions typically take the form of one of three main grammatical categories. For English these are appositions, adjuncts/PPs and conjuncts, also in this order, and for Dutch these are appositions, adjuncts/PPs and premodifiers, also in this order. This means that the languages mainly show differences with respect to the third realisation group. A difference between English and Dutch is that conjuncts are typically presented as separate punctuation units in English, both when they occur sentence-initially and when they occur sentence-medially. In Dutch there is more variation in the use of punctuation for sentence medial conjuncts and those that are not presented as separate punctuation units have not been considered as separate discourse units in this study (see 2.4.3). Another difference is the relatively high frequency of premodifiers in Dutch. This is a typical use of Dutch and hardly occurs in English (cf. 9.2.4 for further explanation of this category in Dutch). Moreover, the grammatical realisation of interruptions is, to a certain extent, also dependent of the position of the interruption, with prefinite interruptions showing a higher frequency of appositions than postfinite interruptions, which, in turn, show a more equal frequency of appositions and adjuncts/PPs. The high frequency of prefinite interruptions realised as appositions can be explained by the fact that these interruptions typically follow the subject of the sentence and the appositions form a way of providing additional information about the subject, a function that is used particularly in newspapers (see below). Moreover, the clausal interruptions also typically take the form of one of three grammatical forms: finite adverbial clauses, non-finite adverbial clauses and non-restrictive relative clauses. In the case of prefinite interruptions, in English, the non-finite clauses show the highest frequency, followed by non-restrictive clauses and finite adverbial clauses, whereas in Dutch the non-restrictive show the highest



frequencies. Dutch shows a surprisingly high number of non-finite clauses, which typically contain a past participle (cf. Chapter 5 on significantly higher frequency of non-finite clauses in English than Dutch). In English about half of the non-finite clauses contain a present participle, with the other half containing a past participle. Furthermore, in both languages most finite adverbial clauses are introduced by *zoals/as*, with adverbial clauses of time forming the second most frequent group and adverbial clauses of condition the third most frequent group.

Finally, differences between the genres cannot only be found in the frequencies of the interruptions, but also in their grammatical realisation. In the English academic prose genre, most phrasal interruptions take the form of appositions, but this is closely followed by adjuncts/PPs and conjuncts. Dutch shows a similar pattern, with the exception of the third group, which is formed by premodifiers. In both languages the newspaper genre shows a particularly high frequency of appositions on the one hand and non-restrictive relative clauses on the other hand. In this genre both grammatical categories appear to have the function of providing additional information about the subject of the sentence, which can be a person or a particular situation. Furthermore, the short stories genre in both languages contains hardly any clauses, and in English most interruptions take the form of an apposition, whereas in Dutch this is an adjunct/PP. In addition, Dutch also shows a high instance of a particular subpattern, the C1X pattern, in which the interruption does not actually interrupt another discourse unit, but occurs in between the nucleus and the appended satellite. The interruption typically takes the form of a reporting clause that 'interrupts' reported speech (a dialogue). Finally, the leaflets genre in both languages again shows a low frequency of clauses and a high frequency of appositions. In Dutch, this genre also shows a considerable instance of premodifiers.

## 9. Discussion

### 9.1 Introduction

The aim of the present study is to establish what the main sentencing patterns are in English and Dutch and to what extent these patterns can be related to the particular linguistic systems of these languages, to their writing cultures, or to an interaction between linguistic system and writing culture. Another aim is to try to establish to what extent genre influences sentence structure. Chapters 2, 3 and 4 described the design and theoretical motivation of the corpus compilation and annotation and Chapters 5, 6, 7 and 8 presented the main results of the sentencing analysis

Whereas the results chapters focused on presenting the mere numbers of the various sentencing patterns, the present chapter will make an attempt to provide more insight into these numbers by taking the linguistic systems of English and Dutch into consideration and the genres to which the sentences belong.

Because the analysis of sentences occurred at a very detailed level, in that both their discourse and grammatical structure were taken into consideration, the results generated were also very detailed. When aggregating the main findings of the results chapters, it does, however, become clear that differences between the languages with respect to sentencing structure can be described on the basis of the following five parameters: 1) sentence length, 2) main sentence patterns, 3) beginning of sentences, 4) interruptions, and 5) end of sentences.

The first part of this chapter will focus on differences, as well as striking similarities, between English and Dutch sentencing structure with regard to these five parameters irrespective of genre. The second part will focus on differences between the languages with regard to these parameters at the level of the individual genres, i.e. the academic prose genre, the newspaper genre, the short stories genre and the public information leaflets genre.

## 9.2 Main findings - languages

The present section will focus on those results for which the statistical analyses showed that the differences found between the languages with respect to a particular aspect of the sentencing analysis were significant irrespective of the genre to which the sentence belongs. An attempt will be made to relate these differences to the respective language systems of English and Dutch and the differences between them.

### 9.2.1 Sentence length

Although differences in sentence length between the languages are to be interpreted with some caution due to the influence of differences in spelling conventions on sentence length, an analysis has shown that English sentences are significantly longer in three of the four genres than Dutch sentences, with the short stories genre being the only exception (see 5.2.2). Even though this exception shows that sentence length is at least to a certain extent genre-dependent (see 9.3 below), the significant difference in length between the languages in the case of three of the four genres under analysis is noteworthy (see Hannay 1997 and Cosme 2007 for similar findings concerning the Dutch and English newspaper genre).

With respect to the distribution across the four sentence length categories (i.e. 1-10, 11-20, 21-30 and 31+ word sentences), the analysis showed that English contains significantly more sentences in the 31+ length category than Dutch in the three genres for which differences were found, whereas Dutch showed higher frequencies in the 1-10 words and the 11-20 words categories.

An explanation for the rather large number of relatively short sentences in especially Dutch, i.e. sentences that do not exceed 20 words, could be sought in the explicit advice given in a variety of Dutch style guides to restrict sentence length by avoiding the use of particular constructions that make sentences more complex. Examples of such constructions are those in which additional information is placed between two elements that belong together, such as a subject and a verb, thereby interrupting the flow of the sentence (*tangconstructies*) (cf. Doeve & Onrust: 1992: 90ff; van der Horst 1999: 38ff; Renkema 2005: 82-83; Tiggeler 2006: 191; Burger & de Jong 2009: 133-135). Another construction that is typically advised against involves sentences with a long start (*lange aanloop*), in which the main clause is either preceded by a very long phrase or one or more subordinate clauses (cf.

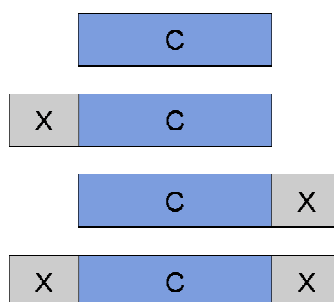
Renkema 2005: 85-86; Hermans 2006: 55-57; Tiggeler 2006: 192; Burger & de Jong 2009: 136). As the suggestion is often made to put this information into two shorter sentences, this might then be linked to the higher frequency of rather short sentences in Dutch. Furthermore, some style guides in Dutch, especially the more genre-specific ones, make reference to the readability tests as designed by Rudolf Flesh (1949), who developed a formula to determine the readability of texts based on the average length of sentences and words. Despite the fact that this formula, and the Dutch Flesh-Douma variant, is typically presented as being far too simplistic in nature, it is interesting to note that the formula does still receive some attention, even to the extent of a reproduction of the original table that contains an overview of various sentence lengths and their associated degrees of complexity (eg. Lamers 1986: 124-131; Donkers & Willems 2002: 184-186; Burger & de Jong 2009: 140-144).

For English, explicit advice on sentence length is typically not provided in general style guides (eg. Sinclair 1992; Ritter 2002; Peters 2004; Hicks 2009), excluding exceptions that suggest that 'if you discover that [your longest sentence] is something like 118 words, do some rewriting' (Trask 2002: 208). Whether the lack of this explicitness could serve as a potential explanation for the longer sentences in English when compared to Dutch is an interesting hypothesis that needs to be studied in more detail.

In line with the relationship that is postulated between long sentences and sentence complexity in many Dutch style guides (e.g. Hermans 2002: 49; Tiggeler 2006: 193), sentence length does indeed constitute a commonly used method – often supplemented by various other ones (cf. Cosme 2007: 205) – to measure complexity. This difference in sentence length between English and Dutch in three of the four genres could, in that respect, be an indication of a difference in sentence complexity. In an earlier small-scale contrastive corpus analysis of English and Dutch sentencing patterns, Hannay has also related sentence length to syntactic complexity and has characterised Dutch as having a 'chopping style', by showing, for instance, a preference for shorter sentences and sentence fragments, and English as having a 'combining style', by showing a preference for longer sentences and certain clause-combining devices (1997: 243, 249). A more recent contrastive analysis of syntactic complexity in Dutch, English and French has been carried out by Cosme (2007), who also characterises Dutch as being the least syntactically complex language of the three languages, based on an analysis of sentence length and the extent and type of subordination and coordination used (2007: 296ff).

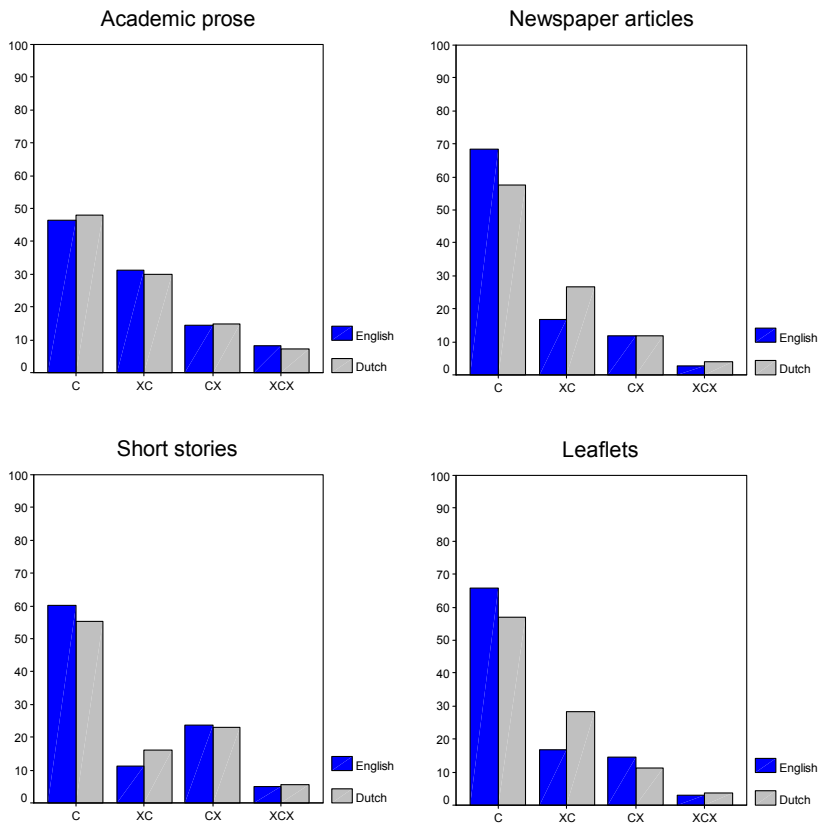
## 9.2.2 Main sentence patterns

In order to find the most frequent sentence patterns in English and Dutch, all sentences in the corpus were analysed at both the levels of discourse and grammar (cf. Chapters 2 and 3). The discourse analysis has shown that all sentences in both languages can be categorised into one of four main patterns. These are sentences that consist of only a nuclear unit, (C); sentences that consist of a nuclear unit and one or more prepended satellites (XC); sentences that consist of a nuclear unit and one or more appended satellites (CX); and sentences that consist of a nuclear unit, one or more prepended satellites and one or more appended satellites (XCX) (see Figure 1 below).



**Figure 1** Main sentence patterns

The analyses have shown that, on the one hand, the languages show considerable similarities in the sense that all sentence patterns in both languages can be reduced to these four broad categories, with the least complex pattern in terms of discourse structure, the C-pattern, showing the highest frequencies and the most complex pattern, the XCX pattern, showing the lowest frequencies in both languages. On the other hand, the analyses have also shown that a comparison between the languages cannot be made without taking genre into consideration, as the genre to which a sentence belongs influences its structure to a considerable extent. Specifically, with the academic prose genre forming the exception, the newspaper, short stories and leaflets genres all show significant differences between the languages with respect to the frequencies of these four main patterns. However, despite the fact that the exact frequencies are thus genre-dependent, the overall picture that emerges is that in the three genres that show differences between the languages, the C-pattern shows relatively high frequencies in English and the XC pattern shows relatively high frequencies in Dutch. The frequencies of these main patterns are presented in Figure 2 below.



**Figure 2 Main sentence patterns in English and Dutch in four genres**

Although the sentences can and have been categorised at this broad level of analysis, a more refined analysis has shown that all sentence patterns can be further broken down into a range of subpatterns, the most frequent of which are the AC, the CD and the ACD patterns in both languages (excluding the patterns formed by interpolated satellites, see 9.2.4 below). This more refined analysis has, for instance, also shown that despite the fact that the XC pattern on the whole occurs more frequently in Dutch than in English, the subpatterns in which the C is preceded by two or more elements occur significantly more frequently in English than in Dutch (see 9.2.3 below). In cases such as these, a closer inspection of the subpatterns, in combination with an analysis of the grammatical realisation of the satellites, might contribute to an explanation of frequency differences for the main patterns between the languages. In other words, although a rough analysis of

sentence patterns at the level of discourse structure proves useful in that all sentences can be categorised into and can thus be reduced to four main patterns, it does not provide sufficient insight into the differences between the languages, as these emerge when a more refined analysis is performed and when genre is taken into account as well. Sections 9.2.3, 9.2.4 and 9.2.5 below will therefore focus on the main findings of these more refined analyses.

### 9.2.3 Beginnings of sentences

In both languages the nuclear unit, the C, can be preceded by up to three prepended satellites, forming a range of subpatterns. By far the most frequent subpattern in both languages is the one in which the nucleus is preceded by one satellite, the AC pattern, followed by the far less frequent ACD subpattern. When preceded by two satellites, A1C(X) and ABC(X) are the most frequently occurring subpatterns (see 6.2 for the infrequent A1AC(X) subpattern). Despite the fact that the XC main pattern on the whole occurs more frequently in three of the four genres in Dutch, with the exception of the academic prose genre, a more refined analysis of the subpatterns of the XC pattern has shown that particularly the ABC(X) subpattern occurs significantly more frequently in English across all genres. Furthermore, the patterns in which the nucleus is preceded by three elements are rare in both languages, but show a higher frequency for English than for Dutch.

The result that differences between the languages arise when looking at the number of elements that occur in sentence-initial position is in itself not surprising, as this can be explained by structural differences between English and Dutch. Specifically, Dutch, unlike English, is a verb-second language, which means that the finite verb is typically placed in second position and that no more than one element is thus allowed in sentence-initial position (Haeseryn et al. 1997: 1261; Smits 2002: 22). It should, however, be noted that this characteristic of Dutch does not merely involve a limit on the number of elements that precede the finite verb, but also the type of elements and the hierarchical relation between them. As the interpolated satellite in the A1C(X) pattern actually just functions as in interruption of the A-satellite, but then one that is positioned at the very end of the satellite instead of literally interrupting it like in the A1AC(X) pattern (cf. 6.4), the combination A1 can be seen as constituting one complex A-satellite instead of two elements (see Haeseryn et al. 1997: 1297 for a similar analysis). It is for this reason not surprising that Dutch easily allows for this sentence pattern (cf. see 6.2). What is more remarkable is that the corpus contains Dutch sentences that follow the

ABC(X) pattern, in which two preposed satellites that are of the same hierarchical rank (cf. 2.5.3) precede the nucleus. That Dutch does not categorically reject the ABC(X) pattern despite its V2 character has, however, already been noted by Haeseryn et al. (1997: 1297ff) and has been further investigated by Smits (2002), who, in her study on complex beginnings in native and learner English, found that Dutch allows for the same adverbial clusters in sentence-initial position as English does, albeit with a lower frequency (p. 168-169). Although they are rare in Dutch, especially in comparison with English, it is interesting to note that the frequency of occurrence differs per genre, which would mean that the occurrence of the ABC(X) pattern is not solely dependent on the language system, but also the genre or text type within the language system. More specifically, in both languages the academic prose genre shows the highest frequency of this pattern in comparison with other genres (cf. 9.3 below). Consider sentence (1), which is taken from the Dutch academic prose genre and follows the ABC pattern. Note that although Dutch *echter* (*however*) can occur in sentence-initial position, Haeseryn et al. explain that it more typically combines with another sentence-initial element which it then follows (1997: 1278, 1297) and that the use as in (1) below is seen by some as constituting an anglicism and therefore not considered correct (1997: 1394).

- (1) <A><conj>Echter<conj><A>, <B><advcl\_cond>als er vaak en veel woede en frustratie bij die conflicten komt kijken<advcl\_cond><B>, <C>heeft dit negatieve effecten voor kinderen<C>. <s5966, academic prose>

(<A><conj>However<conj><A>, <B><advcl\_cond>if there often and much anger and frustration involved is in these conflicts is<advcl\_cond><B>, <C>has this negative effects on the children<C>.)

(<A><conj>However<conj><A>, <B><advcl\_cond>if there is often and much anger and frustration involved in these conflicts<advcl\_cond><B>, <C> this has negative effects on the children<C>.)

Precisely because it is perceived by some as an anglicism, this sentence-initial placement of *echter* – and other related conjuncts that behave similarly in Dutch, such as *immers* (after all) (cf. Haeseryn et al. 1997: 1278, and 9.3 below) – could be interpreted as a reflection of the dominance of English in the academic prose genre and the associated behaviour of the English equivalent of *echter*, i.e. *however*, in this genre, which is not only highly frequent, but also typically occurs in sentence-initial position in the academic prose genre (cf. Biber et al. 1999: 886-887). It could



thus be suggested that at least the academic prose genre in Dutch is to a certain extent affected by English with respect to Dutch conjunct placement.

With respect to the grammatical realisation of the A-satellites and the B-satellites, a close analysis has shown that the languages show similarities with respect to the realisation of the A-satellite, which is typically realised as a conjunct in both languages, but show more variation with respect to the realisation of the B-satellite. In Dutch this typically takes the form of an adjunct/PP, whereas English also shows a considerable number of adverbial clauses of various types in this position. This is similar to the grammatical realisation of the A-satellite in the sentence patterns in which the nucleus is preceded by only one element (see below). Furthermore, in analysing the small number of Dutch ABC(X) sentences, it is also interesting to take note of the type of punctuation mark used to separate the various discourse units, as Dutch, unlike English, uses the colon in a few instances to separate the A-satellites that takes the form of a conjunct from the other units, as in sentence (2) below. Note that the second element to precede the subject has been classified as a so-called zz-element instead of a B-satellite, as it does not constitute a separate punctuation unit (cf. 2.4.3 & 5.5):

(2) <A><advcl\_nonfin>Om te beginnen<advcl\_nonfin><A>: <zz><pp>met een gezond gewicht<pp><zz> <Ca><coord\_a\_asyn>voel je je fitter<coord\_a\_asyn><Ca>, <Cb><coord\_b\_asyn>je conditie is beter<coord\_b\_asyn><Cb> <Cc><coord\_c>en je beweegt soepel<coord\_c><Cc>. <s9432, leaflets>

(<A><advcl\_nonfin>To start<advcl\_nonfin><A>: <zz><pp>with a healthy weight<pp><zz> <Ca><coord\_a\_asyn>feel you yourself more fit<coord\_a\_asyn><Ca>, <Cb><coord\_b\_asyn>you are in better shape<coord\_b\_asyn><Cb> <Cc><coord\_c>and you move more flexibly<coord\_c><Cc>.)

(<A><advcl\_nonfin>To start<advcl\_nonfin><A>: <zz><pp>with a healthy weight<pp><zz> <Ca><coord\_a\_asyn>you feel more fit<coord\_a\_asyn><Ca>, <Cb><coord\_b\_asyn>you are in better shape<coord\_b\_asyn><Cb> <Cc><coord\_c>and you move more flexibly<coord\_c><Cc>.)

In the hierarchy of punctuation marks (cf. 2.4.1), colons are typically classified as reflecting a sharper separation in comparison with the comma and, as Smits suggests, the choice for this punctuation mark might be 'resorted to' in Dutch to adhere to the verb-second principle (2002: 171). Examples such as sentence (2)

illustrate how the interplay between not only discourse units and grammatical realisation, but also the type of punctuation mark used can shed more light on sentence structure and on how sentence structures that are generally considered exceptional are presented in such a way as to make them more conventional.

With respect to the grammatical realisation of the satellites of the more frequent A1C(X) subpattern, the A-satellite takes the form of an adjunct or PP in the vast majority of cases in both languages, which is similar to the realisation of A in the AC and ACD patterns. It is mainly with respect to the realisation of the interpolated satellite that the languages show differences. In English this is predominantly realised as a non-finite adverbial clause, followed by adjuncts/PPs and conjunct/disjuncts, whereas in Dutch it is predominantly realised as either an adjunct/PP or an apposition.

To return to the far more frequent subpatterns in which the nucleus is preceded by one element, the AC and ACD patterns, the analysis has shown that in both languages this A/zz-satellite is realised as a phrase in the vast majority of cases, typically taking the form of an adjunct or prepositional phrase, especially in Dutch. In both languages the second largest realisation group is formed by conjuncts, although it should be noted that considerable and significant differences exist between the different genres between the languages (cf. 9.3 below). Furthermore, despite the fact that clausal A-satellites occur much less frequently than phrasal As, the clausal As are significantly higher in number in both the AC and ACD subpatterns in English in three of the four genres, with the short stories genre forming the exception in the AC subpattern. In both languages these clausal As are mainly realised as finite adverbial clauses, with English showing particularly high frequencies of adverbial clauses of concession and Dutch showing high frequencies of adverbial clauses of condition. The group of non-finite clauses is significantly larger in English, mainly containing present participles, whereas the few instances in Dutch contain a past participle or infinitive. This frequency difference between the languages with respect to non-finite clauses is in line with and thus confirms earlier claims in the contrastive literature on finiteness/non-finiteness in English and Dutch (cf. Aarts & Wekker 1987: 301; De Moor 1998: 309; Hannay & Mackenzie 2009: 93-96), claims that have recently been supported by quantitative findings of Cosme (2007: 279-280). Furthermore, a possible explanation for the significantly higher frequency of clauses in this position in English, and also in the ABC(X) and A1C(X) patterns discussed above, could be linked to the advice given in a large number of Dutch style guides against the use of sentences with a long orientational section (*lange aanloop*) (see 9.2.1 above) (cf. Renkema 2005: 85-86;

Hermans 2006: 55-57; Tiggeler 2006: 192; Burger & de Jong 2009: 136). A long start is typically described as a sentence that starts with a particularly long phrase or one or more subordinate clauses and is discouraged because it has the effect of postponing the nuclear information, typically placed in the independent clause, thereby reducing the readability of the sentence. This would then be a reflection of a difference in writing cultures and not necessarily in the linguistic systems.

In sum, an analysis of sentence beginnings indicates that both languages, but particularly Dutch, show a high frequency of sentences that start with an adjunct or prepositional phrase. English, on the other hand, shows a significantly higher frequency of clauses in initial position, especially non-finite adverbial clauses. Differences between the languages can be found when the nucleus is preceded by more than one satellite, with English showing a significantly higher frequency of the ABC(X) subpattern than Dutch. In both languages, however, the frequency of particular subpatterns and the grammatical realisation of the various elements is influenced by the genre in which the sentence occurs.

#### 9.2.4 Interruptions

In addition to the sentence patterns that can be created by various combinations of prepended and appended satellites, a wide range of sentence patterns are also formed by various combinations of interpolated satellites (cf. Chapter 8). Although these discourse units can interrupt both the nucleus and the prepended and appended satellites (cf. 2.5.3), they typically occur in the nucleus in both languages. On the whole, English contains significantly more interruptions than Dutch across all genres. This applies both to sentences that contain one interruption and to sentences that contain two or more interruptions. With respect to the main sentence patterns formed by interruptions, the languages again show much overlap, with the most frequent ones being the C1C, the AC1C, the C1CX and CX1X, and the XC1CX and XCX1X subpatterns (see Chapter 8 for a full overview). Although the C main pattern shows the highest absolute frequency of sentences with one and two interruptions, the XCX main pattern shows the highest relative frequency of interruptions. This might indicate that the occurrence of interpolated satellites in a sentence is associated with a more complex sentence structure on the whole.

The association between interruptive structures and sentence complexity has also been made by a wide variety of Dutch style guides in their discussion of the so-called *tangconstructie* (see 9.2.1 above) (cf. Doeve & Onrust: 1992: 90ff; van der Horst 1999: 38ff; Renkema 2005: 82-83; Tiggeler 2006: 191; Burger & de Jong

2009: 133-135). Although they come in different forms, a characteristic of this construction is that information is placed in between two elements that belong together, such as a subject and a verb. This can have the effect of decreasing the readability and interrupting the flow of the sentence. Although the present study has only looked at a subset of the *tangconstructies*, namely those that are presented as a separate punctuation units, i.e. here classified as interpolated satellites, the fact that they are so explicitly dispreferred in these style guides could be related to the significantly lower frequency of interpolated satellites in Dutch when compared to English.

With respect to the position of the interruption in relation to the finite verb of the discourse unit it occurs in, English shows a fairly even distribution of interruptions that either precede or follow the finite verb, whereas in Dutch this is dependent on the sentence pattern in which the interruption occurs. More specifically, the Dutch XC and XCX main patterns show hardly any instances of prefinite interruptions. This can be explained by the verb-second criterion in Dutch (see 9.2.3 above), which has the effect of not easily allowing for more than one satellite in sentence-initial position. The exceptional cases in which a subject is followed by an interruption and then by the finite verb are, again, perceived as anglicisms by some and therefore not generally accepted (Haeseryn et al. 1997: 1300). Another effect of the V2 criterion in combination with the XC and XCX sentence patterns is that because the subjects in these sentence patterns are placed in postfinite position, those interruptions that are realised as appositions, which applies to the vast majority (see below), are automatically also placed in postfinite position, typically immediately following the subject. Sentence (3) below presents an example of the typical position of interruptions in Dutch in the XC(X) patterns, i.e. after the finite verb, and sentence (4) is one of a total of 3 sentences in Dutch in which the interruption precedes the finite verb. What is interesting to note is that in two of the three cases the A-satellite is realised as a conjunct and in one case the A-satellite is marked off by means of a colon, also indicating a more separated or detached status of the sentence-initial element with respect to the rest of the sentence (see 9.2.3 above). This appears to indicate that the specific function of the sentence-initial element and choice of punctuation mark thus interact in such a way to make a pattern that the grammar of Dutch does not easily allow for, considering its V2 character, acceptable after all.

- (3) <zz><adj>In 1951<adj><zz> <C><indepcl>werd de Nederlandse Televisie Stichting (<1\_postfv><appos\_NP>NTS<appos\_NP><1\_postfv>) opgericht<indepcl><C>. <s4020, academic prose>
- (<zz><adj>In 1951<adj><zz> <C><indepcl>was the Dutch Television Foundation (<1\_postfv><appos\_NP>DTF<appos\_NP><1\_postfv>) established<indepcl><C>.)
- (<zz><adj>In 1951<adj><zz> <C><indepcl> the Dutch Television Foundation was(<1\_postfv><appos\_NP>DTF<appos\_NP><1\_postfv>) established<indepcl><C>.)
- (4) <A><advcl\_nonfin>Of anders gezegd<advcl\_nonfin><A>, <C><indepcl>de buitenlandse politiek - <1\_prefv><appos\_np>en de Vietnamoorlog in het bijzonder<appos\_np><1\_prefv> - bleek uitermate geschikt om van 'een progressieve grondhouding' te getuigen<indepcl><C>. <s3786, academic prose>
- (<A><advcl\_nonfin>Or put differently<advcl\_nonfin><A>, <C><indepcl>foreign politics - <1\_prefv><appos\_np>and the Vietnam War in particular<appos\_np><1\_prefv> - proved to be particularly suitable for demonstrating 'progressive principles'<indepcl><C>.)

With regard to the grammatical form of the interpolated satellites, the languages show overlap in the most frequently occurring grammatical categories. In both languages, the vast majority of interruptions take the form of a phrase, mainly realised as appositions or adjuncts/PPs, with the minority taking the form of a clause, mainly realised as finite adverbial clauses, non-finite adverbial clauses and non-restrictive relative clauses. What is interesting about this latter result is the relatively high frequency of non-finite adverbial clauses in Dutch. As noted in 9.2.3, non-finite clauses have been shown to occur significantly more frequently in English than Dutch in various studies. Although English does indeed show higher frequencies of non-finite clauses in this position, mainly containing a present participle, Dutch shows a considerable number of non-finite adverbial clauses with a past participle. This may indicate that the preference of Dutch for finite adverbial clauses is to a certain extent dependent on the function and position of this clause in the sentence, with certain positions allowing more easily for non-finite clauses than others. This may be motivated by the fact that non-finite clauses typically present a more condensed way of integrating information than finite clauses (cf. Quirk et al. 1985: 995; Biber et al. 1999: 198), making them suitable candidates for the grammatical realisation of interruptions. Sentence (5) below contains an example of a clausal interruption in Dutch, realised as a non-finite clause.

- (5) <C><indepcl>De SEO, <1\_prefv><advcl\_nonfin>gespecialiseerd in econometrisch onderzoek<advcl\_nonfin><1\_prefv>, weet weinig van integratie<indepcl><C>. <s3163, newspaper articles>

(C<indepcl>The SEO, <1\_prefv><advcl\_nonfin>specialised in econometric research<advcl\_nonfin><1\_prefv>, knows little bout integration<indepcl><C>.)

The most frequent phrasal realisation, the apposition, provides an efficient means of incorporating additional information, often about the subject (cf. 9.3 below on frequency of occurrence of apposition and function across the different genres). This function of appositions is exemplified by sentence (6) below.

- (6) <zz><adj>Yesterday<adj><zz> <C><indepcl>the Shadow Chancellor, <1><g\_appos\_NP>Oliver Letwin<appos\_NP><1>, tried the old con trick again<indepcl><C>. <s286, newspaper articles>

With respect to the less frequently occurring phrasal interruptions, differences between English and Dutch can be found for the conjunct, which is more frequent in English, and for a category that has been labelled ‘premodifier’, which is more frequent in Dutch. The difference in frequencies for the conjunct could, on the one hand, be seen as reflecting a mere difference in punctuation practice between the languages, as conjuncts in Dutch are typically not marked off by means of punctuation when occurring sentence-medially and therefore not considered separate discourse units in the present study (cf. 2.3.3 on the definition of discourse units). On the other hand, precisely because conjuncts in English are presented as separate punctuation units, it could at the same time be argued that they therefore also have a different rhetorical effect, by creating a clear break – i.e. an interruption – in the sentence and thereby placing more emphasis on the elements that precede and follow them. As for the premodifier in Dutch, this use of an interruption can be considered typical of this language and comes in two main types. The first involves the method of presenting the prefix of a word between embedded brackets, thereby creating an efficient way of presenting multiple, yet related forms of a word, (cf. Burrough-Boenisch 2004: 15-21. The other type is when the premodification in a noun phrase is presented between brackets (cf. Renkema 2005: 368). Because this often occurs at the phrase level rather than the clause level, and also because brackets are used to introduce this type of interruption, it could be argued that this particular type only has semi-interruption

status, as it does not really interrupt the flow of the sentence. This would mean that a fairly large group of interruptions in Dutch has a less intrusive status than the interruptions in English. Sentence (7) presents an example of a premodifier interruption in Dutch.

- (7) <C><indepcl>Wissel limonade en sap af met kinder- of  
 (<1\_postfv><premod>vruchten<premod><1\_postfv>) thee, water of niet te sterke  
 bouillon<indepcl><C>. <s9078, leaflets>
- <C><indepcl>Alternate lemonade and juice with children's- or  
 (<1\_postfv><premod>fruit<premod><1\_postfv>) tea, water or not too strong broth  
 <indepcl><C>.)

Apart from the overlap in grammatical realisations of interruptions between the languages, the difference could indicate that at least a subset of them perform a different function in the languages and could indeed point to a difference in status of the interruption with respect to the extent to which it disrupts the flow of the sentence. The idea that at least a group of interruptions could be seen as less intrusive in Dutch is further supported by a difference in punctuation practice between the languages. Specifically, in addition to the commas that both languages typically use to separate interruptions from the rest of the sentence, Dutch uses brackets in a considerable number of cases, also for interruptions other than those realised as premodifiers. A closer look at the interruptions does, however, indicate that this difference in function is not solely dependent on language, but also on genre, as Section 9.3 below will show.

### 9.2.5 Ends of sentences

Although the nucleus could, in principle, be followed by an indefinite number of appended satellites, both languages contain very few sentences in which this number exceeds three. In fact, both English and Dutch show a clear preference for patterns in which the nucleus is followed by only one appended satellite, the CD and ACD subpatterns respectively. What is more, the languages even show a remarkably similar frequency for the CX main pattern as a whole. However, differences between the languages arise when not only grammatical realisation of the discourse units, but also the punctuation marks to link the units are taken into consideration.

With respect to the D-satellite, in both languages this is realised as a clause in the majority of cases in all subpatterns, with the D in the CDE subpattern in Dutch forming an exception and more often taking the form of a phrase. When realised as a phrase, the languages show much overlap in the grammatical form of the D, which is most frequently realised as an apposition, although this is to a certain extent genre-dependent, as Section 9.3 below will show. With respect to the clausal Ds, despite considerable differences between the four genres in both languages, on the whole the largest realisation group is formed by adverbial clauses in both languages. However, a closer analysis of the adverbial clauses in this position shows that significant differences between English and Dutch can be found in the distribution of finite and non-finite adverbial clauses, with the latter again being much more frequent in English, predominantly containing a present participle. Furthermore, when looking at the semantic class of adverbial clauses, it becomes clear that, similar to the clausal prepended satellites, English again shows a preference for adverbial clauses of concession, whereas Dutch shows higher instances of adverbial clauses of time and reason.

With respect to the less frequently occurring grammatical categories, a notable difference between English and Dutch can be found in the relatively high frequency of D-satellites in Dutch that are realised as non-restrictive relative clauses and as independent clauses. As for this latter category, Dutch not only shows higher frequencies of the sentence pattern in which two independent clauses are juxtaposed, it also shows higher frequencies of the colon to link these two units – a punctuation mark that shows significantly higher overall frequencies for Dutch, across the genres. Sentence (8) provides an example of a Dutch sentence in which the nucleus and appended satellite, both realised as independent clauses, are linked by means of colon.

(8) <zz><pp>Bij 60-plussers<pp><zz> <C><indepcl>is de situatie erg slecht<indepcl><C>: <D><indepcl>maar een op de vijf heeft een baan<indepcl><D>. <s2143, newspaper articles>

(<zz><pp>With 60-somethings<pp><zz> <C><indepcl>the situation is very bad<indepcl><C>: <D><indepcl>only one in five has a job<indepcl><D>.)

In line with its typical usage (cf. Quirk et al. 1985: 1620; Onrust et al. 1993: 194; Huddleston & Pullum 2002: 1743), the colon indicates an elaborative interpretation in (8), in which the appended satellite further specifies or elaborates on the information contained in the nucleus. In addition to this, the colon also has an



announcement function, drawing attention to the information contained in the unit that follows the punctuation mark (cf. Koenen & Smits 2004: 290; Tiggeler 2006: 158). In his small-scale contrastive corpus analysis of English and Dutch sentencing patterns, Hannay (1997) already identified this higher frequency of the colon in Dutch and labelled the constructions formed by colons ‘colon constructions’ (p. 252). In his view, their higher frequency, together with the use of other devices, is reflective of a style he refers to as a ‘prominence-promoting style’ (1997: 234). This use of the colon as a prominence device is even better exemplified by sentence (9), in which the nucleus is realised as a fragment and the appended satellite again as an independent clause. Although it should be noted that nuclei realised as fragments are rare and that this particular usage of the colon is not restricted to Dutch, it is more frequent in Dutch.

- (9) <C><fragment\_NP>Het gevolg<fragment\_NP><C>: <D><indepcl>vrijwel alle commercial radio- en tv- stations zijn in meerderheid in handen van buitenlandse uitgevers gekomen<indepcl><D>. <s1990, newspaper articles>

(<C><fragment\_NP>The effect<fragment\_NP><C>: <D><indepcl>almost all commercial radio- en tv- stations are for the majority in the hands of foreign publishers<indepcl><D>.)

In English, on the other hand, the colon typically, though not exclusively, performs the function of introducing a list or an apposition (see 7.3.2). Moreover, in those cases in which the English D-satellite is realised as an independent clause, this is more often linked to the preceding nucleus by means of a dash – a punctuation mark that shows significantly higher frequencies in English than in Dutch. Sentence (10) below contains an example of an English sentence with a dash.

- (10) <C><indepcl>This government must raise the money it needs fairly<indepcl><C> - <D><indepcl>and then explain that it is spending it wisely<indepcl><D>. <s272, newspaper articles>

As was noted in Chapter 7 and by Siepmann et al. (2008), in addition to its specifying function, the dash can also be used to mark off a comment that the writer wants to make about the information contained in the preceding nucleus (p. 211). Siepmann et al. describe this use as ‘illustrative of its interactional flavour’, characterising the dash as ‘a mark of a writer who is very much present in the text, closely monitoring it, aware of the reader’s needs and ready to jump in with a comment, a correction, or a reformulation’ (2008: 211). As Section 9.3.2 below will

show, this use of the dash is particularly frequent in the English newspaper genre. Instead of considering these differences in punctuation marks as mere differences in punctuation style between the languages, a closer analysis of the Dutch sentences linked by colons and the English sentences linked by dashes appears to point to a difference in rhetorical style. It is the identification of such differences in style that is the exact purpose of a sentencing analysis, i.e. analysing how the interplay between discourse structure, grammar and punctuation can achieve particular rhetorical effects.

In short, whereas the languages show remarkable overlap in the frequencies of the discourse subpatterns containing one appended satellite, differences between the languages become visible when the grammatical realisation of this satellite and punctuation mark used to link it to the preceding nucleus are taken into account as well.

### 9.2.6 Summary

A broad analysis of the discourse structure of English and Dutch sentences shows that despite some differences, the languages actually show remarkable similarities at the level of discourse, as all sentences in both languages fall into the same four main patterns. Differences between the languages do, however, arise when a closer look is taken at the length of sentences, the start of sentences, the occurrence and use of interpolated satellites and the grammatical realisation of various discourse units. The main frequency differences between the languages can be summarised as follows:

**Table 1 Differences between sentencing patterns Dutch and English summarised**

English	Dutch
<ul style="list-style-type: none"> <li>• longer sentences</li> <li>• higher frequency of sentences consisting of only nucleus</li> <li>• higher frequency of adverbial clauses, esp. non-finite, in sentence-initial position</li> <li>• higher frequency of sentences starting with two or more satellites, esp. ABC(X)</li> <li>• higher frequency of interpolated satellites</li> <li>• higher frequency of adverbial clauses, esp. non-finite, in sentence-final position</li> </ul>	<ul style="list-style-type: none"> <li>• shorter sentences</li> <li>• higher frequency of XC sentences with phrasal satellite at start</li> <li>• higher frequency of adjuncts/PPs in sentence-initial position</li> <li>• lower frequency of sentences starting with more than one satellite, mainly A1C(X)</li> <li>• lower frequency of interpolated satellites</li> <li>• higher frequency of independent clauses and non-restrictive clauses in sentence-final position</li> </ul>

It is interesting to note that with respect to the differences found between the languages, only a few can be explained by differences in the respective linguistic systems. This applies, for instance, to the lower frequency of the ABC(X) type of complex beginning in Dutch, which can be explained by its verb second principle, and to the higher frequency in English of non-finite clauses taking a present participle, which has been shown to be characteristic of this language. This would mean that the majority of differences between the languages are attributable to something other than their linguistic systems. As has been suggested, certain explanations may be sought in the writing cultures of the languages, which are here considered to be, at least to a certain extent, reflected in style guides and writing manuals of English and Dutch. Although a more thorough investigation of the concept of writing culture is still necessary, what is interesting is that there are a number of issues that receive strong focus in a considerable number of Dutch style guides, i.e. sentences with a long start (*lange aanloop*), sentences with a particular type of interruptive structure (*tangconstructie*) and sentence length. There is typically a strong focus on the notion of readability and the reader is made aware of how the use of certain constructions affects, or even hampers, readability. This would mean that despite the fact that the linguistic system of Dutch allows for certain constructions that could be considered to increase the complexity of sentences, writers are deliberately discouraged from using these. It should be noted that although similar advice can probably be found in some English style guides, it is not formulated as explicitly, at least not with respect to the same constructions. Of course, whether there is a direct relation between what has been found in the corpus designed for this study and what style guides prescribe is another matter, but it is rather remarkable that these are exactly the areas in which the languages show differences – start of sentences, interruptions and sentence length.

It may then be suggested that in the deliberate and explicit attempt to increase readability, Dutch sentences can give the impression of being less complex in nature when compared to English sentences. Whether this is also reflective of the fact that written Dutch is closer to the spoken language than written English, as has been suggested by Hannay & Mackenzie (2009: 39, 219, see also Hannay 1997: 234, 248), might be plausible, as various style guides do recommend that one write texts in a careful speech style ('verzorgde spreektaal') (cf. Doeve & Onrust 1992: 13ff; Van der Horst 1997: 81; Van der Spek 1997: 111), but this would have to be looked at in more detail.

The analysis of sentencing patterns does, in any case, show that writing culture has to be taken into consideration when trying to explain differences in sentencing patterns between the languages. As the second part of this chapter will show, the relevance of writing culture in a sentencing analysis will become even clearer when genre is also taken into account – a sort of writing subculture.

## **9.3 Main findings – languages \* genres**

Even though the four different genres were primarily incorporated in this study to fulfil one of the main criteria in corpus design, i.e. making sure that the corpus could be considered representative of the languages under consideration (cf. 4.3.1), statistical analyses have indicated that the vast majority of results that concern differences between English and Dutch cannot be given without making explicit reference to these genres. For each of these genres differences between the languages will be described on the basis of the five parameters of sentence length, main sentence patterns, beginning of sentences, interruptions and end of sentences.

### **9.3.1 Academic prose**

#### **Sentence length**

As was already indicated above (9.2.1), an analysis of sentence length shows that English sentences are significantly longer than Dutch sentences in three of the four genres, with the short stories genre being the exception. When compared to the other genres, the academic prose genre contains the lowest frequency of sentences in the shortest length category (1-10 words) and the highest frequency in the longest length category (31+ words) in both languages. There are, however, significant differences between the languages, with English showing significantly higher numbers in the 31+ words length category, to which almost 40% of the sentences belong, and Dutch containing significantly more sentences in the 1-10 words and 11-20 categories, to which 10.6% and 35.3% of the sentences belong respectively.

When using sentence length as a method to measure sentence complexity, as was suggested in 9.2.1. above, sentences in the English academic genre could be

considered more complex than the Dutch sentences in this genre. However, as sentence length only constitutes one of the methods to measure complexity, using this in combination with various other methods to analyse the structure of sentences might afford more insight into whether English sentences can indeed be considered more complex than Dutch sentences in this genre, as the following sections will illustrate.

### **Main sentence patterns**

What is remarkable is that despite the significant differences in sentence length between English and Dutch, the academic prose genre is the only genre for which no significant differences between the languages were found with respect to the frequencies of the four main sentence patterns. However, when comparing the frequencies of these main patterns to the other genres, clear differences can be found. Specifically, although about half of the sentences in this genre belong to the main C-pattern, this number is considerably lower than in other genres, with the numbers for patterns in which the nucleus is preceded or followed by satellites being considerably higher. This indicates that in comparison to the other genres, the academic prose genre appears to favour sentence structures in which the nucleus is preceded, followed or interrupted by satellites. Whenever a nucleus is combined with one or more satellites, this results in a more complex structure, in which various discourse units or pieces of information are hierarchically related to each other. In other words, the academic prose genre appears to favour complex sentences (cf. 1.3.3), in which information is hypotactically linked. The extent of hypotaxis – or subordination – is often used as an indicator of sentence complexity (cf. Cosme 2007: 200-201, 206-207). It is this higher frequency of sentences in which nuclei are accompanied by satellites in combination with the particularly long sentences that characterise the academic genre as complex.

With respect to its main pattern, the C-pattern, although no significant differences in frequencies were found between the languages, they did differ significantly in the make-up of this pattern. Not only does this pattern contain significantly more interpolated satellites in English than in Dutch, creating subpatterns such as the C1C and C12C subpatterns (see section below on Interruptions), English also contains significantly more instances of coordinated nuclei than Dutch. Specifically, in English more than one quarter of the nuclei in the C-pattern are coordinated with other nuclei, compared to 15% in Dutch. This higher frequency of coordinated nuclei in English is not restricted to the nuclei in the C-

pattern, but also applies to the nuclei in the most frequent subpatterns, namely the AC and CD subpatterns. With respect to the grammatical realisation of these coordinated nuclei, although they are typically realised as two or more independent clauses that are coordinated with each other in both languages, English shows a significantly higher instance of coordinated subordinate clauses or embedded clauses than Dutch. Sentence (11) contains an example of coordinated embedded clauses.

- (11) <Ca><coord\_a\_subcl>They may be similar to the children in other studies who do not select physical hobbies (Hall, 1988) and are not chosen by others to play games<coord\_a\_subcl><Ca> Smyth, 1992), <Cb><coord\_b\_subcl>and for whom social integration and motor ability are strongly related<coord\_b\_subcl><Cb>. <s5484, academic prose>

The higher frequency of these coordinated subordinate and embedded clauses, even in a sentence pattern that can in other terms be classified as simple, is one of the factors that contributes to the complexity that is associated with the English academic prose genre. This means that despite the fact that the languages show similar frequencies for the main sentence patterns, differences emerge when each of these patterns, including the subpatterns, is looked at in more detail.

### **Beginning of sentences**

Although Section 9.2.3 above already showed that the patterns in which the nucleus is preceded by only one satellite, the AC pattern and ACD pattern, are far more frequent than the patterns in which it is preceded by more than one satellite, the academic prose genre is the genre that in both languages, but especially in English, shows the highest frequency of complex beginnings in comparison with the other genres. In line with the results described in the previous section with respect to the frequencies of the main sentence patterns, but nonetheless worthy of note are the very similar frequencies of the AC pattern and ACD pattern in the academic prose genre in both languages.

Differences can, however, be found when looking at the grammatical realisation of the A-satellite in these subpatterns. In both languages this takes the form of a phrase in the vast majority of cases, with the group of clauses being significantly more frequent in English. As was suggested in 9.2.3, this lower frequency of clauses in sentence-initial position in Dutch could be linked to the

advice given in Dutch to avoid the use of sentences with a long start (*lange aanloop*). As for the phrasal A-satellites, despite the fact that the vast majority in both languages take the form of a phrase realised as an adjunct/PP, this particular realisation group is considerably larger in Dutch, whereas the group of conjuncts is considerably larger in English. This relatively high frequency of conjuncts in the English academic prose genre has also been noted by Biber et al. (1999: 880ff, also on high frequency of conjunct initial position, followed by medial position, p. 890-891), who explain that an important aspect of this genre is the presentation and support of arguments. They suggest that '[t]he higher frequency of linking adverbials in academic prose not only reflects this communicative need but also the characteristic choice of this register to mark the links between ideas overtly, as these arguments are developed' (p. 880, ). Although a similar line of argument would apply to the Dutch academic prose genre, this does not show a higher frequency of conjuncts when compared to the other genres. In fact, the overall frequency of conjuncts in Dutch is relatively high, i.e. accounting for around 25% of all sentence-initial phrases, and consistent across the different genres, with the exception of the short stories genre. This relatively high frequency might be related to the shorter sentence length in Dutch and the suggestion put forward in various style guides to, on the one hand, make sure that sentences do not become too long and, on the other hand, guarantee that the relation between the different sentences remains clear by the use of linking words (cf. Renkema 2005: 78-79; Hermans 2006: 51). Furthermore, a closer look at the type of sentence-initial conjuncts used in both languages shows that in English these are typically presented as separate punctuation units, thereby receiving the label A-satellite, and in Dutch they are almost exclusively realised as so-called zz elements, a label used to annotate those sentence-initial elements that are not presented as separate punctuation units (cf. 2.4.3). Whereas in English *however* is by far the most frequent conjunct (see Biber et al. 1999: 886/887 for a similar observation), Dutch shows a relatively high frequency of the rather short *zo* ( $\approx$  in this way) and *ook* (also) on the one hand, and a group of pronominal adverbs (*voornaamwoordelijke bijwoorden*) on the other hand, which can have a similar clause-linking function as conjuncts (cf. Haeseryn et al. 1997: 462, 490).

When realised as a clause, in both languages this almost exclusively takes the form of an adverbial clause of various types, with English showing higher frequencies of adverbial clauses of concession and of non-finite clauses in comparison to Dutch, and Dutch showing relatively high frequencies of adverbial clauses of condition. The frequency of both semantic types is in line with Biber et

al.'s findings for this genre, who explain that 'clauses of condition and concession are important contributors to the development of arguments, which is a significant goal of academic writing' (1999: 825). Sentence (12) provides an example of a concessive adverbial clause in the English academic prose genre.

- (12) <A><advcl\_conces>Although different studies have used different testing procedures and different cut-off points in classifying children<advcl\_conces><A>, <C><indepcl>there is a developing consensus that around 5-6% of children may have appreciable perceptual-motor problems<indepcl><C> (Smyth, 1992) . <s5285, academic prose>

With respect to complex beginnings, as was noted above, these are not only significantly more frequent in English on the whole, but also show a particularly high frequency in the academic prose genre when compared to the other genres. This is in line with Smits' findings (2002), who looked in great detail at the use and occurrence of complex beginnings in English and the English produced by Dutch learners. In her analysis of learner language, she also looked at the occurrence of complex beginnings in Dutch, but restricted this analysis to newspaper articles. The findings of the present study therefore supplement her extensive study by providing information about complex beginnings in four genres of Dutch. As for the type of complex beginnings in this genre, English shows a relatively high frequency of the ABC(X) pattern, especially in comparison to Dutch. In Dutch the majority of complex beginnings follow the A1C(X) pattern. With respect to the grammatical realisation of the satellites in these patterns, in both languages the A-satellite in the ABC(X) pattern predominantly takes the form of a conjunct and the B-satellite the form of an adjunct/PP and, especially in English, of an adverbial clause. It is this use of the conjunct in Dutch, i.e. in combination with another sentence-initial element, which is considered an anglicism by some, as was suggested in 9.2.3 above (cf. Haeseryn et al. 1997: 1278). Although infrequent in Dutch, its higher frequency in the academic prose genre could be linked the dominance of English in this genre, thereby perhaps influencing Dutch academic discourse. Sentence (1) below contains an example of a Dutch ABC(X) sentence with a conjunct in sentence-initial position.

- (13) <A><conj>Echter<conj><A>, <B><pp>bij kinderen waarbij het gedrag na interventie niet veranderd was<pp><B>, <C><indepcl>verhoogde de omgang met deviante leeftijdgenoten de kans op de ontwikkeling van een gedragsstoornis op latere leeftijd<indepcl><C>. <s6059, academic prose>



(<A><conj>However<conj><A>, <B><pp>with children of which the behaviour had not changed after intervention<pp><B>, <C><indepcl>the association with deviant peers increased the chance of developing a behavioural problem at a later age<indepcl><C>.)

Furthermore, although the A1AC(X) pattern is a rather infrequent subpattern, in both languages the academic prose genre shows the highest frequency of this subpattern when compared to other genres. The English academic prose genre then also contains most instances of the sentence pattern in which the nucleus is preceded by three satellites (11 of 18 occurrences in English in total). Sentences (14) and (15) provide examples of an English ABZC sentence and a Dutch A1AC respectively.

(14) <A><conj>Thus<conj><A>, <B><advcl\_cond>if a child correctly positions a toy bottle near the mouth of a doll<advcl\_cond><B>, <Z><advcl\_nonfin>as if feeding the doll<advcl\_nonfin><Z>, <C><indepcl>this is functional play<indepcl><C>.  
<s5394, academic prose>

(15) <A><advcl\_time><coord\_a\_phr>Nadat zijn moeder in Maxi's afwezigheid - <1\_prefv><coord\_b\_phr>maar zichtbaar voor de proefpersoon<coord\_b\_phr><1\_prefv> - de chocolade naar het witte kastje heeft verplaatst<coord\_a\_phr><advcl\_time><A>, <C><indepcl>komt een nietsvermoedende Maxi terug voor zijn chocolade<indepcl><C>. <s6157, academic prose>

(<A><advcl\_time><coord\_a\_phr>After his mother in Maxi's presence, - <1\_prefv><coord\_b\_phr>but visible for the subject<coord\_b\_phr><1\_prefv> - has moved the chocolate to the white cupboard <coord\_a\_phr><advcl\_time><A>, <C><indepcl>an unsuspecting Maxi returns for his chocolate<indepcl><C>.)

An analysis of sentence beginnings thus shows that in both languages this is the genre that contains most complex beginnings. This relatively high frequency is what gives the genre as a whole, but particularly the English academic prose genre, the impression of being more complex when compared to other genres. An interesting hypothesis is whether the higher frequency of these patterns in the Dutch academic prose genre when compared to the other Dutch genres could be due to the dominance of English in this particular text type or whether this higher frequency can simply be explained by the very nature of this genre as a more complex text type in both languages.

### Interruptions

Interpolated satellites have been shown to occur significantly more frequently in English than in Dutch across all genres (see 9.2.4 above and Chapter 8). In both languages, however, the frequency of interruptions is particularly high in the academic prose genre, with nearly 20% of the English sentences in this genre containing an interruption and 14% of the Dutch sentences. As was also explained in 9.2.4 above, the vast majority of these interruptions occur in the nucleus and can either precede or follow the finite verb, with Dutch showing a preference for the postfinite position due to its V2 character.

An analysis of their grammatical realisation shows that in both languages the vast majority are realised as a phrase. In English the largest phrasal realisation groups are formed by appositions, adjuncts/PPs and conjuncts. In Dutch over half of the interruptions take the form of an apposition, followed by adjuncts/PPs and a fairly large third realisation group is formed by premodifiers (see 9.2.4 above), which especially applies to those interruptions that occur after the finite verb. With respect to the most frequent realisation form, the apposition, both Meyer (1992: 98), in his extensive study of appositions, and Biber et al. (1999: 639ff) have shown that they are indeed particularly common in this genre and in the newspaper genre (see 9.3.2 below). In describing the pragmatics of appositions, Meyer explains that as both these genres can be characterised by the fact that ‘the discourse participants possess a low amount of shared knowledge’, this might make appositions ‘communicatively more necessary’ in these genres than in other genres (1992: 98). With respect to their grammatical form, Biber et al. explain that appositions, particularly those realised as noun phrases, are ‘a maximally abbreviated form of postmodifier’, which are ‘favored in the registers with highest information density’ (1999: 639). Sentence (16) below contains an interruption realised as an apposition, taken from the English academic prose genre.

- (16) <zz><adj>In the second section<adj><zz> <C><indepcl> the other patriotisms of the British Isles -<1\_prefv><appos\_NP>loyalty to Wales, Scotland, Ireland and England<appos\_NP><1\_prefv> - are considered in much the same way<indepcl><C>. <s4176, academic prose>

As for the other realisation groups, it was already suggested above (see 9.2.4) that despite the fact that it could be seen as constituting a mere punctuation difference between English and Dutch, interpolated satellites that take the form of conjuncts in English could be interpreted as creating a clear break in the sentence.

This creates a different rhetorical effect when compared to many Dutch conjuncts that are placed sentence-medially, but are not presented as separate punctuation units. A closer look at the type of conjuncts in English shows that the largest groups are formed by *however* and *for example/for instance*, which is in line with Biber et al.'s findings for this genre (1999: 881, 887). The conjunct in English combined with the relatively large group of premodifiers in Dutch creates the picture that a subset of interruptions performs a different function in the respective languages, with the Dutch subset only having semi-interruption status, as was suggested above (9.2.4). An example of this Dutch premodifier use in this genre is presented in sentence (17).

- (17) <C><indepcl>Deze activiteiten hadden tot doel het product dusdanig te (<1\_postfv><premod>her<premod><1\_postfv>) ontwikkelen en te vormen dat het aansloot op de wensen en behoeften van de consument<indepcl><C>. <s4871, academic prose>

(<C><indepcl>These activities had the function of (<1\_postfv><premod>re<premod><1\_postfv>) developing and forming the product in such a way that it fulfilled the wishes and needs of the consumer<indepcl><C>.)

Furthermore, although it only concerns the minority of interruptions in both languages, there is also a group of interruptions that takes the form of a clause. In both languages these are predominantly realised as finite adverbial clauses, non-finite clauses or non-restrictive relative clauses. The non-finite clauses, both taking a present participle and past participle, are particularly frequent in English, but also surprisingly so in Dutch, then containing a past participle (cf. Biber et al. 1999: 630 for similar findings with respect to post-modification of non-finite clauses in the English academic prose genre). With respect to the semantic role of the finite adverbial clauses, the languages also show overlap, as a large group takes the form of an adverbial clause of comparison, followed by time and condition. In English a small group is also formed by adverbial clauses of concession. This means that with respect to the clausal interruptions the languages show overlap in the type of grammatical form they typically take, but subtle differences in the frequencies between the different groups.

An analysis of interruptions thus shows that in both languages this is a genre that shows a higher frequency when compared to other genres. A reason for this can probably be sought in the characteristic of this text type as having a high

information density (cf. Biber 1999: 639) and interpolated satellites forming an efficient way of incorporating information in a sentence. The analysis also shows that they are particularly frequent in English, with a subset of them disrupting the flow of the sentence more noticeably than in Dutch. The idea that a high number of interruptions can be related to increased complexity of a text, in combination with long sentences and complex beginnings, is supported by the fact that a large number of Dutch style guides discourage the use of them as they associate it with undesirable sentence complexity, as was explained in 9.2.4 above.

### **Ends of sentences**

Despite the fact that the third main sentence pattern, the CX pattern, shows almost identical frequencies for English and Dutch in the academic prose genre, interesting differences arise when the grammatical realisation of the discourse units is also taken into account. For instance, when looking at the D-satellite in the CD, ACD and CDE subpatterns, in both languages this takes the form of a clause in the vast majority of cases, with differences being found in the types of clauses that dominate. In English by far the largest group is formed by adverbial clauses, i.e. nearly 75%, followed by non-restrictive relative clauses, whereas Dutch shows a more even distribution across the three main realisation groups of adverbial clauses, non-restrictive relative clauses and independent clauses. The largest realisation group in English, consisting of adverbial clauses, take the form of a non-finite clause in over half of the cases, which is a highly significant difference with Dutch, and predominantly contain a present participle. This high frequency of non-finite clauses with a present participle is not immediately supported by Biber et al.'s study, who label them supplementive clauses (1999: 782-783, 824-825), but it should be noted that they have included both embedded adverbial clauses, i.e. not marked off by means of punctuation marks, and non-embedded clauses in their study, which might explain the differences between their findings and those of the present study. With regard to the finite adverbial clauses, the largest group in English is formed by adverbial clauses of concession (cf. Biber 1999: 825 for similar findings), whereas in Dutch the frequencies are more evenly divided across adverbial clauses of time, reason and concession. An example of an English sentence containing a D-satellite realised as an adverbial clause of concession is presented in sentence (18) below.

- (18) <C><indepcl>The cognitive skills for the humorous appreciation of incongruity (<1\_prefv><appos\_NP>absurdity or the juxtaposition of different frames of reference<appos\_NP><1\_prefv>) are believed to develop after 18 months<indepcl><C> (McGhee, 1979), <D><advcl\_conces>although primitive precursors to humour can be seen earlier in laughter in response to tickling, peekaboo and chasing<advcl\_conces><D> (Shultz, 1976). <s5821, academic prose>

As for the third largest realisation group in Dutch, the independent clause, it was already explained in 9.2.5 above that this is linked to the preceding nucleus by means of a colon in a considerable number of cases. Interestingly, this is the genre that shows the highest frequency of this pattern in comparison with the other genres. This may be in line with the function of this particular punctuation mark of introducing a further explanation or elaboration of the information contained in the nucleus. Although English also shows some instances of this use, the colon is predominantly used to introduce lists or appositions that occur in sentence-final position. Sentence (19) presents an example of the Dutch use of the colon in this genre.

- (19) <C><indepcl>Het is belangrijk voor hen om contact te maken met nieuwe vrienden en/of de bestaande vriendschappen te versterken<indepcl><C>: <D><indepcl>dat geeft hun de mogelijkheid om reacties te krijgen van vrienden op hun eigen ideeën, meningen en emoties<indepcl><D> (Brown, 1990; Fuligni & Eccles, 1993). <s5941, academic prose>

<C><indepcl>It is important for them to make contact with new friends and/or to strengthen existing friendships<indepcl><C>: <D><indepcl>that gives them the possibility of getting reactions from friends on their ideas, opinions and emotions<indepcl><D> (Brown, 1990; Fuligni & Eccles, 1993).)

With respect to the phrasal appended satellites, the languages show more overlap, with the vast majority in both languages taking the form of an adjunct/PP or an apposition. These realisation groups are exemplified by sentences (20) and (21) below.

- (20) <C><indepcl>Een van de kenmerken van de historische discipline is de verkaveling van de geschiedenis in welomschreven tijdvakken<indepcl><C>, <D><nonrestr\_relcl>waarbij de cesuren belangrijke veranderingen in de ontwikkeling van menselijke samenlevingen markeren<nonrestr\_relcl><D>, <E><appos\_NP\_list>zoals de val van het Romeinse Rijk, de Verlichting, de

Industriële Revolutie, de Franse Revolutie en de Eerste Wereldoorlog<appos\_NP\_list><E>. <s4925, academic prose>

(<C><indepcl>One of the characteristics of the historical discipline is the subdivision of history in well-described time slots<indepcl><C>, <D><nonrestr\_relcl>where the cut-off points mark important changes in the development of human societies <nonrestr\_relcl><D>, <E><appos\_NP\_list>such as the fall of the Roman Empire, the Enlightenment, the Industrial Revolution, the French Revolution and the First World War<appos\_NP\_list><E>.)

- (21) <A><advcl\_nonfin>As expected by the moderation hypothesis<advcl\_nonfin><A>, <C><indepcl>the results of the present study demonstrated that the subjective ease with which participants could bring to mind positive or negative attributes about British Prime Minister Tony Blair affected the favourability of participants' attitudes<indepcl><C>, <D><pp>but only among those participants who were uninterested with British politics<pp><D>. <s5865, academic prose>

An analysis of the satellites occurring in sentence-final position thus shows that in both languages these are predominantly realised as clauses and that the same grammatical categories occur in both languages, but that English and Dutch differ with respect to the frequencies of the different clause types. Whereas English shows a high frequency of non-finite clauses and finite adverbial clauses of concession, Dutch shows higher frequencies of non-restrictive and independent clauses in this position when compared to English.

### Summary

The picture that emerges when analysing the sentencing patterns in the academic prose genre is that of a text type that can, in many respects, be characterised as complex in both languages. The factors that contribute to this notion of complexity are a relatively high number of lengthy sentences, complex sentence beginnings, interpolated satellites and the highest frequency of the most complex sentence pattern, the XCX pattern, in comparison to the other genres.

As each of these constructions are considerably more frequent in English, the English academic prose genre gives the impression of being even more complex than the Dutch academic prose genre. This is also caused by the fact that some constructions are not just higher in number, but also different – and arguably more complex – in type. This applies, for instance, to the make up of the main C-pattern,

which not only contains more instances of coordination, but especially of the coordination of embedded and subordinate clauses. Another example is presented by a difference in the grammatical realisation of a substantial group of interruptions. Whereas these clearly provide a break in the flow of the English sentence, a considerable number in Dutch can hardly be seen as intrusive.

Interestingly, the languages show remarkable similarity with respect to the frequencies of the four main sentences patterns. Whether this similarity in discourse structure and thus at the most coarse-grained level of analysis is reflective of the dominance of English in the academic prose genre, thereby affecting Dutch sentence structure also with respect to, for instance, a higher frequency of complex beginnings in this Dutch genre, would have to be looked into in more detail, but presents an interesting hypothesis. However, differences between the languages arise when taking a closer look at the frequencies of various subpatterns, with English showing higher frequencies of the subpatterns with multiple satellites in sentence-initial position and of the subpatterns created by interruptions. More differences arise when taking a closer look not at the type, but at the frequency of the various grammatical categories. For instance, with respect to the grammatical realisation in sentence-final position, English again shows higher frequency of the non-finite clause and finite adverbial clauses of concession, whereas Dutch shows higher frequencies of the non-restrictive clause and independent clause, typically separated from its preceding nucleus by means of a colon.

### **9.3.2 Newspaper articles**

#### **Sentence length**

Similar to the academic prose genre, the sentences in the newspaper genre are again longer in English than in Dutch. The distribution of sentences across the four sentence length categories is, however, different from the academic prose genre. Specifically, whereas the academic prose genre contains only a very small number of sentences in the 1-10 word category and a relatively large number of sentences in the 31+ word category, especially in English, the newspaper genre contains a relatively large number of sentences in the shortest word category in both languages, around 20% in both languages. Differences between the languages can be found for the 11-20 words category, which contains significantly more sentences in Dutch (50.0% vs. 40.6%), and the 31+ words category, which contains

significantly more sentences in English (11.7% vs. 6.5%). These findings are in line with earlier findings of contrastive analyses of Dutch and English press editorials carried out by Hannay (1997: 243-244) and, more recently, by Cosme (2007: 259). It should be noted that in addition to press editorials, the newspaper subcorpus partly consists of news reportage articles (see 4.3.2 for the exact make-up of newspaper subcorpus).

Just as the high number of longer sentences and low number of particularly short sentences in the academic prose genre could be considered to suit the style of a genre that has in various respects been classified as complex, so it would be interesting to see if the relatively high number of sentences up to 20 words in both languages in the newspaper genre, accounting for 62.3% of the sentences in English and 71.3% in Dutch, could also be related to a particular style or classification of this text type. With respect to their high frequency, a possible explanation could, at least for Dutch, be sought in guidelines provided in writing manuals for journalists, such as Donkers & Willems (2002), who explicitly advise adopting an average sentence length of between 10 and 20 words (p. 183) or in style guides of leading Dutch national newspapers. For instance, the style guide of *de Volkskrant* (van Gessel et al. 2006) explains that sentences exceeding 20 words could impair readability and suggests using these sparingly and cutting sentences off at clause boundaries instead of building compound or complex sentences (p. 34). The style guide of *Trouw*, another Dutch national newspaper, states that a national newspaper aimed at reaching a broad readership should adopt an average sentence length of 15 words (de Berg 2006: 151). An analysis of guidelines provides in style guides of a number of English national newspapers or writing manuals for journalists (eg. Kay 1990; Austin 2003; Pape & Featherstone 2005; March 2007) shows that no explicit reference is made to sentence length or readability of sentences. Instead, the focus is more on the choice of words, use of jargon and, for instance, adhering to the KISS principle: keeping it short and simple (Kay 1990: 104; Pape & Featherstone 2005: 58). In the exceptional cases in which sentence length is explicitly addressed, the advice is formulated along the lines of '[i]n general, prefer the short sentence to the long one, particularly in the intro' and '[a]void over-complex sentences full of subordinate clauses and phrases' (Hicks 2007: 90).

With respect to the function of the shorter sentences, particularly the very short ones, an analysis of their discourse-grammatical features in combination with their semantic content may prove fruitful. In his analysis of English and Dutch press editorials, Hannay (1997) makes an attempt at explaining the function of particularly short sentences and finds that in Dutch a considerable number of these



are either questions or answers to questions (1997: 244). In his view these differences suggest that ‘short sentences do different work in the two languages with regard to the organisation of the discourse and the shaping of the argument’ (*ibid*). Although a frequency analysis of the number of question marks in combination with short sentences does not immediately support Hannay’s findings, a more systematic analysis of these sentences might yield similar results. A quick, non-systematic qualitative analysis of a selection of the sentences belonging to the shortest length category does, however, give the impression that the function of these sentences is not only dependent on language, but also on genre, subgenre and even on source. Whereas the frequency of particularly short sentences does not appear to be dependent on subgenre, i.e. news reportage and editorials, it does appear to be dependent on source, i.e. *the Daily Mirror* vs. *the Guardian* in English and *de Telegraaf* vs. *de Volkskrant* in Dutch, with the English *the Daily Mirror* containing a larger number of sentences in the 1-10 word category (259 (64.8%)) than *the Guardian* (141 (35.2%)). Dutch shows a more even distribution across the two newspapers. A closer look at these short sentences also shows that especially English contains quite a few sentences that start with a sentence-initial coordinator, mostly *but*, in comparison with Dutch (7.5% vs. 3.5% of the sentences). Biber et al. (1999) have also noted this sentence-initial use of a coordinator, against which there is a ‘well-known prescriptive reaction’ (p. 83), and have found that it is indeed more frequent in the newspaper genre than the academic prose genre, often occurring at paragraph boundaries, ‘where they create a marked effect’ (p. 84). In both languages, these coordinators are more frequent in the editorial than the news reportage subgenre and are, in English, more frequent in *the Daily Mirror* than *the Guardian*.

Also noteworthy is the higher number of nuclei of the short sentences in English that are realised as non-independent clauses when compared to Dutch (9.0% vs. 4.0%). It is these discourse-grammatical features in combination with an analysis of their semantic content that give especially the English short sentences the impression of being often critical and powerful writer-present statements or comments. This is in line with, for instance, the advice given by Kay that ‘[s]entence fragments are like exclamation points. They’re emphatic. Good writers use them as they use exclamation points – sparingly’ (1990: 100). Examples (22) and (23) below are taken from the *Daily Mirror* and *Guardian* respectively. For both examples the surrounding sentences are provided to show that the short sentences are often either part of a series of particularly short sentences, making a number of statements in a row, as in (22), or are positioned in between somewhat longer

sentences, in which case the contrast in sentence length adds to its powerful effect, as in (23).

- (22) <zz><conj>Yet<conj><zz> <C><indepcl>President Bush launched a mindless, ill-thought-out adventure in Iraq<indepcl><C>, <D><advcl\_nonfin>aided and abetted by Tony Blair<advcl\_nonfin><D>. <s169, newspaper articles, DM>  
<zz><adj>Each day<adj><zz> <C><indepcl>we see what that means<indepcl><C>. <s170, newspaper articles, DM> <C><fragment\_PP>In deaths, suffering and terror<fragment\_PP><C>. <s171, newspaper articles, DM>
- (23) <C><indepcl>The winner is clearly Charles Kennedy<indepcl><C>, <D><nonrestr\_relcl>who stood aside from the inquiry when it was announced early last month in protest at its restrictive terms<nonrestr\_relcl><D>, <E><pp>despite strong pressure from No 10 and private unease on his own part<pp><E>. <s1086, newspaper articles, GD> <C><indepcl>The losers are Tony Blair and Michael Howard<indepcl><C>. <s1087, newspaper articles, GD>  
<C><indepcl>The former is left with a winged investigation<indepcl><C>, <D><advcl\_nonfin>wounded but not killed off<advcl\_nonfin><D>, <E><advcl\_nonfin>destined now to be seen as a cover-up should it do anything other than directly criticise the prime minister when it reports in the summer<advcl\_nonfin><E>. <s1088, newspaper articles, GD>

Although definitely also present in Dutch, whether this particular use of short sentences is more frequent in the English newspaper genre and whether it does indeed often fulfil this particular function would have to be analysed in more detail. Analyses such as these ones, in which quantitative discourse-grammatical features are combined with more qualitative analyses, do, however, appear to support the claim that text type, extending as far as subgenre and source, has a considerable influence on both the structure and function of sentences.

### Main sentence patterns

With respect to the distribution of sentences across the four main sentence patterns, the newspaper genre follows the overall trend, in which the largest number of sentences belong to the C-pattern and the lowest number of sentences to the XCX pattern. As this applies to three of the four genres, with the exception of the academic prose genre, it is not likely that this distribution is particular for the newspaper genre. However, significant differences between the languages can be found for the frequencies of the C-pattern, which is more frequent in English than Dutch, taking up nearly 70% of all sentences in English, compared to 57.6% of the

sentences in Dutch. Dutch, on the other hand, shows significantly higher numbers of sentences that belong to the XC pattern, namely 26.6% compared with 16.9% in English. In line with the overall picture as presented in 9.2.2, the languages show remarkably similar frequencies for the CX pattern, both containing only around 12% of the sentences in this category, and only a very small number of sentences that follow the XCX pattern (2.5% in English; 4.1% in Dutch).

As for the make-up of the C main pattern, the languages again show much overlap, with the vast majority taking the form of a single, uncoordinated nucleus (around 86.0% in both languages) that is realised as an independent clause. English shows a higher number of nuclei that take the form of a non-independent clause, but those are predominantly reporting clauses. i.e. clause fragments that introduce reported speech.

### **Beginning of sentences**

Similar to the academic prose genre, the patterns in which the nucleus is preceded by only one prepended satellite, i.e. the AC and ACD subpatterns, are the most frequent ones in both languages. A difference between the languages for this genre is that Dutch shows significantly higher frequencies of the XC main pattern as a whole, which is mainly caused by the AC subpattern, whereas English again shows a significantly higher frequency of the sentence patterns in which the nucleus is preceded by two or more elements.

With respect to the grammatical realisation of the A-satellite in the AC and ACD subpatterns, similar to the academic prose genre, this takes the form of a phrase in the vast majority of cases in both languages. The distribution across the different phrasal realisation categories is for English rather different from the academic prose genre, whereas for Dutch these genres show more overlap in this respect. In English the group of adjuncts/PPs is considerably larger in the newspaper genre than in any of the other genres, whereas the group of conjuncts is considerably smaller than in, for example, the academic prose genre. This is in line with Biber et al.'s findings, who characterise the newspaper genre as a text type that is 'particularly concerned with current events' (1999: 785), which, in their view, explains the particularly high frequency of time adverbials, but also process adverbials and place adverbials (*ibid*). A similar explanation seems plausible for the Dutch newspaper genre, in which adjuncts/PPs also form the largest realisation group. However, as no extensive multi-genre corpus analysis comparable to the one carried by Biber et al. (1999) has thus far been performed for Dutch, there is

no reference point against which the frequencies as found in the present study can be compared.

Furthermore, although the frequency of conjuncts in Dutch is rather consistent across the different genres, with the exception of the short stories genre, the relatively high frequency of them in this genre when compared to English (26.6% vs. 17.3%) could, as was already suggested above, perhaps be related to the seemingly deliberate shorter sentence length. In addition to the advice given in various style guides to adopt an average sentence length of between 15 and 20 words in this genre, a number of suggestions are also provided about ways in which sentences can be kept short. One of these is to cut off sentences after the completion of the first independent clause, before the possible insertion of a coordinating or subordinating conjunction that starts a new clause (cf. van Gessel et al. 2006: 34; de Berg 2006: 152). The relatively high frequency of these clause-linking elements in Dutch could thus potentially be related to the suggestions given to cut off sentences at exactly this point in order to reduce their length. Sentence (24) provides an example of a Dutch pronominal adverb *daardoor* that has the function of a conjunct. Although claims such as the following need careful analysis, it could be suggested that in examples such as the one below there may be a stronger tendency in English to link the second clause, now introduced by a conjunct, to the preceding clause by changing it, for instance, into a non-restrictive relative clause. This, in turn, may then also be related to the difference in sentence length between the languages.

- (24) <zz><pp>Door de beschadiging van zijn persoon<pp><zz> <C><indepcl>is automatisch ook de functie die hij vervult in diskrediet gebracht<indepcl><C>. <s2181, newspaper articles> <zz><conj>Daardoor<conj><zz> <C><indepcl>moet worden betwijfeld of Oudkerk nog wel goed kan functioneren als wethouder<indepcl><C>. <s2182, newspaper articles>

<zz><pp>Due to the damage of him as a person<pp><zz> <C><indepcl>this has automatically discredited the function that he fulfils<indepcl><C>. <s2181, newspaper articles> <zz><conj>Because of this<conj><zz> <C><indepcl>it should be questioned whether Oudkerk is still able to function well as alderman<indepcl><C>. <s2182, newspaper articles>

In addition to the large group of pronominal adverbs, two other large groups are formed by the particularly short conjuncts *ook* (also) and *zo* (≈so), similar to the academic prose genre.

For those cases in which the A in the AC and ACD subpatterns take the form of a clause, the numbers are significantly higher for English than for Dutch, which could, again, be linked to the advice given in Dutch style guides to avoid the use of a long start (*lange aanloop*) (see 9.2.3 above). In both languages the largest realisation groups are formed by adverbial clauses of condition, particularly in Dutch, adverbial clauses of time or non-finite adverbial clauses, particularly in English. With respect to the latter category, the largest group is again formed by non-finite adverbial clauses with a present participle, as in sentence (25) below.

- (25) <A><advcl\_nonfin>Having won the war<advcl\_nonfin><A>, <C><indepcl>Mr Bush is perilously close to losing the peace<indepcl><C>. <s250, newspaper genre>

With respect to the sentences with more than one sentence-initial element, the A1C(X) pattern shows the highest frequencies in both languages, with English again showing a significantly higher overall frequency of complex beginnings, especially the ABC(X) subpattern. It should be noted, however, that in both languages the sentences with complex beginnings are markedly lower in number when compared to the academic prose genre, with the newspaper genre ranking third. As for the grammatical realisation of the A-satellite in the A1C pattern, this almost exclusively takes the form of an adjunct/PP, in line with the high frequency of the adjunct/PP in the AC and ACD subpatterns. Furthermore, whereas the interpolated satellite in the A1C subpattern in the English academic prose genre appeared a popular position for conjuncts, this is predominantly realised as an adverbial clause, mainly non-finite, adjunct/PP or apposition in the newspaper genre. Sentence (26) presents an example of a sentence with the A1C subpattern in the English newspaper genre.

- (26) <A><adj>Later<adj><A>, <1><advcl\_nonfin>beginning in 1915<advcl\_nonfin><1>, <Ca><coord\_a>it occupied Haiti for 19 years<coord\_a><Ca> <Cb><coord\_b>and then abruptly left<coord\_b><Cb>. <s1238, newspaper articles>

### Interruptions

As Section 9.2.4 above already indicated, sentences containing one or more interpolated satellites occur significantly more frequently in English than in Dutch across all genres. Following the academic prose genre, the newspaper genre contains the second largest number of interruptions when compared to the two

remaining genres, irrespective of language. In both languages the vast majority of sentences containing an interruption follow the C1C subpattern and English shows a stronger, but non-significant, preference for interruptions that are positioned before the finite verb of the nucleus than Dutch.

Similar to the academic prose genre, the vast majority of interruptions in both languages in the newspaper genre, i.e. two-thirds, take the form of a phrase. What is more, but particularly in this genre, a high number of interruptions are realised as appositions (see section on interruptions in academic prose and reference to Meyer 1992 and Biber et al. 1999 above). With respect to the other phrasal interruptions, subtle differences between the languages can be found for the less frequently occurring grammatical realisations: the premodifier in Dutch and the situation in which a second coordinate is presented as an interruption in English. The interesting situation in the latter case is that two units are in a paratactic relation when viewed from a grammatical perspective, but in a hypotactic or hierarchical relation when viewed from a discourse perspective, an interaction that has been captured by the annotation system as it analyses at both these levels (see 3.4.1 for more information on interpolated coordination). Although this occurs in both languages, this realisation form shows higher frequencies for English than for Dutch. Sentence (27) below presents an example of this type of interruption (see 3.3.1, footnote 10 for an explanation and the use of coordination labels in examples such as (27)).

(27) <C><coord\_a\_phr>The decision to release five of the nine -  
<1\_prefv><coord\_b\_phr>and the admission by Home Secretary David Blunkett  
that they pose no threat<coord\_b\_phr><1\_prefv> - is shameful recognition they  
were wrongly detained<coord\_a\_phr><C>. <s339, newspaper articles>

In addition to showing differences in grammatical realisation when compared to the academic prose genre, a closer analysis of the interruptions in the newspaper genre reveals that a subset of them appears to have a genre-specific function. Besides the general function of providing additional, factual information in a condensed format, the interruptions in the newspaper genre also seem to be used by writers as a position in the sentence in which they comment or reflect on the subject at hand (cf. Siepmann et al. 2008: 168ff on different functions of interruptions). This function is exemplified by sentence (27) above. On first reading, the information contained in the interruption may appear to present purely factual information that is presented as an aside, but a closer reading of the sentence and

the article as a whole shows that it is not purely intended as a piece of factual information. Specifically, the focus of the article is to direct harsh criticism towards the Home Secretary's unjust treatment of detainees. Sentence (27) states that 5 detainees were wrongly detained, a blunder the author wants to underline by explaining in the interruption that it was a mistake to detain them as they posed no threat. It is precisely the interaction between, or perhaps even the mismatch between, the information contained in the interruption – considered a relevant detail by the writer – and the discourse status of interruptions, one that is lower in hierarchy than the other elements, by which the writer achieves the effect of making this information stand out. Although this function needs to be looked at in more detail, a first analysis appears to indicate that is used more frequently in the English newspaper genre than in the Dutch newspaper genre.

Moreover, in addition to forming a nice sentence to exemplify the interaction between discourse status and grammatical realisation, sentence (27) is also representative of a difference in punctuation practice between English and Dutch that applies to a considerable number of interruptions. Specifically, despite the fact that the vast majority of interruptions in both languages are typically marked off from the rest of the sentence by means of commas, a substantial number of them in English are introduced by paired dashes, whereas in Dutch a substantial number of them, not just those with the function of a premodifier, are surrounded by brackets. The distinction between these punctuation marks, especially in combination with interpolated satellites, is something that many style guides in both languages draw explicit attention to. For instance, *the Oxford Guide to Style* advises using the dash to 'express a more pronounced break in sentence structure than commas, and to draw more attention to the enclosed phrase than parentheses (Ritter 2002: 141; see also Anson & Schwegler 1998: 504; Fowler & Aaron 2010: 471, and for Dutch: Koenen & Smits 2004: 294; Tiggeler 2006: 164; Burger & de Jong 2009: 267). As the difference in effect of using either punctuation mark is explicitly addressed in the style guides of both languages, it could be suggested that the choice of one particular punctuation mark over another is a deliberate one and in that sense reflective of the writer's rhetorical intentions concerning the information contained in the interpolated satellite. With respect to the English dash in this genre, an analysis shows that its use could be described as 'the mark of a writer who is very much present in the text', to use Siepmann et al.'s words (2008: 211). Siepmann et al. also note that the status of the information that is contained between paired dashes is significantly different from the information contained between brackets: '[w]hereas brackets are used for information which is

considered really backgrounded, dashes are used to mark off information which the writer wishes to give special prominence' (2008: 209). This function is exemplified by sentence (27) above and by (28) below, where the information between dashes serves to remind the reader of controversial and damaging information about the subject of this sentence, Aitken. It should be noted that it is the combination of an interruption, with the choice of punctuation marks and particular choice of words (*sensational*) that create an added effect of making the information stand out.

- (28) <C><indepcl>Aitken - <1\_postfv><advcl\_nonfin>jailed for perjury after a sensational libel trial<advcl\_nonfin><1\_postfv> - announced yesterday his desire to get back into front-line politics as an MP<indepcl><C>. <s625, newspaper articles>

Since in almost half of the cases in the Dutch newspaper genre and an even larger number in the Dutch leaflets genre the interruptions are surrounded by brackets, the choice of this punctuation mark could be considered representative of a difference in function of at least a subset of the interruptions in the English and Dutch newspaper genre, with English using the interruptions as a foregrounding device of particular information and Dutch as a backgrounding device. This function is exemplified by sentence (29) below.

- (29) <A><pp>Ook op ander terrein dat te maken heeft met onderwijs en geld<pp><A>, <C><indepcl>heeft minister Van der Hoeven (<1\_postfv><appos\_NP>Onderwijs<appos\_NP><1\_postfv>) gisteren een verstandig standpunt ingenomen<indepcl><C>. <s2197, newspaper articles>.

<A><pp>Also on a different territory that has to do with education and money<pp><A>, <C><indepcl>has Minister Van der Hoeven (<1\_postfv><appos\_NP>Education<appos\_NP><1\_postfv>) yesterday a sensible approach adopted<indepcl><C>.)

<A><pp>Also on a different territory that has to do with education and money<pp><A>, <C><indepcl>Minister Van der Hoeven (<1\_postfv><appos\_NP>Education<appos\_NP><1\_postfv>) adopted a sensible approach yesterday<indepcl><C>.)



### Ends of sentences

As was already noted in 9.2.5 above, the languages show remarkably similar frequencies for the CX main pattern across the different genres, with the vast majority of sentences belonging to the CD subpattern. However, interesting differences between the languages for this genre arise when looking not only at the grammatical realisation of especially the appended satellite in the CD and ACD subpatterns, but particularly at the interaction between grammatical category and punctuation mark used to separate the appended satellite from its preceding nucleus.

As for the grammatical realisation of the D-satellite, this takes the form of a clause in the majority of cases in both languages. In English, the largest realisation group is formed by adverbial clauses, which are fairly evenly distributed across the finite and non-finite categories, with the latter group again occurring significantly more frequently in English than in Dutch. As for the type of non-finite adverbial clause, this genre shows a higher frequency for those with a past participle in comparison with the other genres in which the type with a present participle dominates. When finite, a large group in English is again formed by adverbial clauses of concession, followed by adverbial clauses of time. A second large clausal realisation group is formed by the rather subgenre-specific reporting clause, which, on closer inspection, almost exclusively occurs in the news reportage articles subgenre (cf. Biber 1999: 923 on the high frequency of these clause types in the news genre). A third, again considerable, genre-specific realisation group that is particularly frequent in English is formed by appended clauses. This realisation group was discussed in quite some detail in Chapter 3 (3.4.1), as it not only constitutes a label that was applied to a group of clauses and clause fragments that did not fit any of the other realisation categories, but particularly because it constitutes a group of appended satellites that are positioned on the coordination-subordination gradient. Similar to the particular type of interruptions described above that take the form of second coordinates, often marked off from the surrounding sentence by paired dashes, a considerable number of the appended clauses are introduced by coordinators and marked off from the preceding nucleus by a single dash – a punctuation mark that is considerably more frequent in the English newspaper genre when compared to the other genres, as was noted above (see also 7.2). As was explained in 3.4.1, the reason for classifying this type of appended clause as a satellite instead of a coordinated nucleus is partly motivated by their particular grammatical form, which has been labelled appended coordination by Quirk et al. (1985: 975), who describe it as a ‘loose kind of

coordination' in which the second coordinate is added as an afterthought, and as supplementation by Huddleston & Pullum (2005: 1350), used for constructions that have the character of appendages or interpolations. The present approach to these constructions as satellites is further motivated by the fact that the majority of these clauses are marked off from the preceding nucleus by means of a dash, which on the hierarchy of punctuation marks (cf. 2.4.1, Quirk et al. 1985: 1612; Huddleston & Pullum 2002: 1731) marks a stronger break from the surrounding text than, for instance, a comma. Similar to their use of foregrounding the information that is contained in interpolated satellites, they also place additional emphasis on the information contained in appended satellites separated from the nucleus by means of a single dash. This use is exemplified by sentences (30) and (31) below.

- (30) `<Ca><coord_a_asyn>Bosses earn massively more than they used to<coord_a_asyn><Ca>, <Cb><coord_b_asyn>get enormous bonuses<coord_b_asyn><Cb>, <Cc><coord_c_asyn>pay comparatively little tax<coord_c_asyn><Cc> - <D><g_indepcl><coordinator>and<coordinator> now we learn that they have millions extra poured into their pensions funds<indepcl><D>. <s465, newspaper articles>`
- (31) `<C><indepcl>This message, <1_prefv><advcl_time_nonfin>once filtered down to the Communist party's grass roots<advcl_time_nonfin><1_prefv>, may ameliorate an obsession with growth at any costs<indepcl><C> - <D><appcl>an obsession that has too often been uncritically applauded by observers in the west<appcl><D>. <s1146, newspaper articles>`

Examples such as these and the ones with the interpolated satellites in the previous section exemplify the rhetorical use of this particular punctuation mark, which, in combination with discourse structure and grammatical realisation, appear to be used as rhetorical devices by writers. With respect to this particular pattern, the analysis shows that it appears not only to be more prominent in English than Dutch, but also specific to the newspaper genre.

When looking at the grammatical realisation of the clausal D-satellites in Dutch, even though the largest group is also formed by adverbial clauses, unlike English, these are almost exclusively finite clauses, with the largest semantic types being reason and comparison. A second large group in Dutch is formed by non-restrictive relative clauses and a third one by reporting clauses, which, similar to English, have a rather subgenre-specific function and occurrence. Moreover, whereas English shows higher frequencies of the dash to separate nuclei from

appended satellites, Dutch shows significantly higher frequencies of colons in this position (see 9.2.5 above). Although this only applies to a limited number of sentences, the frequency difference in itself can be considered representative of a style that has been characterised as a ‘prominence-promoting style’ in 9.2.5 above (cf. Hannay 1997). It should, however, be noted that the frequency difference between the languages is more pronounced for the academic prose genre and leaflets genre and also that a closer analysis of the sentences with colons shows that this punctuation mark performs a similar function in both languages. With respect to the grammatical realisation of the appended satellite the languages also show overlap, with the independent clauses and appositions forming large groups, although English shows a considerably higher frequency of lists in this position. In pinpointing the exact behaviour and function of the colon in this genre, it is interesting that in English there appears to be a difference between the two main sources in this respect, with *the Guardian* showing a higher frequency of colons in this position than *the Daily Mirror*. In Dutch the distribution of sentences with this pattern is divided more evenly across *de Volkskrant* and *de Telegraaf*. Furthermore, both languages show a considerably higher frequency of colons in the editorial subgenre than the news reportage subgenre. Sentence (32) is taken from a Dutch editorial from *de Volkskrant*.

(32) <C><indepcl>Maar de manier waarop de burgemeester nu te werk is gegaan, lijkt op het tegenovergestelde<indepcl><C>: <D><appos\_NP>de weg van de meeste weerstand<appos\_NP><D>. <s2924, newspaper articles>

<C><indepcl>But the way in which the mayor went about this seems like the opposite approach<indepcl><C>: <D><appos\_NP>the line of most resistance<appos\_NP><D>.

The D-satellite can also take the form of a phrase. Similar to the academic prose genre, the largest realisation groups are formed by appositions and adjuncts/PPs in both languages. In English but not in Dutch, this position is also taken up in a few cases by conjuncts that are presented as separate discourse units, with the conjunct *too* accounting for most sentences. An example is presented by the second sentence in (33) below, which, in addition to the adverb *also* contains the conjunct *too*, presumably for extra emphasis.

(33) <zz><adj>Yesterday<adj><zz> <C><indepcl>Mr Hoon admitted British forces will still be in Iraq in a year's time<indepcl><C>. <s20, newspaper articles>

<C><indepcl>He also confessed that the invasion by US and UK troops has led to many foreign terrorists moving in<indepcl><C>, <D><conj>too<conj><D>. <s21, newspaper articles>

### Summary

In comparison with the academic prose genre, this genre may be characterised as being considerably less complex when looking at, for instance, sentence length, frequency of complex beginnings and number of interruptions. In both languages, the overall sentence length is considerably shorter than the academic prose genre, with the Dutch newspaper genre showing significantly shorter sentences than English. This shorter sentence length is in line with the advice given in Dutch prescriptive style guides, which even express preferred sentence length in the number of words that sentences should consist of and link this advice to the text's intended broad readership. This characteristic of the text type might also explain the lower number of sentences in which the nucleus is preceded by a subordinate clause, for which English again shows higher numbers, and the lower number of interruptions. In the sentences that do contain interruptions, the interruption often provides an efficient means of integrating additional, background information, for instance about the subject of the sentence, into the sentence as a whole. The lower frequency of constructions that could be associated with increasing the complexity of sentences might thus be a reflection of the difference in readership of this genre in comparison with, for instance, the academic prose genre.

A closer look at the constructions used in this genre, especially the interaction between discourse structure, grammatical realisation and punctuation, has also provided insight into how various constructions may be used as rhetorical devices. This applies, for instance, to the English use of short sentences as powerful statements or the use of dashes in combination with interpolated satellites or appended clauses, which the writer appears to be using to make certain information stand out. Although further analyses are still needed, the analyses carried out give the impression that the interplay of discourse, grammar and punctuation is not only dependent on language and genre, but even on subgenre and source.

### 9.3.3 Short stories

#### Sentence length

As Section 9.2.1 already indicated, in contrast to the other three genres included in this study, the short stories genre is the only genre for which no significant difference in sentence length was found between English and Dutch. Also in contrast to the other genres, the largest number of sentences are concentrated in the 1-10 words length category, which applies to around 50% of the sentences in both languages. In fact, when looking at the distribution of sentences across the four length categories, it becomes clear that around 80% of the sentences do not exceed 20 words. In terms of sentence length, the short stories genre thus constitutes a text type in which the two languages behave similarly.

An explanation for this high number of short sentences can be found in the fact that a large proportion of the texts in this genre consists of simulated dialogue, which, in turn, can be characterised by its high frequency of clausal and phrasal fragments (cf. 3.5) that often consist of only a few words.

#### Main sentence patterns

Similar to the other genres discussed so far, the largest number of sentences in this genre belong to the C-pattern. However, unlike the other genres, this is not followed by the XC pattern, but by the CX pattern, to which around one quarter of the sentences belong. Despite the fact that the languages show similarities with respect to which sentence patterns dominate, significant frequency differences between the languages can be found for the XC pattern, which again shows higher frequencies in Dutch than in English.

With respect to the make-up of the most frequent pattern, the C-pattern, although this consists of one, uncoordinated nucleus in the vast majority of cases, around one quarter of the sentences consist of coordinated nuclei. In comparison to the other genres, with the exception of the English academic prose genre, the short stories genre shows a relatively high number of coordinated nuclei, the vast majority of which take the form of coordinated independent clauses, similar to the other genres.

Even though the dominance of the uncoordinated C-pattern as opposed to the coordinated C-pattern is similar to the other genres, considerable differences arise when looking at the grammatical realisation of the nucleus in this pattern. Specifically, this is the genre that shows the highest percentage, i.e. around 25%, of

nuclei that do not take the form of an independent clause in both languages. Instead, these are realised as clausal or phrasal fragments. The languages show overlap in the distribution across the main realisation categories, with the largest group being formed by phrasal fragments (around 60%), followed by clausal fragments (around 22%), and, lastly, by a category that contains discourse markers, vocatives and question words (around 12%) (see 3.5.2 for the exact make-up of this latter category). As was already indicated above, the high frequency of precisely these grammatical categories can be explained by the long stretches of simulated dialogue that make up considerable parts of these text types. And, as was explained in detail in Chapters 2 and 3 (e.g. 2.5.2, 3.5), these particular stretches of text posed certain challenges to an annotation system that was primarily designed for analysing the sentence structure of written texts.

### **Beginning of sentences**

In line with the academic prose and newspaper genre, the sentence patterns in which the nucleus is preceded by one satellite, the AC and the ACD subpatterns, are the most frequent ones in this genre in both languages as opposed to sentences beginning with more than two satellites. Similar to the newspaper genre, Dutch again shows a higher frequency of the main XC pattern, with the exception of the subpatterns in which the nucleus is preceded by two or more elements, which are again significantly more frequent in English.

Also in line with the genres discussed so far, the A-satellite in the AC and ACD subpatterns takes the form of a phrase in the vast majority of cases. With respect to the distribution across the main phrasal realisation categories, Dutch shows a higher preference for adjuncts/PPs at the start of a sentence than English, which applies to around 80% of the Dutch sentences, whereas English contains more conjuncts and disjuncts in this position. It should be noted, however, that this genre posed particular challenges in classifying certain discourse units grammatically. Section 3.3.2 dealt with the important step in the categorisation process of determining whether a discourse unit should be classified on the basis of its syntactic form or function, where it was explained that function is seen as giving information about the semantic content or function of a discourse unit. As for the short stories genre, one such challenge was presented by adverbials. In an example such as the following, the word *now* could be classified as a circumstance adverb or adjunct, but also as a discourse marker (cf. Quirk et al.: 526ff, 625; Biber et al. 1999: 1086).

- (34) <Ca><coord\_a>I was forced to agree to having three courts at the far end of the sports hall kept open for badminton<coord\_a><Ca> <Cb><coord\_b>and then he came out with it<coord\_b><Cb>: <s12620, short stories>  
<zz><dm>Now<dm><zz> <C><fragment>what about the rock climbers<fragment><C>? <s12621, short stories>

As the categorisation of these words is highly dependent on contextual factors, for instance on whether they occur in prose sections or simulated dialogue sections of the texts, strict guidelines were at times difficult to set for this genre. One way in which this was dealt with was by extending a number of categories in order to allow for a wider range of structures. The main motivation for this was easing the categorisation process to increase consistent annotation. Differences between the languages with respect to these particular sentence-initial phrases might therefore be a reflection of this annotation difficulty and not necessarily of an actual difference in frequency of the different phrasal categories. A closer inspection of these categories does indeed show that the occurrence of certain adjuncts and conjuncts is genre-specific. Of course this applies to the entire discourse marker, question word and vocative category, but also to conjuncts such as *so*, *then* and *anyway* for English, which also happen to be highly frequent in conversation (cf. Biber et al. 1999: 886). With respect to the latter category, there seems to be a difference between the languages. Specifically, whereas the frequency of particular types of conjuncts appears to be largely determined by genre in English, with, for example, *however* being highly frequent in the academic prose genre and *so* and *then* highly frequent in the short stories genre (also see Biber et al. 1999: 886 for similar findings), the occurrence of particular types of conjuncts in Dutch does not appear to be as strongly influenced by genre as it is in English, with the conjuncts *zo* (*so*), *ook* (*also*) and *daardoor/daarom* (*this is how*, *this is why*) being the most frequent ones irrespective of genre.

Although much lower in number, it is interesting that the clausal A-satellites take the form of an adverbial clause of time in over half of the cases in both languages, thereby forming the largest realisation group. The high frequency of this type of adverbial clause is in line with Biber et al.'s findings for the fiction genre; they account for this by explaining that '[a]dverbial clauses of time are particularly useful for describing beginnings, endings, and the duration of activities, or for describing concurrent events' (1999: 822). A second large group of clauses is formed by adverbial clauses of condition, especially in Dutch, which, when compared to Biber et al.'s findings for English, are not particularly frequent in the

fiction genre, but are frequent in conversation (1999: 821). In line with those findings, the high frequency of this category could then be explained by the large sections of simulated dialogue in the short stories genre.

As for the sentence patterns in which the nucleus is preceded by two or more satellites, it becomes clear that short stories have the lowest number of complex beginnings of all genres. In both languages the A1C subpattern again shows higher frequencies than the ABC pattern, similar to the newspaper genre, with English showing a significantly higher overall frequency of both types. With respect to the A1C subpattern, similar to the high frequency of adjuncts/PPs and adverbial clauses of time in the AC and ACD subpatterns, the A-satellite is almost exclusively realised as an adjunct/PP and the interpolated satellite either as an adjunct/PP or adverbial clause of time or a non-finite clause. A closer inspection of the sentences with the A1C subpattern does indeed show that especially in this genre the beginning of the sentence often seems to locate the event in time or place, which is in line with Biber et al.'s characterisation of the function of these clauses in this genre (see above). Sentences (35) and (36) below are taken from the English and Dutch short stories genre respectively.

(35) <A><adj>Back in the waiting room<adj><A>, <1><advcl\_nonfin>looking round at what Flo's parents would have no hesitation in calling losers<advcl\_nonfin><1>, <C><indepcl>I began to worry again<indepcl><C>. <s12412, short stories>

(36) <A><adj>Later<adj><A>, <1a><advcl\_time><coord\_a\_subcl>toen mijn zus vrolijk en overmoedig geworden was door de drank<coord\_a\_subcl><1a>, <1b><coord\_b\_subcl>en ik loom en een tikje droevig<coord\_b\_subcl><subcl\_advcl\_time><1b>, <C><indepcl>vertrokken we in de speciaal voor ons gemaakte rode feestjurken<indepcl><C>, <D><PP>in het gezelschap van een grote groep vrienden naar een café<PP><D>. <s15716, short stories>

<A><adj>Later<adj><A>, <1a><advcl\_time><coord\_a\_subcl>when my sister had become happy and overconfident with the booze<coord\_a\_subcl><1a>, <1b><coord\_b\_subcl>and I drowsy and a bit sad<coord\_b\_subcl><subcl\_advcl\_time><1b>, <C><indepcl>we left in the red dresses especially made for us<indepcl><C>, <D><PP>in the company of a large group of friends to a café<PP><D>. <s15716, short stories>



## Interruptions

Although English again contains more interruptions than Dutch in this genre, in both languages the short stories genre shows the lowest frequency of interruptions when compared to the other genres. Similar to the other genres, the interruption typically occurs in the nucleus, again making the C1C subpattern the most frequent subpattern, accounting for half over the occurrences in Dutch and as many as over 60% in English. Also in line with the other genres is that the interruptions that occur before or after the finite verb of the nucleus show a fairly even distribution in English, whereas Dutch shows a strong preference for interruptions occurring after the finite verb, thereby adhering to the verb-second principle (see Section 9.2.3 above).

With respect to the grammatical realisation of these interruptions, English shows a slightly stronger preference than Dutch for phrases over clauses, whereas in Dutch this is more evenly divided. When realised as a phrase, Dutch shows a very high frequency of adjuncts/PPs followed by appositions, whereas English shows somewhat more variation in the realisation of these phrasal interpolations, with a number of them taking the form of conjuncts, disjuncts and second coordinates (see section above in newspaper section for this latter category). The high frequency of adjuncts/PPs, particularly in Dutch, could be related to Biber et al.'s study (1999), who found that circumstance adverbials are particularly frequent in the fiction genre and who linked this to their usefulness in contributing to describing and creating an imaginary setting (199: 766, 785). It may therefore not be surprising that the interruptions, which occur predominantly in the descriptive sections of the short stories, also take the form of circumstance adverbials. As for the clausal interruptions, in both languages these predominantly take the form of adverbial clauses, with the non-finite ones showing a higher frequency in English, followed by non-restrictive relative clauses.

Furthermore, a closer look at the interruptions shows that they occur almost exclusively in the non-simulated dialogue parts of the short stories, i.e. in the descriptive parts, and that a number of them appear to be interrupting the flow of discourse in a more pronounced way than others. This applies especially to the interruptions that take the form of a place or time adverbial. The choice of presenting these adverbials as interpolated satellites has the effect of placing prominence on information that typically merely provides background information and is therefore often integrated into the sentence. This subset of interruptions could then be seen to function as focusing devices, often used to mark changes or developments in the story. This occurs in both languages, but appears to be used

even more frequently in Dutch, although this would have to be looked into in much more detail to confirm this. Sentence (37) illustrates this function, taken from the Dutch short stories genre.

(37) <C><indepcl>En Alma richtte, <1\_postfv><pp>vanaf toen<pp><1\_postfv>, al haar aandacht op haar dochtertje en het huishouden<indepcl><C>. <s16655, short stories>

<C><indepcl>And Alma focused, <1\_postfv><pp>from then on<pp><1\_postfv>, all her attention on her daughter and the housework<indepcl><C>.

Furthermore, a subset of interruptions that is particularly frequent in Dutch are exclusively realised as reporting clauses and interrupt reported speech, i.e. the simulated dialogue parts of short stories. These interruptions also form their own sentence pattern, i.e. the C1X subpattern, in which the interpolated satellite actually occurs between the nucleus and the first appended satellite instead of literally interrupting either one. Sentence (38) follows this subpattern and is taken from the Dutch short stories genre.

(38) <C><reportedcl><indepcl>Ik begrijp je niet<indepcl><reportedcl><C>, <1><reportingcl><fragment>zei ik<fragment><reportingcl><1>, <Da><reportedcl><coord\_a>Ariane is geen aanstelster<coord\_a><reportedcl><Da> <Db><reportedcl><coord\_b>en waarom zeg je dat van Erik<coord\_b><reportedcl><Db>? <s15322, short stories>

(<C><reportedcl><indepcl>I don't understand you<indepcl><reportedcl><C>, <1><reportingcl><fragment>I said <fragment><reportingcl><1>, <Da><reportedcl><coord\_a>Ariane is no exaggerator<coord\_a><reportedcl><Da> <Db><reportedcl><coord\_b>and why would you say that of Erik<coord\_b><reportedcl><Db>?)

The interesting thing to note about this subpattern is that it occurs almost exclusively in Dutch. Although the languages show overlap for another common position of reporting clauses, i.e. clause-final position, they behave differently with respect to the occurrence of reporting clauses in clause-medial position, occurring between two discourse units. This is in line with Biber et al.'s analysis of the position of reporting clauses in the English fiction genre, who found that final position is preferred in English, following reported speech (1999: 923). They account for this by suggesting that 'writers seem to prefer to leave the quoted element intact', which they relate to processing ease (*ibid*). The languages thus appear to behave

differently, at least to a certain extent, in the representation of simulated dialogues in short stories.

In short, although the short stories genre shows the lowest frequency of interruptions in both languages when compared to the other genres, English still shows a higher frequency than Dutch. Furthermore, in addition to sharing one of the main functions of interruptions with the other genres, the short stories also contains a number of interruptions that seem to be particular to this genre.

### **Ends of sentences**

As was already indicated above, the short stories genre shows a higher frequency of the CX main pattern in comparison to other genres. Similar to the other genres, the vast majority of these sentences belong to the CD subpattern, but a substantial number of them, between 17% and 19% in both languages, belong to the CDE subpattern. The third most frequent subpattern is the ACD pattern.

Not only with respect to discourse structure, but also with respect to the grammatical realisation of the discourse units, this genre behaves differently when compared to the other genres discussed so far. Similar to the nucleus in the main C-pattern (see above), the percentage of nuclei that take the form of non-independent clauses in the CD and CDE subpatterns is considerably higher when compared to other genres. This applies less strongly to the nuclei in the ACD subpattern, of which only a few percent take the form of a non-independent clause. As for the appended satellites, similar to the other genres, this takes the form of a clause in around two thirds of the cases in the most frequent subpattern, the CD pattern. Differences between the languages can be found in the realisation of this clausal D, which in English predominantly takes the form of an adverbial clause or reporting clause and in Dutch mainly of a reporting clause, followed by an independent clause and only in the third instance as an adverbial clause. Noteworthy is that the largest realisation group in English, the adverbial clause, takes the form of a non-finite clause in almost 75% of the cases, thereby forming a much larger group than in the other genres. Despite the fact that this is a clause type that occurs less frequently in Dutch (see Section 9.2.3 above), for this language too the short stories genre shows the highest frequency in comparison to the other genres, i.e. nearly 25%. In English the vast majority of these non-finite clauses contain a present participle, whereas in Dutch the largest group contain past participles. This high frequency of non-finite clauses is in line with Biber et al.'s findings for supplementary clauses (1999: 820), i.e. non-finite clauses with a

present participle. They suggest that because these particular clauses ‘leave the relation between the adverbial clause and main clause inexplicit’, they are especially useful in fiction, as this ‘requires descriptive details to create an imaginary world, but does not have the need for the explicitness of the more expository registers’ (p. 822). Sentence (39) below contains an example of a D-satellite realised as a non-finite clause, taken from the English short stories genre.

- (39) <C><indepcl>Remarks passed unheard<indepcl><C>, <D><advcl\_nonfin>dissolving in the steam of my tea<advcl\_nonfin><D>. <s11877, short stories>

Furthermore, in both languages the largest semantic class for the finite adverbial clauses is that of time, which is also in line with Biber et al.’s findings (1999: 820). A second large group is formed by adverbial clauses of comparison, and a third group is more language specific, i.e. adverbial clauses of concession in English and reason in Dutch.

As for the second largest realisation group in Dutch, the independent clause, a closer analysis of their function shows that this is varied and dependent on the grammatical realisation of the preceding nucleus and on the type of punctuation mark used to separate the two discourse units. For instance, in a considerable number of cases, the nucleus is realised as a non-independent clause, such as in sentence (40) below.

- (40) <C><fragment\_NP>An un-African experience<fragment\_NP><C>, <D><indepcl>zo heb ik een rit over de wegen van Malawi eens horen noemen door een in een Zambiaanse modderkuil gestrande automobiliste<indepcl><D>. <s15383, short stories>

(<C><fragment\_NP>An un-African experience<fragment\_NP><C>, <D><indepcl>that is how I heard the trip down the roads of Malawi being called by a motorist who got stuck in a Zambian pothole<indepcl><D>.)

As was explained above, this genre presented certain challenges with respect to determining the hierarchical status of discourse units. One such challenge was presented by the situation in which two independent clauses are juxtaposed. The decision then involves determining whether one of the two functions as the nucleus and the other as a satellite or whether both should be analysed as having nuclear status. As was explained in Chapter 2 (2.5.1), the identification of the nuclear unit is determined on the basis of a combination of the type of punctuation

mark that is used to link the units, their syntactic status and their semantic content. With respect to the punctuation marks used to link two independent clauses, these are usually the dash, more frequent in English, the colon, more frequent in Dutch, the semi-colon and the comma. The first two punctuation marks generally prove useful in determining the hierarchical status of units, as these typically link a nucleus and an appended satellite (cf. 7.3 & 7.5). The latter two marks are, however, more ambiguous in this respect and an analysis of the type of punctuation marks used between the independent clauses showed that it is precisely these marks that showed higher frequencies in the short stories genre, which explains why these sentences were considered more difficult to annotate. In comparison to Dutch, English showed significantly higher frequencies of the semi-colon, typically linking *asyndetically* coordinated sentences, i.e. two coordinated nuclei, such as in the following example:

- (41) <Ca><coord\_a\_asyn>The Shuttle terminal at Cheriton slipped  
by<coord\_a\_asyn><Ca>; <Cb><coord\_b\_asyn>the train manager announced that  
they were approaching the Channel<coord\_b\_asyn><Cb>. <s10827, short stories>

Both languages, but especially Dutch, contained a number of cases in which the two independent clauses were linked by commas. This use of the comma is typically referred to as a *comma splice* (cf. 7.6). As Huddleston and Pullum explain, ‘the absence of any grammatical link strongly favours a stronger indicator than the comma to separate the clauses’ and characterise clauses linked by commas as ‘*infelicitous in varying degrees*’ (2002: 1742). In the case of unacceptable comma splices, i.e. those that cannot be referred to as ‘*quasi-syndetic*’ (*ibid*), as they are, for instance, not parallel in structure or do not make use of lexical markers that make explicit the relation between the two clauses, the interpretation of the relation between the clauses can be coordinative or elaborative, which then, in the absence of other indicators, is largely dependent on the context. Consider in this respect the second sentence in (42) below:

- (42) (1) <A><advcl\_time>Nadat ik de merels minutieus had  
bestudeerd<advcl\_time><A>, <Ca><coord\_a>werd ik zelf een vogel<coord\_a><Ca>  
<Cb><coord\_b>en voelde me vrij<coord\_b><Cb>. <s14039, short stories>  
(2) <C><indepcl>Ik bemerkte dat ik anders om me heen begon te  
kijken<indepcl><C>, <D><indepcl>ik voelde me zelfstandig en  
ongebonden<indepcl><D>. <s14040, short stories>

- ((1) <A><advcl\_time>After I had studied the blackboards  
minutely<advcl\_time><A>, <Ca><coord\_a>I became a bird myself<coord\_a><Ca>  
<Cb><coord\_b>and felt free<coord\_b><Cb>.  
(2) <C><indepcl>I noticed that began to look at things in a different  
way<indepcl><C>, <D><indepcl>I felt independent and unattached<indepcl><D>.)

The motivation for classifying the second independent clause in (42.2) as an appended satellite instead of the second coordinate of the preceding nucleus is that the second independent clause is interpreted as providing a further elaboration on the first. An alternative explanation for the higher number of independent clauses linked by commas compared to the other genres might be related to Biber et al.'s characterisation of this genre as not having as strong a need for 'explicitness' when compared to the other genres (1999: 822). In other words, it may be more typical of this genre that the interpretation of the relation between units is left to the reader, even more so than in other genres. However, the relevance for this study of still being able to annotate such sentences consistently is that it affects their classification with respect to the four main sentence patterns.

Although much less frequent in number, in both languages the D-satellite also takes the form of a phrase. In Dutch the largest group is formed by adjuncts/PPs, whereas English shows a high number of vocatives, discourse markers and tag questions in this position, with appositions forming a considerable group in both languages. What is also noteworthy is that the short stories genre shows the widest variety of phrasal realisations when compared to the other genres.

As stated above, although the CD subpattern shows much higher frequencies than the other subpatterns, a substantial number of sentences in this genre belong to the CDE subpattern. Although a more thorough analysis is needed, there is one sentencing pattern for each language that stands out, not necessarily just in terms of frequency, but also in terms of being particular to this genre. For English this is the pattern in which the first appended satellite is realised as a reporting clause, with the second one providing further information about the preceding reporting clause. Although this second appended satellite, the E-satellite, often takes the form of a non-finite clause, it can be realised by other grammatical categories. The pattern has a few occurrences in Dutch, but is considerably more frequent in English. Biber et al. have also identified this pattern as being particularly common to the English fiction genre and state that reporting clauses that occur in initial or final position, with the latter position being much more

common (see above), 'often contain expansions of different kinds' (1999: 924). Sentence (43) below presents an example of this pattern.

- (43) <C><reportedcl><fragment\_NP>Pringles<fragment\_NP><reportedcl><C>?  
<D><reportingcl><fragment>she asks<fragment><reportingcl><D>,  
<E><advcl\_nonfin>offering me the bowl<advcl\_nonfin><E>. <s11020, short stories>

For Dutch the notable pattern is similar to the one presented in sentence (42) above, but this time the independent clause does not follow the nucleus, instead following the D-satellite. In this pattern, the independent clause presents some sort of summary or consequence of the information contained in the preceding two units. Sentence (44) below provides an example of this pattern.

- (44) <C><fragment>Twee keer fout gereden op dezelfde dag<fragment><C>,  
<D><advcl\_time>terwijl het nog middag moet worden<advcl\_time><D>,  
<E><indepcl>dat belooft niet veel goeds<indepcl><E>. <s13640, shor stories>

(<C><fragment>Taking the wrong turn twice on the same day<fragment><C>,  
<D><advcl\_time>while it was not even noon yet<advcl\_time><D>,  
<E><indepcl>that is not very promising<indepcl><E>.)

### Summary

Not only when comparing the languages with each other, but especially when comparing this genre to the other genres, the short stories genre stands out. It can be characterised as having an intermediate status with respect to the distinction between spoken and written language, as considerable portions of these texts consist of simulated dialogue (cf. 2.2). It is precisely this intermediate status that presented certain challenges with respect to the annotation procedure, which mainly affected such decisions as determining the hierarchical status of discourse units and classifying certain grammatical categories unique to this genre.

This is the only genre in which the languages do not show any differences with respect to sentence length and even show overlap in that 80% of the sentences do not exceed 20 words. Another similarity is that this is the genre in which the CX main pattern shows higher frequencies when compared to the other genres. Differences between the languages arise when looking at the discourse patterns, with the XC main pattern again being more frequent in Dutch and the subpatterns in which the nucleus is preceded by more than one element being

more frequent in English. In both languages, however, this is the genre that shows the lowest frequency of complex beginnings. This also applies to the interruptions. The combination of particularly short sentences, a small number of complex beginnings and interruptions give the genre the impression of being much less complex with respect to sentence structure when compared to especially the academic prose genre, but also the newspaper genre. Whether this impression is mainly based on, and should thus be restricted to, the simulated dialogue parts of these text types should be analysed in more detail, as these are sections that can be characterised by their fragmentary nature.

Besides certain genre features that appear to be characteristic of both languages, such as the high frequency of time and place adverbials occurring in various positions in the sentence, the Dutch and English short stories genre each show instances of certain patterns that appear to be more language-specific. For English, an example of this is the very high frequency of non-finite clauses in sentence-final position, also forming a particular pattern in combination with reporting clauses. For Dutch, this is the high frequency of independent clauses in sentence-final position. In both languages it is again an analysis of the interplay between discourse units, grammatical realisation and punctuation mark used that identifies the sentence patterns that are more particular to the two languages under investigation.

### **9.3.4 Public information leaflets**

#### **Sentence length**

In both languages the largest group of sentences can be found in the 11-20 words length category, although this group is larger in Dutch than in English. Rather similar to the short stories genre, in this genre too 85% of the Dutch sentences do not exceed 20 words, the largest concentration of which can be found in the 11-20 words length category instead of the 1-10 words length category. In English around 66% of sentences do not exceed 20 words. In comparison to Dutch, English contains significantly more sentences in the longer length categories, i.e. those containing 21-30 words and 31+ words.

As was suggested in 9.2.1 above, an explanation for the high number of rather short sentences, especially in Dutch, could be linked to the advice given in a considerable number of style guides in Dutch to restrict sentence length to an average of 15-20 words per sentence. As this advice is typically not as explicitly



given in the English style guides, at least not expressed in terms of number of words, this might explain the significant difference in sentence length between the languages. Furthermore, as public information leaflets are typically intended for a wide and varied readership, simplifying sentence structure, for instance with respect to sentence length, may be done deliberately in an attempt to increase the readability of texts. The advice to keep sentences short typically forms one of several items on a list of various suggestions to increase the readability in texts, others being the use of direct questions, personal pronouns, subheadings and lists (cf. van den Boomen & van der Lans 1991: 103, 114-115; Woerkum & Kuiper 1995: 128ff; Huigen 2004: 27ff; Hopster & Tiggeler 2007). With respect to the latter item, when looking at the frequency of punctuation marks that are typically used in lists, i.e. serial commas or the serial use of semi-colons, it should be noted that this is indeed the genre that contains significantly more lists than any of the other genres (see section below on *Ends of sentences*). And many longer sentences do indeed contain lists, especially in Dutch. For instance, a considerable number of the Dutch sentences that belong to the 31+ word category consist of lists, i.e. 44% of the sentences, compared to 27% of the English sentences in this category. This means that on the basis of a mere word count alone, sentences containing lists could easily be perceived as constituting complex sentences, but in a large number of cases these consist of a list, of which the items are presented on different lines. In other words, the use of particular layout devices contributes to the readability of lengthy sentences in the leaflets genre. An example of a long Dutch sentence, i.e. containing 49 words, is presented in (45) below.

- (45) <C><fragment\_comp\_list>Indirecte factoren,  
 <1\_pfv><nonrestr\_relcl><coord\_a\_sub>die op zichzelf niet tot RSI  
 leiden<coord\_a\_sub>, <coord\_b\_sub>maar in combinatie met fysieke factoren  
 kunnen bijdragen aan het ontstaan van  
 RSI<coord\_b\_sub><nonrestr\_relcl><1\_pfv>, zijn<fragment\_comp\_list><C>:  
 - <Da><coord\_a\_phr\_asyn\_list>te weinig hersteltijd<coord\_a\_phr\_asyn\_list><Da>;  
 - <Db><coord\_b\_phr\_asyn\_list>psychische belasting (<1><appos\_NP\_list>hoge  
 werkdruk, werkstress, hoog werktempo, werk met hoge mentale  
 eisen<appos\_NP\_list><1>) <coord\_b\_phr\_asyn\_list><Db>;  
 - <Dc><coord\_c\_phr\_asyn\_list>weinig sociale ondersteuning  
 (<2><appos\_NP\_list>relaties met collega's, chef en  
 management<appos\_NP\_list><2>) <coord\_c\_phr\_asyn\_list><Dc>;  
 - <Dd><coord\_d\_phr\_asyn\_list>kou, tocht<coord\_d\_phr\_asyn\_list><Dd>. <s9790,  
 leaflets>

(<C><fragment\_comp\_list>Indirect factors,  
 <1\_pfv><nonrestr\_relc><coord\_a\_sub>which in themselves cannot lead to  
 RSI<coord\_a\_sub>, <coord\_b\_sub>but in combination with physical factors can  
 contribute to the development of RSI<coord\_b\_sub><nonrestr\_relc><1\_pfv>,  
 are<fragment\_comp\_list><C>:  
 - <Da><coord\_a\_phr\_asyn\_list>too little time to  
 recover<coord\_a\_phr\_asyn\_list><Da>;  
 - <Db><coord\_b\_phr\_asyn\_list>psychological strain(<1><appos\_NP\_list>high work  
 pressure, occupational stress, high work pace, work with high mental  
 effort<appos\_NP\_list><1>) <coord\_b\_phr\_asyn\_list><Db>;  
 - <Dc><coord\_c\_phr\_asyn\_list>little social support (<2><appos\_NP\_list>relations  
 with colleagues, chef and management<appos\_NP\_list><2>)  
 <coord\_c\_phr\_asyn\_list><Dc>;  
 - <Dd><coord\_d\_phr\_asyn\_list>cold, draught<coord\_d\_phr\_asyn\_list><Dd>.)

### Main sentence patterns

In both languages, overall frequencies of the main sentence patterns show remarkable similarities with the newspaper genre. For instance, the frequency of the main C-pattern in the English newspaper genre is 68.5% and in the leaflets genre 65.7%, and for Dutch this is 57.6% compared to 56.9% in the leaflets genre. A possible explanation might be sought in a comparable goal and readership of these genres: sharing information with a wide and varied readership. This is different for the academic prose genre, which typically has a very well-defined, limited and typically highly educated readership, which, as was suggested above, may explain why this genre can in various respects be classified as complex with respect to its sentence structure. Related to their different purpose and readership is, of course, also the different type of information these genres present to the readers and the different lines of argumentation used, all of which are likely to be reflected in sentence structure. As was noted above, the short stories genre can in some respects be considered the 'odd one out', in that it takes an intermediate position on the spoken-written continuum as it contains various features that are typically associated with the spoken language, such as a high frequency of fragments.

Dutch and English also show differences for the main sentence patterns, with Dutch showing a significantly higher frequency of the XC main pattern and English of the C-pattern. With respect to the main C-pattern, in both languages the subpattern in which the nucleus is uncoordinated dominates, but English shows a significantly higher number of coordinated nuclei. Although the differences between the languages in this genre are less pronounced than in the academic

prose genre, they are still considerable. Unlike the English academic prose genre, in the leaflets genre the coordinates predominantly take the form of syndetically coordinated independent clauses. As for the subpattern that does not consist of coordinated nuclei, Dutch shows a significantly higher frequency of nuclei that take the form of non-independent clauses, which applies to almost 10% of the Dutch sentences compared to 3.5% in English. In English, these few cases mainly take the form of phrasal fragments, whereas Dutch also shows a relatively large group of clausal fragments and nuclei that are realised as adverbial clauses, as in (46) below.

- (46) <zz><adj>Soms<adj><zz> <C><indepcl>kan cannabisgebruik leiden tot  
afhankelijkheid<indepcl><C>. <s8818, leaflets>  
<C><advcl\_cond>Met name als iemand al eerder een verslaving heeft  
gehad<advcl\_cond><C>. <s8819, leaflets>
- (<zz><adj>Sometimes<adj><zz> <C><indepcl>cannabis use can lead to  
dependence<indepcl><C>. <s8818, leaflets>  
<C><advcl\_cond>Especially when someone has had an addiction  
before<advcl\_cond><C>. <s8819, leaflets>)

This higher number of fragments in Dutch is in line with Hannay's (1997) findings in his study on sentencing patterns in English and Dutch newspaper texts and fundraising letters. Hannay associates this use with a style that he characterises as a prominence-promoting style (see 9.2.5 above). As the present study has only found a significantly higher frequency of fragments in the Dutch leaflets genre, its frequency of occurrence appears to be dependent on genre. It is interesting to consider the advice given in Dutch style guides with respect to sentence fragments. Although there are some that explicitly discourage the use of sentence fragments or adverbial clauses such as the one in (46) above (eg. Tiggeler 2006: 193), others explain that the occasional use of elliptical clauses can make a particular message more powerful (eg. Renkema 2005: 143).

### Beginning of sentences

As with the other genres, the subpattern in which the nucleus is preceded by one prepended satellite, i.e. the AC and ACD subpatterns, is significantly more frequent in Dutch and the subpatterns in which the nucleus is preceded by two or more satellites are more frequent in English.

With respect to the grammatical realisation of the A-satellite, Dutch behaves similarly to the other genres, with the vast majority of the satellites taking the form of phrases, but English deviates somewhat, as it shows a relatively high frequency of clauses in sentence-initial position, i.e. nearly 40%. As was already suggested in 9.2.3 above, the dominance of phrases instead of clauses especially in Dutch and especially in the Dutch leaflets genre could be linked to the explicit advice given in a wide variety of Dutch style guides to avoid the construction of clauses with a so-called long start (*'lange aanloop'*), in which the main clause is either preceded by a very long phrase or one or more subordinate clauses (cf. Renkema 2005: 85-86; Hermans 2006: 55-57; Tiggeler 2006: 192; Burger & de Jong 2009: 136). As was also suggested above, because one of the solutions to the avoidance of long starts is to split the sentence into two shorter sentences, this might then also be linked to the high frequency of shorter sentences in Dutch.

The phrasal realisation category that dominates in both languages is the group of adjuncts/PPs, followed by conjuncts, with the latter showing significantly higher frequencies in English. Although the absolute frequency is lower, the English leaflets genre contains a similar relative frequency to the English academic prose genre. A closer analysis of these conjuncts shows that the leaflets genre shows a much smaller range of different types than, for instance, the academic prose genre, with a few types showing somewhat higher frequencies. Specifically, similar to the academic prose genre, the most frequent one is *however*, followed by *for instance/for example* and *so* – a conjunct that is particularly frequent in the spoken language (Biber et al. 1999: 886-887). The mix of conjuncts that occur in both the spoken and written language, combined with the high frequency of relatively short sentences and a relatively low frequency of interruptions (see below) might be reflective of a deliberate attempt to make this genre accessible to a wide readership, thereby adopting some features that may be associated with a more informal style. A much more thorough analysis at various levels, including a lexical one, would, however, be necessary to substantiate such a claim.

With respect to the clausal As, in both languages the largest realisation group is formed by adverbials of condition, especially in Dutch. Other than this semantic type, no other type of adverbial clauses shows particularly high frequencies. In both languages the vast majority of these sentence-initial adverbial clauses are finite, also in English, which shows a higher frequency of non-finite clauses in the other genres. As these are all public information leaflets that deal with a wide variety of health-related matters, a relatively frequent sentence construction is to explain what steps people need to take or what people should do

in the case of certain events. An example of this sentence pattern is provided by (47) below.

- (47) <Aa><coord\_a\_subcl>If you snort drugs<coord\_a\_subcl><Aa>, <Ab><coord\_b\_subcl>and you use a note or a straw to snort through<coord\_b\_subcl><Ab>, <C><indepcl>you shouldn't share it with anyone else<indepcl><C>, <D><advcl>as blood can be passed from the inside of a person's nose to another<advcl><D>. <s7848, leaflets>

In this genre English again contains more complex beginnings than Dutch. In addition, this is also the genre in English that contains the second largest percentage of complex beginnings, following the academic prose genre, whereas in Dutch this is the genre that shows the lowest frequency. As has been suggested various times in this chapter, the low frequency in Dutch could not only be explained by its verb-second character, but could also be linked to the advice to keep the start of sentences short, thereby explaining the particularly small number of occurrences in this genre, i.e. 12. As for the type of complex beginning, Dutch shows a preference for the A1C pattern and English shows a more even distribution across the categories of the A1C(X) and the ABC(X) patterns, similar to the academic prose genre. Rather particular to the English leaflets genre is the relatively high frequency of interpolated satellites that take the form of conjuncts in the A1C subpattern, which in Dutch are mainly realised as appositions. Despite the fact that they are very infrequent, similar to the English academic prose genre, the English leaflets genre contains a few instances of sentences in which the nucleus is preceded by three prepended satellites, an example of which is given in (48) below.

- (48) <A><conj>Nevertheless<conj><A>, <B><advcl\_cond>if a person feels upset or hurt by any action<advcl\_cond><B> <zz><conj>then<conj><zz> <C><indepcl>this is a problem which must be addressed<indepcl><C>. <s7506, leaflets>

In certain respects the English leaflets genre thus shows some overlap with the academic prose genre in terms of a relatively high frequency of complex beginnings, sentence-initial adverbial clauses and conjuncts, although a closer analysis shows that the types of adverbial clauses and conjunct are rather genre-specific.

### Interruptions

Although English again contains more interruptions than Dutch, similar to the short stories genre, this genre shows a lower frequency of interruptions when compared to the academic prose genre and newspaper genre in both languages. The interruptions again typically occur in the nucleus, with the C1C pattern being the most frequent subpattern. English shows a fairly even distribution of interruptions occurring before and after the finite verb of the nucleus, whereas Dutch again shows a preference for the position after the finite verb. A reason for the significantly lower frequency of interruptions in Dutch could again be sought in the advice given in the style guides to avoid the use of so-called *tangconstructies* (see 9.2.4 above), information that is typically placed between two elements that belong together, such as a subject and a verb.

Rather similar to the newspaper genre, in this genre the vast majority of interruptions take the form of phrases that are realised as appositions. This applies especially to the interruptions that precede the finite verb of the nucleus, with the ones that occur after the finite verb showing somewhat more variation across the different realisation categories. A closer look at the appositions shows that a large number have the very practical function of giving abbreviations of diseases, explanations or synonyms of health terms, and so on. In other words, they have a clear function in a genre whose main goal is to provide information. An example of this use is provided in sentence (49) below.

- (49) <C><indepcl>HIV (<1\_prefv><appos\_NP>Human Immunodeficiency  
Virus<appos\_NP><1\_prefv>) is a virus which attacks the body's immune  
system<indepcl><C>. <s7025, leaflets>

Moreover, when looking at the type of punctuation mark used to mark off these interruptions, it becomes clear that both languages, but particularly Dutch, show a very high frequency of brackets. As this has been classified as a punctuation mark that in both languages is typically placed around information that could be characterised as extra, less relevant or as an aside, it would mean that the interruption in this genre mainly has the function of providing practical information (see 9.3.2, section on interruptions in newspaper genre).

In addition to this overlap in grammatical realisation between the languages, Dutch shows a relatively high frequency, especially of the interruptions occurring after the finite verb, of interpolated satellites that are realised as premodifiers, i.e. nearly one third. As was suggested above (9.2.4), it could be

argued that these interruptions are less disruptive than others, not only because of their grammatical form, but also because they are surrounded by brackets.

Furthermore, a closer look at the function of interruptions in this genre shows that a subset of them perform a rather genre-specific function in both languages. The information contained in these interruptions typically provides references to other sections in the text or to tables and graphs. Sentence (50) below exemplifies this function.

- (50) <C><indepcl>The requirements of the Noise Regulations at each of these action levels are summarised in Table 1 (<1\_postfv><indepcl>see the centre pages of the leaflet<indepcl><1\_postfv>) <indepcl><C>. <s6809, leaflets>

It should be noted that (50) is representative of one of the annotation difficulties that arose in this genre, which again concerned determining the hierarchical status of the discourse unit, here classified as an interruption, that occurs in sentence-final position. As it does not actually interrupt the nucleus, but follows it, it could also be argued that this is an appended satellite. However, the present analysis as an interpolated satellite is mainly based on a combination of the type of punctuation marks used, i.e. brackets, and the semantic content.

In comparison to the other genres it could thus be argued that the interpolated in the leaflets genre performs a function that can be characterised as being mainly practical. This analysis is based on a combination of the grammatical realisation, i.e. typically appositions, the type of punctuation mark used and the semantic content.

### **Ends of sentences**

In terms of frequency and type, the leaflets genre behaves similarly to the other genres, with the subpattern in which the nucleus is followed by one appended satellite, the CD subpattern, being by far the most frequent one in both languages.

A subtle difference between the languages can be found for the grammatical realisation of the nucleus, which in English consists more often of coordinated nuclei than in Dutch. In both languages, this is also the genre that shows a higher frequency of nuclei that take the form of non-independent clauses when compared to other genres. A closer analysis of these nuclei shows that their main function is to introduce lists. A subset of these types of fragments have

received a particular label, <fragment\_complement\_list>, as the lists often form the complement of the incomplete nuclear sentence, as in (51) below.

- (51) <A><conj>However<conj><A>, <B><pp>after several sleepless nights<pp><B>, <C><fragment\_comp\_list>you will start to find that<fragment\_comp\_list><C>:  
 - <Da><coord\_a\_emb\_asyn\_list>you are tired all the  
 time<coord\_a\_emb\_asyn\_list><Da>  
 - <Db><coord\_b\_emb\_asyn\_list>you drop off during the  
 day<coord\_b\_emb\_asyn\_list><Db>  
 - <Dc><coord\_c\_emb\_asyn\_list>you find it difficult to  
 concentrate<coord\_c\_emb\_asyn\_list><Dc>  
 - <Dd><coord\_d\_emb\_asyn\_list>you find it hard to make  
 decisions<coord\_d\_emb\_asyn\_list><Dd>  
 - <De><coord\_e\_emb\_asyn\_list>you start to feel  
 depressed<coord\_e\_emb\_asyn\_list><De> <s7369, leaflets>

A more pronounced difference between the languages can be found in the grammatical realisation of the D-satellite. In English this takes the form of a clause in two-thirds of the cases, whereas in Dutch the situation is exactly reversed, with the phrases dominating. When realised as a phrase, the largest groups in both languages are again formed by appositions, especially in Dutch. A second large group is formed by adjuncts/PPs. When realised as a clause, the largest groups in both languages are formed by adverbial clauses, particularly in English, and independent clauses. A third large group is formed by non-restrictive relative clauses, especially in Dutch. As for the subordinate clauses, English again shows a higher frequency of the non-finite ones than Dutch and the finite clauses are distributed across the various semantic categories, with the concession type again showing higher frequencies in English and the reason type showing higher frequencies in Dutch. With respect to the second largest realisation group in both languages, the independent clause, it should be noted that in Dutch this is typically separated from the nucleus by means of a colon, a punctuation mark that is particularly frequent in this genre to introduce lists, whereas English uses a dash to separate the two independent clauses. In Dutch the colon appears to be predominantly used as a focus device in these cases. This means that the independent clause following the colon not only serves as a further elaboration or explanation of the information contained in the nucleus, but is also clearly the information that should be focused on in the sentence. An example of this use is provided in (52) below.



- (52) <Ca><coord\_a\_fragment>Gemiddeld<coord\_a\_fragment><Ca>, <Cb><coord\_b>want de slaapbehoefte van mensen varieert<coord\_b><Cb>: <D><indepcl>de een heeft weinig, de ander veel slaap nodig<indepcl><D>. <s8378, leaflets>

(<Ca><coord\_a\_fragment>On average<coord\_a\_fragment><Ca>, <Cb><coord\_b>because the need for sleep of people varies<coord\_b><Cb>: <D><indepcl>some need little, others need much sleep<indepcl><D>.)

The use of the dash in the English leaflets genre appears to be more varied than its use in other genres. Specifically, a closer analysis of its uses shows that in certain cases it appears to be used interchangeably with the colon and the semi-colon, for instance in the case of contrast relations. An example of a dash that occurs between the nucleus and the appended satellite is presented in (53) below:

- (53) <C><indepcl>Don't accept lifts from cruising cabs or touts<indepcl><C> - <D><indepcl>these are illegal<indepcl><D>. <s7787, leaflets>

In comparison to the other genres, the leaflets genre showed considerable variation in the quality of some of these texts with respect to style. Examples of what could be classified as stylistic errors are situations in which the items on a list do not form a parallel structure or cases in which independent clauses are linked by a comma, i.e. comma splices. Although not frequent on the whole, the Dutch leaflets genre does contain most instances of 'unacceptable' comma splices. These are situations in which the clauses that are linked are not parallel in structure nor supported by lexical markers to make the relation between the clauses explicit (cf. 7.6). Sentence (53) below presents an example of a comma splice.

- (53) <du\_Ca><coord\_a\_asyn>Pijn kan variëren tussen licht tintelend gevoel of brandende pijn<coord\_a\_asyn><Ca>, <Cb><coord\_b\_asyn>soms is er sprake van krachtverlies<coord\_b\_asyn><Cb>. <s9814, leaflets>

(<du\_Ca><coord\_a\_asyn>Pain can vary from a light tingling feeling to a burning pain<coord\_a\_asyn><Ca>, <Cb><coord\_b\_asyn>sometimes there is loss of strength<coord\_b\_asyn><Cb>.)

### Summary

The leaflets genre in both languages can in many respects be characterised as a genre that has the clear purpose of presenting information intended for a wide and varied readership in an accessible format. With respect to sentence structure, this is, for instance, reflected in the high frequency of short sentences, the high frequency of lists and the interruptions that have a very practical function of providing additional information, often in the form of appositions, in a very condensed format.

Differences between the languages can, however, again be found in sentence length, with English containing a significantly higher number of sentences in the 31+ word category, of which a relatively small part contains lists. Differences can also be found with respect to the start of sentences, with Dutch again containing not only a higher number of sentences in which the nucleus is preceded by one element, but also a higher number of phrases in this position. English, on the other hand, shows a higher number of clauses in this position and of sentences with a complex beginning. With respect to interruptions, English again contains significantly higher instances than Dutch, but this is a genre in both languages that contains considerably fewer interruptions than the academic prose and newspaper genre. And because of their predominantly practical function, a large number of the interruptions can be considered less intrusive or disruptive to flow of the sentence when compared to the other genres. This applies particularly to the interruptions that are surrounded by brackets, which are again larger in number in Dutch. As for the end of sentences, differences between the languages can mainly be found in grammatical realisation, with English containing a significantly higher number of clauses in this position and Dutch of phrases. It is again the non-finite clauses and the adverbial clauses of concession that show higher frequencies in English. A relatively large group in both languages is also formed by independent clauses in this position, which are often linked to the preceding nucleus by means of a colon, especially in Dutch, or by a dash, mainly in English.

As was noted before, the leaflets genre can thus be characterised as a genre that is very much reader-oriented, which is reflected in its sentence structure and in its layout. Certain devices used to present much information in a clear format, for instance by means of long and complex lists, at times presented certain challenges to the annotation system. A close analysis of these texts also showed variation in style between the differences texts, more so than in other genres.

## 9.4 Conclusion

An analysis of the discourse structure of English and Dutch sentences has shown that they can all be categorised into four main types, i.e. sentences following a C-pattern, an XC pattern, a CX pattern and an XCX pattern. This means that at the most general level of analysis the languages show a considerable amount of overlap. Differences between the languages can, however, be found when analysing the sentences not only at the level of discourse structure, but also at the level of grammar and punctuation, and especially when the interplay between these different levels of analysis is closely examined. Further differences arise when the writing subculture – or genre – is also taken into consideration. Taking into account all these aspects of sentence structure is exactly what an analysis of *sentencing* patterns involves: identifying the decisions writers have made at the levels of discourse, grammar, punctuation and genre to construct orthographic-rhetorical sentences.

When looking at the differences in sentence structure irrespective of genre, it becomes clear that English, on the one hand, contains a significantly larger number of long sentences, a higher frequency of sentences that belong to the basic C-pattern, a higher frequency of interpolated satellites and a different structure at the start of the sentence. Specifically, with respect to this latter aspect, English not only shows a significantly higher number of sentences that start with two or more preposed satellites, but it also contains more sentences that start with a subordinate clause to precede the main clause. A large number of these clauses contain a non-finite verb, which also applies to the clauses occurring in sentence-final position. Dutch, on the other hand, contains a higher number of shorter sentences, a particularly large number of sentences that start with a phrasal element, often an adjunct/PP, which precedes the nucleus, and a higher number of independent clauses and non-restrictive relative clauses in sentence-final position.

It is interesting to note that only a few of these main differences between the languages can be attributed to differences in their respective linguistic systems. For instance, the lower frequency of a particular type of complex beginning in Dutch, the ABC(X) pattern, can be explained by its verb second principle. And the relatively high frequency of non-finite clauses in English, especially in comparison to Dutch, has already been noted by a number of other contrastive studies (cf. Aarts & Wekker 1987: 301; De Moor 1998: 309; Cosme 2007: 279-280; Hannay & Mackenzie 2009: 93-96). This means that the majority of differences are attributable to something other than the linguistic systems of these languages. This study has looked for a possible

explanation in books that can be considered to reflect, at least to a certain extent, the writing cultures of English and Dutch, i.e. style guides and writing manuals. On first impression, the Dutch writing manuals appear to be slightly more prescriptive in nature, especially with respect to a number of key issues that are addressed in a large number of these books. It is interesting to note that a number of the differences between the languages can be related to precisely these key issues in the Dutch style guides. For instance, in an attempt to increase the readability of texts, a number of these books explicitly discourage the use of long sentences, sentences with a long start and sentences with a particular type of interruptive structure. Whether the differences between the languages as identified in the present study can indeed be related to the advice given in style guides would have to be looked at in more detail, but provides an interesting hypothesis. Furthermore, the explicit and deliberate attempt to increase the readability of sentences may be what gives Dutch sentences the impression of being less complex when compared to English sentences.

Writing culture can and has been taken one step further, by including genre in the sentencing analysis as well. Genre has been shown to influence sentence structure to a considerable extent, and often in a different way in the two languages. With respect to the academic prose genre, for instance, this can be characterised as the most complex genre of the four in both languages. This is reflected in its sentence structure by containing a particularly large number of longer sentences, a high number of interpolated satellites and a considerable number of sentences that belong to the XCX pattern. Because the frequencies for each of these aspects are significantly higher in the English academic prose genre, this gives the impression of being even more complex than the Dutch academic prose genre. Differences at the level of sentencing patterns – the interaction between discourse, grammar and punctuation – concern the relatively high frequency of adverbial clauses of concession in sentence-final position in English and relatively high frequency of non-restrictive relative clauses and independent clauses in this position in Dutch, the latter of which are typically introduced by a colon. Perhaps even more interesting than the differences between the languages at the sentencing level is the remarkable overlap at the level of discourse structure. Specifically, with respect to the four main discourse patterns, the languages show no significant differences in frequencies. The dominance of English in the academic prose genre has been put forward as a possible explanation of this overlap, in which case English would influence the discourse structure of Dutch to a certain extent.

In comparison with the academic prose genre, the newspaper genre can be described as being less complex in various respects. The first obvious difference

between the genres is sentence length, with sentences being considerably shorter than in the academic prose. Despite this difference between the genres, the difference in sentence length between the languages still holds, as English sentences are again significantly longer than Dutch sentences. This has again been linked to the explicit advice given in style guides in Dutch, but this time not just in the general style guides, but also the ones designed for the writers in this particular genre, i.e. journalists. As these point to the broad readership of newspaper texts and link this to a focus on readability, this might also explain the lower number of so-called complex beginnings, i.e. either sentences that start with an adverbial clause or sentences that start with multiple satellites, and a lower number of interruptions. Although the frequencies of these complex beginnings and interruptive structures are also lower in English in this genre when compared to the academic prose genre, they are still considerably higher than in Dutch. Moreover, the difference between the languages with respect to interruptions cannot only be described in terms of frequency, but also in terms of function. Whereas a large number of interruptions in Dutch typically provide backgrounded information, in English a considerable number contain foregrounded information, the status of which is reinforced by the interplay between particular types of punctuation marks (dashes), grammatical realisation and semantic content. A closer inspection of not only interruptive structures, but also particularly short sentences and the occurrence of certain grammatical categories shows that differences cannot only be found between the languages and genres, but also between the subgenres and even between different newspapers. Text type, its purpose and readership thus appear to influence sentence structure to a considerable extent.

One of the most characterising features of the short stories genre is that it consists of two very different types of text, i.e. descriptive prose on the one hand and simulated dialogue on the other hand. It is especially the latter type that has a distinct influence on the sentences in this genre, which can be characterised as being particularly short and often fragmentary in nature. In comparison to the other genres, this genre contains considerably shorter sentences and considerably lower frequencies of complex beginnings and interruptions, a high frequency all three of which has been associated with sentence complexity. In this respect, the short stories genre can be classified as less complex with respect to its sentencing patterns than the other genres. Differences between the languages can again be found in the higher number of complex beginnings and interruptions in English, but also in a higher frequency of particular grammatical categories in either language, such as the non-finite clauses in various positions in English and the independent clauses in Dutch. When looking

closely at the use of particular discourse subpatterns, the interplay between grammatical categories, punctuation marks and the way in which, for instance, reported speech is presented subtle differences between the languages arise.

In the leaflets genre the purpose of the text type – presenting information in an accessible format for a broad readership – is very clearly reflected in its sentence structure. Specifically, in both languages the genre can be characterised by its high frequency of short sentences and bullet point lists. Despite these similar aims of the genre in the two languages, differences can again be found in sentence length, the number of complex beginnings and interruptions – all of which are again more frequent in English. As these differences keep recurring in the different genres and as they happen to coincide with the key issues that are addressed in the majority of Dutch style guides, they may be associated with this explicit prescriptive advice. However, again in addition to frequency differences, a closer analysis of the sentences also reveals a difference in function of certain elements, with the interruptions in this genre fulfilling, for instance, mainly a practical function of providing information in a condensed format.

In conclusion, an analysis of the interplay between discourse structure, grammatical realisation and punctuation devices provides insight into and reveals differences in sentencing patterns not only between the two languages, but also between the four genres within these languages. It is only when the interplay between these various aspects is closely examined that subtle differences between the rhetorical purposes of different types of sentences in the two languages can be uncovered.

The key differences between the languages are summarised in Table 2 and the differences between the languages at the level of each of the four genres in Table 3.

**Table 2 Differences between sentencing patterns Dutch and English summarised**

English	Dutch
<ul style="list-style-type: none"> <li>• longer sentences</li> <li>• higher frequency of sentences consisting of only nucleus</li> <li>• higher frequency of adverbial clauses, esp. non-finite, in sentence-initial position</li> <li>• higher frequency of sentences starting with two or more satellites, esp. ABC(X)</li> <li>• higher frequency interpolated satellites</li> <li>• higher frequency adverbial clauses, esp. non-finite, in sentence-final position</li> </ul>	<ul style="list-style-type: none"> <li>• shorter sentences</li> <li>• higher frequency of XC sentences with phrasal satellite at start</li> <li>• higher frequency of adjuncts/PPs in sentence-initial position</li> <li>• lower frequency of sentences starting with more than one satellite, mainly A1C(X)</li> <li>• lower frequency interpolated satellites</li> <li>• higher frequency independent clauses and non-restrictive clauses in sentence-final position</li> </ul>

**Table 3 Differences between sentencing patterns Dutch and English in four genres summarised**

		Academic prose	
		English	Dutch
<b>Length</b>	•	longer sentences (longest sentences in comparison to other genres)	• shorter sentences (longest sentences in comparison to other genres)
<b>Coordination</b>	•	higher frequency coordinated nuclei in main sentence patterns, realised not only as independent clauses, but also as coordinated subordinate/embedded clauses	• lower frequency of coordinated nuclei in main sentence patterns
<b>Sentence beginning</b>	•	higher frequency of adverbial clauses, esp. concession and non-finite, in sentence-initial position	• lower frequency of clauses in sentence-initial position, but condition main type
	•	higher frequency of conjuncts in sentence-initial position	• higher frequency of adjuncts/PPs in sentence-initial position
	•	higher frequency of sentences starting with two or more satellites, esp. ABC(X) (highest frequency complex beginnings in comparison to other genres)	• lower frequency of sentences starting with more than one satellite, mainly A1C(X)
<b>Interpolated satellites</b>	•	higher frequency interpolated satellites (highest frequency in this genre in comparison to other genres)	• lower frequency interpolated satellites (highest frequency in this genre in comparison to other genres)
	•	main function of interruptions is to provide additional information in condensed format, appositions, adjuncts/PPs high frequency. Conjuncts also high frequency.	• main function of interruptions is to provide additional information in condensed format, appositions, adjuncts/PPs and premodifiers high frequency
<b>Sentence end</b>	•	particularly high frequency adverbial clauses in sentence-final position	• more even distribution across three realisation groups: adverbial clauses, independent clauses and non-restrictive clauses in sentence-final position
	•	higher frequency non-finite clauses	• independent clause often linked by colon
	•	high frequency adv cl concession	• lower frequency non-finite clauses
			• frequencies more evenly distributed across adv cl time, reason, concession

Newspaper articles				
	English		Dutch	
<b>Length</b>	•	longer sentences	•	shorter sentences
<b>Main patterns</b>	•	higher frequency main C-pattern	•	higher frequency XC pattern
<b>Sentence beginning</b>	•	higher frequency of adverbial clauses, esp. time and non-finite, in sentence-initial position	•	lower frequency of clauses in sentence-initial position, but condition main type
	•	lower frequency of conjunct in sentence-initial position	•	higher frequency of conjuncts in sentence-initial position
	•	higher frequency of sentences starting with two or more satellites, esp. ABC(X)	•	lower frequency of sentences starting with more than one satellite, mainly A1C(X)
<b>Interpolated satellites</b>	•	higher frequency interpolated satellites	•	lower frequency interpolated satellites
	•	interruption often used as focus device, then typically surrounded by dashes	•	interruption provides mainly backgrounded information, often between brackets and premodifier frequent grammatical realisation
<b>Sentence end</b>	•	particularly high frequency adverbial clauses in sentence-final position	•	lower frequency non-finite clauses
	•	higher frequency non-finite clauses	•	main types of adverbial clauses: reason and comparison
	•	high frequency adv cl concession	•	higher frequency non-restrictive clause
	•	higher frequency appended clauses, typically introduced by dash	•	higher frequency of colon used to separate nucleus and appended satellite
	•	higher frequency of dash used to separate nucleus and appended satellite		



Short stories		
	English	Dutch
<b>Length</b>	<ul style="list-style-type: none"> <li>high frequency short sentences</li> </ul>	<ul style="list-style-type: none"> <li>high frequency short sentences</li> </ul>
<b>Coordination</b>	<ul style="list-style-type: none"> <li>relatively high no of coordinated indep clauses</li> </ul>	<ul style="list-style-type: none"> <li>relatively high no of coordinated indep clauses</li> </ul>
<b>Main patterns</b>	<ul style="list-style-type: none"> <li>high frequency main CX pattern</li> <li>high frequency nucleus taking form of fragments (highest frequency in comparison with other genres)</li> </ul>	<ul style="list-style-type: none"> <li>high frequency CX pattern</li> <li>higher frequency XC pattern</li> <li>high frequency nucleus taking form of fragments (highest frequency in comparison with other genres)</li> </ul>
<b>Sentence beginning</b>	<ul style="list-style-type: none"> <li>mainly adverbial clauses of time in sentence-initial position</li> <li>higher frequency of conjuncts/disjuncts in sentence-initial position</li> <li>higher frequency of sentences starting with two or more satellites, esp. ABC(X) (lowest frequency complex beginnings in comparison with other genres)</li> </ul>	<ul style="list-style-type: none"> <li>mainly adverbial clauses of time&amp;condition in sentence-initial position</li> <li>higher frequency of adjunct/PP in sentence-initial position</li> <li>lower frequency of sentences starting with more than one satellite, mainly A1C(X) (lowest frequency complex beginnings in comparison with other genres)</li> </ul>
<b>Interpolated satellites</b>	<ul style="list-style-type: none"> <li>higher frequency interpolated satellites (lowest frequency in comparison to other genres)</li> <li>occur predominantly in prose sections short stories</li> </ul>	<ul style="list-style-type: none"> <li>lower frequency interpolated satellites</li> <li>occur predominantly in prose sections short stories</li> <li>subset of interruptions used as focusing device (shifts time/place) in prose sections short stories</li> <li>subset of interruptions realised as reporting clause, interrupting reported speech</li> </ul>
<b>Sentence end</b>	<ul style="list-style-type: none"> <li>particularly high frequency adverbial clauses &amp; reporting clause in sentence-final position</li> <li>very high frequency non-finite clauses</li> <li>higher frequency semi-colon to link asynchronously coordinated clauses</li> <li>higher frequency discourse marker/vocative sentence-final position</li> <li>substantial frequency CDE subpattern, in which D often realised as reporting clause and E modifying reporting clause</li> </ul>	<ul style="list-style-type: none"> <li>particularly high frequency reporting clause &amp; non-independent clause in sentence-final position</li> <li>lower frequency non-finite clauses (but highest frequency in comparison with other genres)</li> <li>higher frequency colon and comma (comma splice) to link independent clauses</li> <li>higher frequency adjunct/PP sentence-final position</li> <li>substantial frequency CDE subpattern, in which C often fragment and E independent clause</li> </ul>

<b>Leaflets</b>		
	<b>English</b>	<b>Dutch</b>
<b>Length</b>	<ul style="list-style-type: none"> <li>• high frequency short sentences</li> <li>• higher frequency of longer sentence</li> </ul>	<ul style="list-style-type: none"> <li>• high frequency short sentences</li> <li>• lower frequency of longer sentences</li> </ul>
<b>Coordination</b>	<ul style="list-style-type: none"> <li>• higher frequency of coordinated nuclei, mainly of independent clauses</li> </ul>	<ul style="list-style-type: none"> <li>• lower frequency of coordinated nuclei, mainly of independent clauses</li> </ul>
<b>Main patterns</b>	<ul style="list-style-type: none"> <li>• higher frequency main C-pattern</li> <li>• lower frequency nucleus taking form of non-independent clause</li> </ul>	<ul style="list-style-type: none"> <li>• higher frequency XC pattern</li> <li>• higher frequency nucleus taking form of non-independent clause</li> </ul>
<b>Sentence beginning</b>	<ul style="list-style-type: none"> <li>• higher frequency of clauses in sentence-initial position</li> <li>• higher frequency of conjuncts in sentence-initial position</li> <li>• higher frequency of sentences starting with two or more satellites, esp. ABC(X)</li> </ul>	<ul style="list-style-type: none"> <li>• lower frequency of clauses in sentence-initial position, adv cl condition main type</li> <li>• lower frequency of conjuncts in sentence-initial position</li> <li>• lower frequency of sentences starting with more than one satellite, mainly A1C(X) (lowest frequency complex beginnings in comparison with other genres)</li> </ul>
<b>Interpolated satellites</b>	<ul style="list-style-type: none"> <li>• higher frequency interpolated satellites</li> </ul>	<ul style="list-style-type: none"> <li>• lower frequency interpolated satellites</li> <li>• substantial no of interruptions between brackets</li> <li>• substantial no of interruptions realised as premodifier</li> </ul>
<b>Sentence end</b>	<ul style="list-style-type: none"> <li>• higher frequencies of clause in sentence-final position</li> <li>• higher frequencies of adverbial clauses in sentence-final position</li> <li>• higher frequencies adv cl concession</li> </ul>	<ul style="list-style-type: none"> <li>• higher frequencies of phrase in sentence-final position</li> <li>• higher frequencies of independent clause in sentence-final position, typically introduced by colon</li> <li>• large group formed by non-restrictive clauses</li> <li>• higher frequencies adv cl reason</li> </ul>



# 10. Conclusion

## 10.1 Introduction

This concluding chapter will briefly state the main findings of this study in relation to the research aim as outlined in the Introduction (10.2). This will be followed by a brief reflection on how some of the choices and decisions that were made in order to achieve this aim may have, on the one hand, laid bare some of the limitations of the present approach (10.3), while at the same time identified a number of directions for future research (10.4). The chapter will conclude with a short discussion of some of the practical implications of this study (10.5).

## 10.2 Main findings

It was the main aim of the present study to identify the main sentence patterns in English and Dutch and to determine to what extent the two languages differ from each other in this respect. A related aim was to establish whether potential differences between the languages are attributable to differences in the linguistic systems of English and Dutch or whether these are attributable to other factors, such as the genre in which a sentence is written or to differences between the English and Dutch writing cultures. The sentence has been taken to constitute an orthographic-rhetorical unit, in the construction of which a writer has to make a number of decisions about, for instance, the type of information he wants to put in the sentence, the order in which he wants to present this information and the grammatical form different pieces of information should take. In this view the sentence can thus be seen as a unit that reflects the choices a writer has made about the way in which he wants to package information in linguistic units in such a way that the final product fulfils his communicative aim.

Another reason for comparing the languages at the sentence level was because a consistent, multi-genre contrastive analysis of English and Dutch sentences has not yet been carried out. Instead, most contrastive studies of these two languages focus on an analysis at or below the level of the clause. Another characteristic of the existing studies of these two languages is that they tend to be

more qualitative in nature, with the effect that the observations made are often based on limited datasets or are, for instance, mainly intuition-based. The corpus-based, quantitative studies that do exist, such as the ones by Hannay (1997) and Cosme (2007), have revealed a number of interesting differences between the languages at the sentence level, but both focussed on the analysis of sentences within one particular genre, making it difficult to generalise the findings to the two languages in general. However, these studies did serve as a valuable basis for the present study, which performed a systematic analysis of a considerably larger number of sentences. Specifically, in order to carry out an analysis of English and Dutch sentencing patterns, a corpus consisting of nearly 17,000 sentences was compiled especially for this study. The sentences included in the corpus were taken from four different genres, namely academic prose, newspaper articles, short stories and public information leaflets. All sentences were manually annotated at the levels of discourse and grammar. The discourse analysis involved providing a definition of what constitutes a discourse unit in the written language and how one discourse unit can be distinguished from another (Chapter 2), and a grammatical analysis involved providing each discourse unit with a grammatical label (Chapter 3). The analysis at these levels provides insight into the decisions the writer has made not only at the level of discourse segmentation, but also at the level of the grammar. Moreover, as punctuation marks functioned as one of the main criteria for identifying discourse units boundaries in the discourse segmentation process, an analysis could be made of how a writer uses the interaction between discourse structure, grammatical realisation and punctuation to achieve his communicative aim.

This systematic analysis of a large number of sentences has made it possible to characterise sentences in both languages in a number of ways. One of the main findings is that all sentences in both languages can be reduced to following one of four main discourse patterns, i.e. the pattern in which a sentence consists of a nuclear message alone; the pattern in which this nucleus is preceded by one or more prepended satellites; the pattern in which this nucleus is followed by one or more appended satellites, and the pattern in which the nucleus is both preceded and followed by one or more prepended and appended satellites. In both languages by far the largest group is formed by sentences that consist of a nucleus alone and the smallest group is formed by sentences that follow the pattern in which the nucleus is both preceded and followed by one or more satellites. However, despite this general overall finding and similarity between the languages at the level of discourse structure, interesting differences arise when a more

refined analysis is made of a combination of discourse structure and grammatical realisation. For instance, whereas Dutch contains a higher overall number of sentences that belong to the pattern in which the nucleus is preceded by a prepended satellite, which is typically realised as an adjunct or prepositional phrase, in English this satellite is significantly more frequently realised as an adverbial clause. Another difference with respect to the start of sentences is that English contains a significantly higher number of sentences that start with multiple satellites in sentence-initial position. Differences between the languages, mainly in the preference for particular grammatical categories, can also be found for the satellites occurring in sentence-final position. Specifically, whereas English shows a high number of adverbial clauses in this position, especially non-finite ones, Dutch contains a higher number of independent clauses and non-restrictive clauses in this position. With respect to interpolated satellites, English shows a significantly higher overall frequency. Another overall difference between the languages can be found at the level of sentence length, with English sentences being significantly longer than Dutch sentences.

When relating these main findings to either differences in linguistic systems or to differences in writing culture, it is interesting to note that only a small number of differences between the languages appear to be attributable to differences between the linguistic systems. This applies, for instance, to the higher frequency of multiple satellites in sentence-initial position in English, which can be explained by the fact that Dutch is a verb-second language. This means that the finite verb is typically placed in second position and that no more than one element is thus allowed in sentence-initial position (Haeseryn et al. 1997: 1261, Smits 2002: 22). Another difference that can be related to a difference between the linguistic systems is the significantly higher frequency of non-finite clauses in English, which in other studies have also been shown to be rather characteristic of English (cf. Aarts & Wekker 1987: 301; De Moor 1998: 309; Cosme 2007: 279-280; Hannay & Mackenzie 2009: 93-96). This might mean that most differences between the linguistic systems are either to be found at or below the level of the clause instead of at the sentence level, or it could mean that the model of analysis developed for this study has not been able to capture other differences between the linguistic systems at the sentence level.

However, if the differences identified cannot be related to differences in the respective linguistic systems, it is possible that they are related to differences in the English and Dutch writing cultures, where writing culture is here defined in a narrow sense and seen as being reflected, at least to a certain extent, in the writing

handbooks and style guides of English and Dutch. Although it is acknowledged that the concept of writing culture needs to be further elaborated and refined, the differences in sentence length, frequency of interruptions and clauses in sentence-initial position could then possibly be seen a consequence of the explicit advice that is provided in a wide range of Dutch style guides with respect to precisely these aspects of sentence construction. Specifically, in explaining how the readability of sentences can be increased, the use of long starts (*lange aanloop*), a particular type of interruptive structures (*tangconstructie*), and the construction of lengthy sentences is explicitly discouraged. A number of style guides also formulate this advice in such a way that one is encouraged to write in a careful speech style, imitating the spoken language in certain respects. As the aspects of sentence construction that are explicitly discouraged in many Dutch style guides are precisely the aspects that are associated with increasing sentence complexity, the lack of this explicit advice in English style guides might then serve as a possible explanation for why English sentences on the whole appear to be more complex than Dutch sentences.

However, despite the fact that the differences just outlined do characterise sentences in the two languages in a general way, the discussion needs to be modified, as sentence structure is not only influenced by the language in which a sentence is written, but also to a considerable extent by the particular genre within that language in which it is written. For instance, in both languages the academic prose genre can be characterised as the most complex genre of the four, as it contains the highest number of particularly long sentences, complex beginnings, interruptions and the highest number of sentences that belong to the most complex sentence pattern in which the nucleus is modified by both prepended and appended satellites. However, despite the fact that sentences that belong to the academic prose have a number of characterising features that distinguish them from the sentences in the other genres, the overall differences between the languages also apply at the level of genre. This means that English shows higher frequencies of all the aspects that are associated with sentence complexity.

Furthermore, when looking at sentences in the newspaper genre, English sentences are again significantly longer than Dutch ones, again containing a higher number of complex beginnings and interruptions, but in both languages the frequencies of these different aspects are not only lower when compared to the academic prose genre, but the sentence as a whole appears less complex. This lower degree of complexity could be related to the purpose and readership of this

particular genre, which can be described as the aim of sharing information with a wide and broad readership. Interesting rhetorical differences between the languages can be found when a closer look is taken at, for instance, the function of interruptions in either language or at the use and occurrence of particular focusing devices, the latter of which are more frequently used in Dutch.

The third genre included in this study, the short stories genre, also produces very characteristic sentences. One of the main distinguishing features of the sentences in this genre is that a large number of them are very short and fragmentary in nature, which can to a large extent be explained by the fact that this genre contains considerable sections of simulated dialogue. At the same time, the genre also contains a considerable number of sentences that belong to the more descriptive parts of the stories. Sentences in this genre thus fall into two main categories, those belonging to the simulated dialogue sections and those belonging to the prose sections, either of which have a considerable effect on the structure of sentences. On the whole, the genre contains much lower frequencies of the aspects of sentence construction that are associated with sentence complexity, such as complex beginnings and interruptions. However, in this genre it, too, is the frequency of these aspects that is again greater in English than in Dutch.

The sentences that belong to the fourth and final genre that was included in the corpus, the leaflets genre, also have clearly identifiable features. The purpose of this text type – presenting information in an accessible format for a broad readership – is clearly reflected in its sentence structure. Specifically, this genre contains a high frequency of short sentences and information that is presented in the form of bullet point lists. Despite the fact that the same differences between the languages can again be found for this genre – the higher frequency of longer sentences, complex beginnings and interruptions in English – a closer look at each of these aspects show that they do have a clear genre-specific function. With respect to interruptions, for instance, although more frequent in English, in both languages their main function is to provide additional information in a condensed format, as opposed to, for instance, their more commentary and evaluative function in the English newspaper genre.

In short, an analysis of sentencing patterns in English and Dutch has revealed a number of clear differences between the two languages. On the whole, English contains a higher frequency of various aspects that can clearly be associated with sentence complexity, thereby giving English sentences the characteristic of being more complex than Dutch sentences. Although one of these differences is clearly attributable to a difference in linguistic structure, i.e. the



differences in frequency of complex beginnings, the majority of differences appear to be related to differences between the languages at another level, which could, for instance, be the writing culture of the respective languages. Moreover, this study has also clearly shown that frequency differences alone do not present the whole story when trying to describe sentencing patterns in these two languages. Specifically, it has shown that sentence structure cannot be described or characterised without making explicit reference to the particular genre within which the sentence is written. Genre has a clear effect on sentence structure – one that cannot always be expressed in terms of mere frequency differences alone, but one that especially comes out when looking at the *function* and precise nature of some of these frequency differences. Sentencing patterns in English and Dutch can best be described when a close analysis is made of the interaction between genre, discourse structure, grammatical realisation and use of punctuation.

### 10.3 Limitations

In analysing nearly 17,000 sentences manually, in a systematic and consistent way, a wide range of decisions had to be made, some of which were more practical in nature than others. It is these decisions that may have steered the study in particular directions at times, thereby focusing on certain aspects and not touching upon others. In other words, the particular method adopted and the choices and decisions that were made automatically lead to certain limitations of the present study. These will be briefly described in this section, which is divided into two main parts, the first of which presents the methodological limitations and the second the limitations in relation to the interpretation of the results.

#### 10.3.1 Methodological limitations

One of the questions that could be raised with respect to the model of analysis developed for the present study is whether it was suitable for the analysis of two different languages. As punctuation was, for instance, used as one of the main segmentation criteria and as the English and Dutch punctuational system do not show exact overlap, this led at times to rather ad hoc annotation decisions. Examples of such decisions include the introduction of the zz-label for sentence-initial elements that are not presented as separate punctuation units and the comma-NL label to

mark the punctuation practice in Dutch in which a subject or object is separated from a finite verb by means of a comma, but not used to indicate discourse unit status. Although differences such as these did not lead to any insurmountable annotation difficulties, meaning that they did not jeopardise the consistency of analysis, they did cause the model of analysis to contain certain elements that are clearly ad hoc in nature and would need to be refined and rethought in an updated version.

Another related question that could be raised is whether the choice of one English grammar (i.e. Quirk et al. 1985) as a basis for the grammatical classification of discourse units worked well for two different languages and, notably, for sentences taken from four different genres within those languages. The first part of this question was already addressed in 3.2, which explained that this particular English grammar in comparison to the most comprehensive grammar of Dutch (i.e. Haeseryn et al. 1997) provides a more detailed categorisation of a number of aspects that received focus in this study. However, the second part of the question – whether the model used for grammatical categorisation could be applied to sentences taken from different genres – deserves more attention. Precisely because one of the main outcomes of this study is that genre affects sentence structure to a considerable extent, it is not surprising that the main annotation difficulties that arose were to be found at the level of grammatical categorisation of genre-specific grammatical features. Examples of these are the high frequency of different types of fragments in the short stories genre, especially in the simulated dialogue parts, but also the use of various fragments in the leaflets genre. As the annotation model was not genre-specific, categories such as ‘fragments’ have become container categories – qualitative analyses of which were necessary to gain more insight into their precise make-up. By using a grammar in the design of the annotation model that already pays more attention to genre-specific linguistic features, such as Biber et al.’s (1999) grammar of English, genre-specific grammatical categorisation issues may not only have been easier to deal with, but it could also have created a more refined analysis of a number of categories, such as fragments.

Furthermore, although the development of any model of analysis always involves making certain choices and decisions, there is one decision in particular that requires explicit evaluation. This is the decision to exclude the use and occurrence of embedded clauses from the sentence analysis. Despite the fact that this decision was motivated by the fact that embedded clauses are typically not presented as separate punctuation units and typically do not constitute messages in their own right (cf. 2.4.3 & 3.3.1), the occurrence of embedded clauses has been associated with increasing sentence complexity (cf. Cosme 2007 for the relation between embedded

clauses and sentence complexity and detailed contrastive analysis of these and other clauses in English, French and Dutch newspaper editorials). As one of the main outcomes of this study is that English sentences contain a higher frequency of a number of aspects that are associated with sentence complexity, the inclusion of an analysis of embedded clauses would have proved valuable in this respect.

A final limitation of the present study that deserves attention is the decision to limit the number of genres included in the corpus designed for this study to four. It should be noted that this decision was mainly driven by practical limitations, as the inclusion of more than four genres that would then consist of a sufficient number of sentences to still be considered representative of that particular text type was simply not feasible due to time restrictions. As genre has been shown to play a prominent role in sentence structure, it would, however, be desirable for future studies to further extend the number of genres when performing a sentencing analysis.

### **10.3.2 Limitations in relation to the interpretation of results**

In addition to a number of methodological limitations, a number of limitations also present themselves in relation to the interpretation of results. Because the main aim of the study was to identify and compare the most frequent sentence patterns in the two languages, it is essentially quantitative in nature. This strong focus on a quantitative analysis does, however, mean that a detailed qualitative analysis of a selection of the data lay outside the scope of the present study. One concept that still needs further elaboration and specification is that of writing culture. For the purposes of this study, an impression of the Dutch and English writing cultures was gained from an analysis of a range of style guides of these two languages, but could be further expanded by taking other aspects of writing culture into account as well, such as by conducting a contrastive analysis of a number of Dutch and English writing educational programmes with respect to sentence construction. The notion of writing culture could also be further elaborated by using certain tools of analysis as developed within the field of contrastive rhetoric, or intercultural rhetoric, as it has recently been renamed (cf. Connor 1996; Connor et al. 2008).

Another aspect that a more qualitative analysis of the data would have made possible is determining whether subgenre, such as news reportage vs. editorials in the newspaper genre or journal articles taken from psychology journals vs. history journals in the academic genre, or source, such as different types of academic journals or different types of newspaper, have an effect on sentence

structure as well. This type of analysis was beyond the scope of the present study, but would be interesting to address in future research.

Furthermore, such qualitative analyses could perhaps also provide more insight into the particularly high frequency of, for instance, adverbial clauses of concession in English vs. adverbial clauses of condition and result in Dutch. Specifically, a qualitative analysis in which the notion of writing culture has been further elaborated and which takes subgenre and source into consideration as well might provide more insight into the precise nature and occurrence of particular grammatical categories.

## 10.4 Future research

Because the present study has focused on carrying out a quantitative analysis of sentence structure, a number of the differences identified between the languages and genres raise interesting questions for future research with respect to more qualitative analyses of these quantitative findings. As was suggested above, one of these would be a further elaboration of the concept of writing culture in general and relating some of the main findings that are not attributable to differences between the linguistic systems to this refined concept of writing culture.

An example of a result that would benefit from a more qualitative analysis is the occurrence of interpolated satellites in English. These have not only been shown to be particularly frequent in English in general, but also to have very specific functions in the different genres. Some of these functions have been described and identified, but could still be analysed in much more detail, also by including an analysis of interpolated satellites in other written genres, and perhaps even by comparing them to the occurrence and function of interpolated satellites (typically referred to as parentheticals, cf. Burton-Roberts 2005) in the spoken language.

Another aspect that would be interesting to analyse in more detail in a future study is the question whether certain similarities between the English and Dutch academic prose genre could be related to the increasing dominance of English in this genre. The question would then be whether certain aspects of sentence structure in this genre, such as the high frequency of particularly long sentences, complex beginnings and interruptions, are attributable mainly to the text type and in that sense independent of language, or whether English influences

Dutch sentence structure in certain respects in this genre. One way of looking into this would be by performing a diachronic analysis of academic texts in both languages to determine whether there have been changes over time, especially because English has not always been the dominant language of publication in this genre.

Another suggestion for future research is again related to an extended qualitative analysis of certain research results. It was suggested that one of the reasons why Dutch sentences may appear to be structurally less complex than English sentences could possibly be related to the advice given in Dutch style guides to avoid certain constructions with the aim of increasing the readability of sentences. As some of these style guides formulate this advice in such a way as to make it sound that the purpose of written text is to imitate a careful speech style in some respects, thereby reducing unnecessary sentence complexity, it would be interesting to further look into this and determine whether Dutch written texts do indeed share certain linguistic features that are more typically associated with spoken language.

A final suggestion for future research would be to extend the contrastive analysis of sentence structure by including differences between the linguistic systems of the two languages that have already been identified in previous studies and which are often to be found at or below the clause level. It would then be interesting to see whether some of the differences that are identified in this study could be related to a lower level of analysis, i.e. clause-internal instead of at the sentence level, after all.

## **10.5 Practical implications**

The main findings of this study have practical implications both at the level of the individual languages and at the level of these two languages in contrast. At the level of the individual languages, thus either at the level of English or at the level of Dutch, this study has contributed to gaining insight into how sentences are typically structured in each language, what the most frequent sentence patterns are and how writers combine discourse structure, grammatical realisation and punctuation to construct rhetorically effective sentences. As this is one of the first corpora in which sentences from various genres have been consistently analysed at two main levels, i.e. discourse and grammar and the interaction between these two, the

findings of this study could prove useful for anyone writing in either language. Specifically, if one wants to know what characterises sentences of the Dutch academic prose genre in comparison with the Dutch newspaper genre, it could, for instance, be useful to know that the Dutch academic prose genre allows for lengthy sentences; that it contains hardly any sentence fragments; that a frequently used method to combine clauses is by means of juxtaposing two independent clauses separated by a colon; that the top three of clauses that typically occur in sentence-final position is formed by non-restrictive relative clauses, adverbial clauses and independent clauses, having roughly similar frequencies, and so on. With respect to the newspaper genre, on the other hand, valuable information might be that the majority of sentences do not exceed 20 words; that 30% of all the sentences in this genre start with something other than the subject, often an adverbial of time or place or a conjunct; that this genre allows for a relatively large number of interpolated satellites, as long as they take the form of an apposition and are presented between brackets; that the clauses that occur in sentence-final position are often adverbial clauses of reason or comparison, non-restrictive relative clauses or again independent clauses that are separated from the preceding clause by means of colon, and so on. In other words, this study has made implicit, genre-specific knowledge explicit, providing, as it were, a checklist of the main features and characteristics of sentence structure in four genres of Dutch and four genres of English.

Apart from the practical implications and applications at the level of the individual languages, the results of this study also have practical implications when looking at the languages in contrast. Specifically, as was already suggested in the Introduction to this research, the results of a study such as the present one could be relevant for especially a Dutch society in which English is playing a more and more prominent role. Because of this ongoing development, a considerable number of Dutch people with a wide range of professions are confronted with English in their work. This, in turn, has the effect that English language professionals are also faced with an increased demand for extensive and detailed knowledge of various varieties of English. Specifically, detailed knowledge about differences between the languages at the sentence level could prove useful for the language professionals who deal with English texts written by Dutch writers (editors); the ones who translate Dutch texts into English or vice versa (translators), or the ones who instruct Dutch learners of English about the differences between these two languages (teachers). Despite the fact that a number of professionals working in this field might already be familiar with at least some of these results, at

least at the level of introspection, this study provides a systematic, consistent overview of differences between the languages at the sentence level for four different genres.

## References

- Aarts, B. (2007). *Syntactic gradience: The nature of grammatical indeterminacy*. Oxford: Oxford University Press.
- Aarts, F.G.A.M. & Wekker, H.C. (1987). *A contrastive grammar of English and Dutch: Contrastieve grammatica Engels/Nederlands*. Leiden: Martinus Nijhoff.
- Acuña-Fariña, J.C. (1996). *The puzzle of apposition: On so-called appositive structures in English*. Santiago de Compostela: Servicio de Publicacións de la Universidad de Santiago.
- Acuña-Fariña, J.C. (1999). On apposition. *English Language and Linguistics*, 3(1), 59-81.
- Altenberg, B. & Granger, S. (2002). *Lexis in contrast: Corpus-based approaches*. Amsterdam & Philadelphia: Benjamins.
- Anson, C.M. & Schwegler, R.A. (1998). *The Longman handbook for writers and readers*. New York: Longman.
- Aston, G. & Burnard, L. (1998). *The BNC Handbook*. Edinburgh: Edinburgh University Press.
- Austin, T. (2003). *The Times style and usage guide*. London: Harper.
- Beaugrande, R. de (1999). Sentence first, verdict afterwards: On the remarkable career of the "sentence". *Word*, 50(1), 1-31.
- Berg, J. de (2006). *Trouw schrijfboek*. Amsterdam: Dagblad Trouw/Muntinga Pockets.
- Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, 62, 384-414.
- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243-257.
- Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S. & Finnegan, E. (1999). *The Longman grammar of spoken and written English*. London: Longman.



- Boomen, M. van den & Lans, J. van der (1991). *Schrijfwerk: Een handleiding voor de non-profitsector*. Houten/Antwerpen: Bohn Stafleu Van Loghum.
- Brazil, A. (1995). *A grammar of speech*. Oxford: Oxford University Press.
- Brooks, C. & Warren, R. (1979). *Modern rhetoric*. New York: Harcourt.
- Burger, P. & Jong, J. de (2009). *Handboek stijl: Adviezen voor aantrekkelijk schrijven*. Groningen: Noordhoff Uitgevers.
- Burrough-Boenisch, J. (2004). *Righting English that's gone Dutch*. Voorburg: Kemper Conseil Publishing.
- Burton-Roberts, N. (1975). Nominal apposition. *Foundations of Language*, 13, 391-419.
- Burton-Roberts, N. (1999). Apposition. In E. K. Brown & J. Miller (Eds), *The Concise Encyclopaedia of syntactic categories* (pp. 25-30). Amsterdam: Elsevier
- Burton-Roberts, N., 2005. Parentheticals. In E.K. Brown (Ed.), *Encyclopaedia of Language and Linguistics* (pp. 179-182). Oxford: Elsevier.
- Carlson, L. & Marcu, D. (2001). *Discourse tagging reference manual*. Retrieved from: <http://www.isis.edu/~marcu/discourse/>
- Carlson, L., Marcu, D. & Okurowski, M. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. van Kuppevelt & R. Smith (Eds), *Current directions in discourse and dialogue* (pp. 85-112). Dordrecht: Kluwer Academic Publishers.
- Carter, R. & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide: Spoken and written English: Grammar and usage*. Cambridge: Cambridge University Press.
- Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 35-53). Norwood, NJ: Ablex.
- Chafe, W. & Danielewicz, J. (1987). *Properties of spoken and written Language* (Technical Report No. 5 of the Center for the Study of Writing). University of California, Berkeley: Berkeley.
- Chafe, W. (1988). Punctuation and the prosody of written language. *Written Communication*, 5, 396-426.

- Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: The University of Chicago Press.
- Connor, U. (1996). *Contrastive rhetoric: cross-cultural aspects of second language writing*. Cambridge: Cambridge University Press.
- Connor, U., Nagelhout, E. & Rozycki, W.V. (2008). *Contrastive rhetoric: reaching to intercultural rhetoric*. Amsterdam/Philadelphia: Benjamins.
- Cosme, C. (2007). *Clause linking across languages: A corpus-based study of coordination and subordination in English, French and Dutch*. Unpublished doctoral dissertation, University of Louvain, Belgium.
- Crystal, D.C. & Davy, D. (1975). *Advanced conversational English*. London: Longman.
- Dale, R. (1991). Exploring the role of punctuation in the signalling of discourse structure. In *Proceedings of a workshop on text representation and domain modelling: Ideas from linguistics and AI* (pp.110-120). Technical University of Berlin: Berlin.
- Degand, L. & Simon, A.C. (2005). Minimal Discourse Units: Can we define them, and why should we? In M. Aurnague, M. Bras, A. Le Draoulec & L. Vieu (Eds). *Proceedings of SEM-05: Connectors, discourse framing and discourse structure: from corpus-based and experimental analyses to discourse theories* (pp. 65-74). Biarritz.
- Degand, L. & Simon, A.C. (2009). Mapping prosody and syntax as a strategic choice. In A. Wichmann, D. Barth-Weingarten & N. Dehé (Eds). *Where Prosody Meets Pragmatics* (pp. 81-107). Bangalore: Emerald.
- Dik, S.C. (1997). *The theory of Functional Grammar: Part 2: Complex and derived constructions*. Berlin: Mouton de Gruyter.
- Doeve, R.E. & Onrust, M.G. (1992). *Helder schrijven: praktische adviezen voor duidelijk taalgebruik*. Amsterdam: Prometheus.
- Donkers, H. & Willems, J. (2002) *Journalistiek schrijven voor krant en vakblad*. Bussum: Coutinho.
- Downing, A. & Locke, P. (2002). *A University course in English grammar*. Hertfordshire: Prentice Hall.
- Flesch, R.F. (1949). *The art of readable writing*. New York: Harper.

- Field, A.P. (2005). *Discovering statistics using SPSS: and sex and drugs and rock 'n' roll*. London: Sage.
- Ford, C., Fox, B. & Thompson, S. (1996). Practices in the construction of turns: The 'TCU' revisited. *Pragmatics* 6(3), 427-454.
- Ford, C. & Thompson, S. (1996). Interactional units in conversation: syntactic, intonational and pragmatic resources. In E. Ochs, E.A. Schegloff & S. Thompson (Eds), *Interaction and Grammar* (pp. 138-184). Cambridge: Cambridge University Press.
- Ford, C., Fox, B. & Thompson, S. (2002). Constituency and the grammar of turn increments. In C. Ford, B. Fox & S. Thompson (Eds), *The language of turn and sequence* (pp. 14-38). Oxford: Oxford University Press.
- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics* 21(3), 354-375.
- Fowler, H.R. & Aaron, J.E. (2010). *The Little, Brown Handbook*. New York: Longman.
- Gessel, H. van, Hulsbosch, J.K., Hurdeman, H. (Eds). (2006). *De Volkskrant Stijlboek*. Den Haag: SDU.
- Granger, S. (2003). The corpus approach: A common way forward for Contrastive Linguistics and Translation Studies. In S. Granger, J. Lerot & S. Petch-Tyson (Eds), *Corpus-based approaches to Contrastive Linguistics and Translation Studies* (pp. 17-30). Amsterdam & Atlanta: Rodopi.
- Haeseryn, W., Romijn, K, Geerts, G., Rooij, J. de & Toorn, M.C. van den (1997). *Algemene Nederlandse Spraakkunst*. Groningen: Martinus Nijhoff.
- Halliday, M.A.K. (1989). *Spoken and written language*. Oxford: Oxford University Press.
- Halliday, M.A.K. (1994). *An introduction to Functional Grammar*. London: Arnold.
- Halliday, M.A.K. & Matthiessen, C.M.I.M. (2004). *An introduction to Functional Grammar*. London: Arnold.
- Hannay, M. (1997). Sentencing in English and Dutch. In J. Aarts, I. de Mönnink & H. Wekker (Eds), *Studies in English Language and Teaching* (pp. 231-256). Amsterdam: Rodopi.

- Hannay, M. & Kroon, C.H.M. (2005). Acts and the relation between grammar and discourse. *Functions of Language*, 12(1), 87-124.
- Hannay, M. & Keijzer, M.E. (2005). A discourse treatment of English non-restrictive nominal appositions in Functional Discourse Grammar. In J.L. Mackenzie & M.L.Á. Gómez-González (Eds), *Studies in Functional Discourse Grammar* (pp. 159-194). Bern: Peter Lang.
- Hannay, M. & Mackenzie, J.L. (2009) *Effective writing in English: A sourcebook*. Bussum: Coutinho.
- Hengeveld, K. (2004). The architecture of Functional Discourse Grammar. In J.L. Mackenzie & M.L.Á. Gómez-González (Eds), *A new architecture for Functional Grammar* (pp. 1-21). Berlin: Mouton de Gruyter.
- Hermans, M. (2006). *Schrijven met effect: Stijlcursus doeltreffend formuleren*. Bussum: Coutinho.
- Hicks, W. (2007). *English for journalists*. London: Routledge.
- Hicks, W. (2009). *The basics of English usage*. London: Routledge.
- Hopster, J. & Tiggeler, E. (2007). *Checklist folders en brochures*. Unpublished course material. Amsterdam: Taalcentrum-VU.
- Horst, P.J. van der (1997). *Leesbaar schrijven voor iedereen*. Den Haag: SDU uitgevers/ Standaard Uitgeverij.
- Horst, P.J. van der (1999). *Stijlwijzer: praktische handleiding voor leesbaar schrijven*. Den Haag: SDU.
- Huddleston, R. & Pullum, G. (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
- Huigen, M. (2004). *Zelf brochures schrijven*. Alphen aan den Rijn: Kluwer.
- Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Hunt, K. (1966). Recent measures in syntactic development. *Elementary English*, 43, 732-739.
- Jespersen, O. (1961). *A modern English grammar on historical principles*. London: Allen & Unwin.

- Jones, B. (1996). *What is the point? A (computational) theory of punctuation*. ICCS PhD Thesis Collection, The University of Edinburgh.
- Johansson, S. & Hasselgård, H. (1999). Corpora and cross-linguistic research in the Nordic countries. *Le Langage et l'Homme*, 34(1), 145-162.
- Kane, T.S. (1988). *The new Oxford guide to writing*. Oxford: Oxford University Press.
- Kaplan, R.B. (1966). Cultural thought patterns in intercultural education. *Language Learning*, 16, 1-20.
- Kay, V. (1990). *The essential feature: Writing for magazines and newspapers*. New York: Columbia University Press.
- Kennedy, G. (1998). *An Introduction to corpus linguistics*. London & New York: Longman.
- Koenen, L. & Smits, R. (2004). *Handboek Nederlands*. Bijleveld: Utrecht.
- Kroll, B. (1977). Combining ideas in written and spoken English: A look at subordination and co-ordination. In E. Ochs & T. Bennett (Eds), *Discourse across time and space*. SCOPIL (Southern California Occasional Papers in Linguistics) No. 5.
- Langacker, R. W. (2001). Discourse in cognitive grammar. *Cognitive Linguistics*, 12(2), 143-188.
- Lamers, H.A.J.M. (1986). *Hoe schrijf ik een wetenschappelijke tekst? Een handleiding om scripties, onderzoeksverslagen, dissertaties en literatuurrapporten te schrijven*. Bussum: Coutinho.
- Leech, G. (1991). The state of the art in corpus linguistics. In K. Aimer & A. Altenberg (Eds), *English Corpus Linguistics* (pp. 8-29). London: Longman.
- Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech & A. McEnery (Eds), *Corpus Annotation* (pp. 1-18). London: Longman.
- Leech, G. (2004). Adding linguistic annotation. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 17-29). Oxford: Oxford University Press.
- Loban, W. (1966). *The language of elementary school children* (Research Report No. 1). Champaign, Ill: National Council of Teachers of English.

- Mackenzie, J.L. (1997). *Principles and pitfalls of English grammar*. Bussum: Coutinho.
- Mann, W., & Thompson, S. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3), 243-281.
- Marsh, D. (Ed.). (2007). *Guardian style*. London: Guardian Books.
- Matthews, P.H. (1981). *Syntax*. Cambridge: Cambridge University Press.
- McEnery, T. & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London & New York: Routledge.
- Meyer, C.F. (1992). *Apposition in contemporary English*. Cambridge: Cambridge University Press.
- Meyer, C.F. (2002). *English Corpus Linguistics: An introduction*. Cambridge: Cambridge University Press.
- Moor, W. de (1998). *A contrastive reference grammar English /Dutch*. Kappellen: Uitgeverij Pelckmans.
- Nederhoed, P. (2000). *Helder rapporteren: Een handleiding voor het opzetten en schrijven van rapporten, scripties, nota's en artikelen*. Houten/Diegem: Bohn Stafleu Van Loghum.
- Nunberg, G. (1990). *The linguistics of punctuation*. Chicago: University of Chicago Press.
- Oakes, M.P. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.
- Onrust, M., Verhagen, A. & Doeve, R. (1993). *Formuleren*. Houten/Zaventhem: Bohn Stafleu Van Loghum.
- Pape, S. & Featherstone, S. (2005). *Newspaper journalism: A practical introduction*. London: Sage.
- Permentier, L. (2003). *Stijlboek: Onmisbaar voor wie helder wil schrijven*. Roeselare: Roularta Books.
- Peters, P. (2004). *The Cambridge guide to English usage*. Cambridge: Cambridge University Press.

- Pica, T., Halliday, L., Lewis, N. & Morgenthaler, L. (1989). Comprehensible outputs as an outcome of linguistic demands on the learner. *Studies in Second Language Acquisition*, 11(1), 63-90.
- Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.
- Renkema, J. (2005). *Schrijfwijzer*. Den Haag: SDU Uitgevers.
- Ritter, R.M. (2002). *The Oxford guide to style. The style bible for all writers, editors, and publishers*. Oxford: Oxford University Press.
- Sacks, H., Schegloff, E. & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696-735.
- Sato, C. (1988). Origins of complex syntax. *Studies in Second Language Acquisition*, 10(3), 371-95.
- Selting, M. (2000). The construction of units in conversational talk. *Language in Society* 29(4), 477-517.
- Siepmann, D., Gallagher, J.D., Hannay, M. & Mackenzie, J.L. (2008). *Writing in English: A guide for advanced learners*. Tübingen & Basel: Francke.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sinclair, J. (1992). *English Usage*. London: Collins.
- Sinclair, J. (2004a). *Trust the text: Language, corpus and discourse*. London: Routledge.
- Sinclair, J. (2004b). Corpus and text: basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1-16). Oxford: Oxford University Press.
- Sinclair, J. (2004c). Appendix: How to build a corpus. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 79-83). Oxford: Oxford University Press.
- Smits, A.M. (2002). *How writers begin their sentences: Complex beginnings in native and learner English*. LOT Dissertation Series 67. Utrecht: LOT.
- Spek, E. van der (1997). *Helder en pakkend schrijven: schrijven zonder problemen*. Houten: Spectrum

- Steen, G.J. (2005). Basic discourse acts: Towards a psychological theory of discourse segmentation. In F. Ruiz de Mendoza Ibáñez & M. S. Peña Cervel (Eds), *Cognitive linguistics: Internal dynamics and interdisciplinary interaction* (pp. 283-312). Berlin: Mouton de Gruyter.
- Stenström, A.B. (1994). *An introduction to spoken interaction*. London: Longman.
- Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics*. Boston: Allyn & Bacon.
- Tannen, D. (1982). *Spoken and written language: Exploring orality and literacy*. Norwood, N.J.: Ablex.
- Tiggeler, E. (2006). *Vraagbaak Nederlands: Van spelling tot stijl: Snel een helder antwoord op praktijkvragen over taal*. Den Haag: SDU Uitgevers.
- Trask, R.L. (2001). *Mind the gaffe: The Penguin guide to common errors in English*. London: Penguin.
- Verhagen, A. (1991). Oud en nieuw in interpunctie. In J. Noordegraaf & R. Zemel (Eds), *Accidentia. Taal- en Letteroefeningen voor Jan Knol* (pp. 77-86). Amsterdam: Stichting Neerlandistiek-VU.
- Williams, J. M. (1990). *Style: Towards clarity and grace*. Chicago: University of Chicago.
- Woerkum, C.M.J. & Kuiper, D. (1995). *Voorlichtingskunde: Een inleiding*. Houten: Bohn Stafleu Van Loghum.
- Wynne, M. (Ed.) (2004). *Developing linguistic corpora: A guide to good practice*. Oxford: Oxford University Press.



# Samenvatting in het Nederlands (Summary in Dutch)

## Zinspatronen in het Engels en het Nederlands: een contrastieve corpus analyse

### Introductie

Door de ontwikkeling van het Engels tot wereldtaal moeten Nederlanders met verschillende professionele achtergronden zich steeds vaker goed in het Engels kunnen uitdrukken, zowel mondeling als schriftelijk. Daarbij lijkt de vraag hoe deze nauw verwante talen zich tot elkaar verhouden, ofwel wat de overeenkomsten en wat de verschillen zijn, steeds vaker gesteld te worden. Tot voor kort was het antwoord op deze vraag vaak gebaseerd op intuïties: taalkundigen beschreven de verschillen tussen het Engels en het Nederlands meestal op basis van wat zij meenden waar te nemen (zie Cosme 2007). Hoewel waarnemingen op basis van intuïties zeker waardevol zijn, heeft de komst van corpusonderzoek het mogelijk gemaakt om dit type vragen vanuit een meer kwantitatieve benadering te beantwoorden. Het doel van het huidige onderzoek is dan ook om verschillen tussen het Engels en het Nederlands in kaart te brengen door middel van de methode van corpusonderzoek, waarbij de studie zich richt op de verschillen tussen deze talen op zinsniveau. Voor het onderzoek zijn bijna 17.000 Engelse en Nederlandse zinnen handmatig geanalyseerd, waarbij de zinnen afkomstig zijn uit vier verschillende genres: academisch proza, krantenberichten, korte verhalen en informatiefolders.

De reden om de zin als eenheid van analyse te selecteren komt voort uit de gedachte dat er veel gebeurt op zinsniveau. Zo moet de schrijver verschillende beslissingen nemen op het moment dat hij een zin opstelt. Hoofdstuk 1 begint met het geven van een definitie van de zin, die in deze studie wordt beschouwd als een orthografisch-retorische eenheid (zie Siepmann et al. 2008), en beschrijft vervolgens in meer detail wat voor type beslissingen een schrijver moet nemen bij het opstellen van een zin. Zo moet hij bijvoorbeeld een beslissing nemen over het type informatie dat hij in een zin verpakt, de hoeveelheid informatie, maar ook welke informatie de kernboodschap bevat en welke meer achtergrondinformatie geeft; ook moet hij beslissen in welke grammaticale vorm hij deze verschillende stukjes informatie wil

gieten en bepalen in hoeverre de zin past in het genre waarin hij de zin schrijft. In een eerdere, kleinschalige corpusstudie naar de verschillen in zinsstructuur tussen het Engels en het Nederlands, heeft Hannay (1997) dit beslissingproces aangeduid met de term *sentencing*. Dit onderzoek toonde aan dat er op zinsniveau interessante verschillen zijn te vinden tussen beide talen, waarbij hij het Engels typeerde als een taal met een zogenaamde *combining style*, door de hogere frequentie van langere zinnen die door verschillende manieren van nevenschikking en onderschikking tot stand zijn gekomen, en het Nederlands als een taal met een zogenaamde *chopping style*, vanwege juist de hogere frequentie van relatief korte zinnen en zinsfragmenten. In de huidige studie werd het onderzoek van Hannay als uitgangspunt genomen, maar is zijn analysemodel verder uitgebreid door naar een veel groter aantal zinnen te kijken, afkomstig uit vier verschillende tekstsoorten. Naast het in kaart brengen van verschillen in zinsstructuur en zinspatronen in beide talen, heeft de huidige studie tot doel na te gaan in welke mate deze verschillen zijn terug te brengen tot verschillen tussen beide taalsystemen, verschillen tussen beide taalculturen, of tot een *interactie* tussen taalsysteem en taalcultuur.

### **Analysemethode**

Om de zinsstructuur in beide talen structureel in kaart te brengen is een analysemodel ontwikkeld dat zich zowel richt op de discoursestructuur van zinnen als de grammaticale realisatie van de verschillende zinsdelen. Hoofdstuk 2 begint met een overzicht van verschillende theoretische en praktische benaderingen met betrekking tot het beschrijven van een discourse-eenheid – of een informatie-eenheid – in zowel de gesproken als de geschreven taal. Dit overzicht dient enerzijds om inzicht te geven in de complexiteit van een sluitende definitie van een discourse-eenheid en om de voorgestelde criteria te evalueren waarmee deze eenheden geïdentificeerd, en daarmee van elkaar ontscheiden, kunnen worden. Anderzijds heeft het overzicht de functie te bepalen welke van de verschillende benaderingen de basis kan vormen voor de eenheid van analyse die kan worden toegepast om de zinnen in dit onderzoek op consequente wijze te kunnen segmenteren in discourse eenheden. Het overzicht heeft geleid tot een nieuwe analyse-eenheid, de *Sentence Information Unit* (SIU), die grotendeels gebaseerd is op de *Punctuation Unit* zoals beschreven door Chafe (1988), Hannay (1997) en Hannay en Kroon (2005). In overeenstemming met deze benaderingen worden interpunctietekens tot op zekere hoogte opgevat als grensmarkeringen van eenheden die de schrijver bewust geplaatst heeft en daardoor inzicht kunnen geven in zijn discourse-intenties. Naast interpunctie als middel om

eenheden van elkaar te onderscheiden, zijn hiervoor ook syntactische en semantische criteria toegepast. Een verdere functie van het identificeren van *Sentence Information Units* is om inzicht te geven in hoe de verschillende discourse-eenheden zich hiërarchisch tot elkaar verhouden, waarbij een onderscheid gemaakt wordt tussen eenheden die kern- of satellietstatus hebben. Van de eenheden met satellietstatus wordt ook genoteerd of deze voorafgaan aan de kerneenheid, deze onderbreken of erop volgen. Naast de definitie van de SIU geeft het hoofdstuk ook een overzicht van de segmentatierichtlijnen, waarbij de nadruk ligt op de segmentatieproblemen en de oplossingen hiervoor. Het hoofdstuk geeft de theoretische fundering van de SIU, maar licht ook nader toe dat bepaalde richtlijnen meer praktisch gefundeerd zijn met het oog op een consequente analyse van de discourse structuur binnen zinnen te bewerkstelligen.

Om een *sentencing* analyse van zinnen te kunnen uitvoeren, zijn naast de discoursestructuur van zinnen ook hun grammaticale kenmerken in kaart gebracht. Hoofdstuk 3 beschrijft hoe aan iedere discourse-eenheid ook een grammaticaal label is toegekend en concentreert zich met name op de moeilijkheden bij het proces van grammaticale classificatie van discourse-eenheden. Dat de processen van discoursesegmentatie en grammaticale classificatie in aparte hoofdstukken worden behandeld betekent niet dat deze processen ook in de analyse zuiver gescheiden zijn. Hoewel de discoursesegmentatie in principe vooraf ging aan de grammaticale classificatie, wordt in beide hoofdstukken opgemerkt dat er vaak juist een interactie plaatsvindt tussen beide analyseniveaus, waarbij discoursekenmerken de grammaticale classificatie beïnvloeden en vice versa. De thema's die bij het proces van grammaticale classificatie aan de orde komen zijn, onder andere, vaststellen op basis van welke criteria een discourse-eenheid het grammaticale label 'zin' (*clause*) toegekend krijgt, wanneer dit het label 'zinsdeel' (*non-clause*) toegekend krijgt en hoe er in het annotatieproces consequent omgegaan kan worden met de glijdende schaal van zinstatus (zie Aarts 2007). Een ander thema betreft de vraag of een discourse-eenheid geclassificeerd moet worden op basis van zijn grammaticale vorm, zijn grammaticale functie of zijn semantische functie en, wederom, hoe het onderscheid tussen deze verschillende mogelijkheden consequent kan worden toegepast. Het kunnen maken van een structureel onderscheid tussen discourse-eenheden die aan elkaar gekoppeld zijn door onderschikking of nevenschikking vormt een ander thema, waarbij de ambigue status van appositie speciale aandacht heeft gekregen. Tot slot komt in het hoofdstuk nog een aantal categorisatieproblemen aan de orde die voortkomen uit het feit dat de zinnen in het corpus afkomstig zijn uit vier verschillende genres. Zo brengt de annotatie van de nagebootste dialoogpassages uit

het korte verhalen genre bepaalde vragen met zich mee en dit geldt ook voor de wijze waarop de informatie meestal wordt gepresenteerd in informatiefolders, namelijk in de vorm van opsommingen. Het hoofddoel van het hoofdstuk is enerzijds om de categorisatieproblemen in kaart te brengen en anderzijds om toe te lichten hoe hier op structurele wijze mee is omgegaan tijdens het annotatieproces om zo een consequente analyse te kunnen realiseren.

Na de bespreking van de theoretische grondslag van het analysemodel volgt een toelichting op de praktische uitvoering van de analyse en de keuzes die zijn gemaakt in de samenstelling van het corpus. Dit komt uitgebreid aan de orde in Hoofdstuk 4, waarin de onderzoeksmethoden nauwkeurig beschreven worden. Er is gekozen voor corpusonderzoek, omdat deze methode het mogelijk maakt op kwantitatieve wijze patronen in kaart te brengen in een grote dataset – het doel van het huidige onderzoek. Het corpus bestaat uit teksten afkomstig uit vier verschillende genres, te weten academische tijdschriften, krantenberichten, korte verhalen en informatiefolders. De teksten en tekstfragmenten zijn met grote nauwkeurigheid geselecteerd, aan de hand van binnen de methode van de corpuslinguïstiek opgestelde selectiecriteria (zie bijvoorbeeld Biber 1993; Sinclair 2004; McEnery et al. 2006). Belangrijke punten daarbij zijn hoe een gebalanceerd en representatief corpus kan worden samengesteld, waarbij beslissingen moeten worden genomen over bijvoorbeeld het aantal tekstfragmenten dat opgenomen wordt in het corpus, de lengte van de tekstfragmenten, de bron van de fragmenten en de hoeveelheid genres. Naast een beschrijving van de precieze samenstelling van het corpus, geeft het hoofdstuk ook inzicht in de annotatieprocedure, waarbij aandacht wordt besteed aan hoofdbegrippen als nauwkeurigheid en betrouwbaarheid van analyse en beschreven wordt in welke mate deze zijn gerealiseerd.

### **Bevindingen**

Het hoofddoel van dit onderzoek is om zinspatronen in het Engels en het Nederlands in vier verschillende genres in kaart te brengen, waarvan de bevindingen worden gepresenteerd in Hoofdstuk 5. Verder is er ook gekeken naar overeenkomsten en verschillen in de opbouw van het begin van de zin (Hoofdstuk 6), interpunctiegebruik in beide talen (Hoofdstuk 7) en het gebruik en de frequentie van constructies die de zin onderbreken (Hoofdstuk 8).

Een analyse van de discoursestructuur van zinnen heeft laten zien dat alle zinnen in het corpus onderverdeeld kunnen worden in vier hoofdtypen: zinnen die alleen uit een kernboodschap bestaan (C-patroon), zinnen waarbij de kern

voorafgegaan wordt door een of meer satellietboodschappen (XC-patroon), zinnen waarbij de kern gevolgd wordt door een of meer satellietboodschappen (CX-patroon) en zinnen waarbij de kern zowel voorafgegaan als gevolgd wordt door een of meer satellietboodschappen (XCX-patroon). Wanneer er alleen op dit analyseniveau naar zinnen wordt gekeken, vertonen de talen een aanzienlijke mate van overlap. Verschillen tussen de talen worden pas zichtbaar wanneer er niet alleen naar de discoursestructuur van zinnen wordt gekeken, maar ook de grammaticale realisatie van de verschillende discourse-eenheden wordt meegenomen en het gebruik van bepaalde interpunctietekens om de verschillende eenheden van elkaar te scheiden. Wanneer genre ook wordt meegenomen in de analyse, worden er nog meer verschillen tussen de talen zichtbaar. Het meenemen van al deze factoren in een analyse van zinspatronen is precies waar een *sentencing* analyse om draait: de beslissingen blootleggen die de schrijver heeft genomen op het niveau van discourse, van grammatica, interpunctie en genre om zo een effectieve zin samen te stellen.

De statistische analyses laten zien dat bepaalde verschillen tussen de talen niet beïnvloed worden door het genre waarbinnen een zin geschreven wordt. Zo bevat het Engels een significant groter aantal langere zinnen dan het Nederlands, maar ook een groter aantal zinnen dat tot het C-patroon behoort, een groter aantal zinnen dat onderbroken wordt door een satellietboodschap en een groter aantal zinnen waarbij de kernboodschap door twee of meer satellietboodschappen wordt vooraf gegaan. Ook bevat het Engels een groter aantal niet-finiete bijzinnen die op verschillende posities in de zin voorkomen, zowel aan het begin als aan het eind. Het Nederlands bevat daarentegen een groter aantal korte zinnen, kernboodschappen die voorafgegaan worden door een korte satellietboodschap, vaak in de vorm van een bepaling van tijd of plaats; verder laat de taal een groter aantal zinnen zien waarbij de satellietboodschap die volgt op de kern de vorm aanneemt van een hoofdzin of een uitbreidende bijzin.

Aangezien een van de hoofdvragen van het onderzoek is of de verschillen die worden waargenomen te verklaren zijn door verschillen in het taalsysteem of verschillen in taalcultuur of genre, is het relevant om op te merken dat slechts een aantal verschillen te verklaren is door verschillen tussen de taalsystemen. Een verschil dat wel te verklaren is door verschillen tussen de systemen betreft het aantal satellieten dat vooraf kan gaan aan de kernboodschap. Het feit dat het finiete werkwoord in de Nederlandse zin in principe de tweede positie inneemt, verklaart waarom het aantal satellietboodschappen vóór de kern in het Engels een aanzienlijk hogere frequentie laat zien. Een ander verschil tussen de taalsystemen is de significant hogere frequentie van niet-finiete bijzinnen in het Engels, een verschijnsel

dat ook al in andere studies is vastgesteld (zie bijvoorbeeld Aarts & Wekker 1987: 301; De Moor 1998: 309; Cosme 2007: 279-280; Hannay & Mackenzie 2009: 93-96). Omdat een relatief gering aantal van de gevonden verschillen te verklaren is op taalsysteemniveau, is er voor een nadere verklaring gezocht naar verschillen op het niveau van de Engelse en Nederlandse taalcultuur. In het huidige onderzoek is gesteld dat taalcultuur tot op zekere hoogte wordt weerspiegeld in de handboeken en schrijfgidsen die voor beide talen zijn geschreven, waarbij opvalt dat de Nederlandse varianten een iets sterker prescriptief karakter lijken te hebben dan de Engelse schrijfgidsen. Opvallend is dat een aantal van de waargenomen verschillen tussen de talen precies overlapt met een aantal hoofdthema's dat aan de orde komt in de Nederlandse schrijfgidsen. Om de leesbaarheid van teksten te vergroten, wordt bijvoorbeeld het gebruik van lange zinnen, zinnen met een lange aanloop of het gebruik van tangconstructies in een aantal boeken expliciet afgeraden. Nader onderzoek zal echter moeten uitwijzen of de gevonden verschillen tussen de talen inderdaad verband houden met de verschillen in schrijfadvis in de schrijfgidsen. Op dit moment heeft deze verklaring nog de status van een interessante hypothese.

De analyse van zinspatronen kan een stap verder worden gebracht door ook de rol van genre in de analyse te betrekken, waarvan is gebleken dat dit een aanzienlijke invloed heeft op de structuur van een zin, maar in beide talen wel vaak een verschillende. Het academische genre kan in beide talen gekenmerkt worden als het meest complexe. Dit komt tot uiting in de hoge frequentie van lange zinnen, de hoge frequentie van onderbrekingen in de zin, en de hoge frequentie van zinnen die tot het meest complexe zinspatroon behoren, het XCX-patroon. Omdat de frequenties van elk van deze aspecten van zinsstructuur in het Engels nog hoger zijn dan in het Nederlands, wekt het Engelse academische genre de indruk nog complexer te zijn dan het Nederlandse. Naast de verschillen, die ook te vinden zijn op het gebied van de grammaticale realisatie van de verschillende discourse-eenheden in de zin, is de overlap op het niveau van de vier hoofdpatronen in dit genre ook opvallend. Deze overlap zou eventueel kunnen worden gerelateerd aan de invloed van het Engels – als dominante voertaal binnen de academische wereld – op de discoursestructuur van Nederlandse zinnen. Een dergelijke hypothese vereist echter nader onderzoek.

In vergelijking met het academische genre is het krantenberichtgenre in een aantal opzichten te karakteriseren als minder complex. Een eerste, duidelijk verschil is dat de zinnen in dit genre aanzienlijk korter zijn dan de zinnen in het academische genre. Hoewel dit geldt voor beide talen, zijn de Engelse zinnen nog steeds aanzienlijk langer dan de Nederlandse. Dit verschil in zinslengte kan wederom gekoppeld worden aan het expliciete schrijfadvis in de speciaal voor journalisten

ontwikkelde Nederlandse schrijfgidsen, namelijk om zinnen bij voorkeur kort te houden om zo de leesbaarheid te vergroten. De nadruk op de brede en uiteenlopende doelgroep in deze gidsen kan ook een mogelijke verklaring bieden voor de lagere frequentie van zinnen met een lange aanloop en zinnen die een of meer onderbrekingen bevatten. Hoewel ook het Engelse krantenberichtengene een lagere frequentie heeft van deze verschillende aspecten die aan de zinscomplexiteit bijdragen, is de frequentie van deze aspecten nog steeds hoger dan het Nederlandse krantenberichtengene. Daarnaast laat een analyse van bijvoorbeeld het type onderbrekingen in een zin zien dat verschillen tussen de talen niet alleen uitgedrukt kunnen worden in termen van frequentie, maar ook in termen van functie beschreven moeten worden. Waar onderbrekingen in de Nederlandse zin vaak achtergrondinformatie geven, lijkt de onderbreking in het Engels juist gebruikt te worden om bepaalde informatie op de voorgrond te plaatsen, waarbij de *interactie* tussen het gebruik van bepaalde interpunctietekens (bv. gedachtestreepjes), grammaticale realisatie en semantische inhoud bijdraagt aan dit effect. Een nadere analyse van onderbrekingen, zinslengte en grammaticale structuren laat zien dat niet alleen verschillen tussen talen en tussen genres zinsstructuur beïnvloeden, maar dat type tekst (kwaliteitskrant vs. niet-kwaliteitskrant) en katern binnen de krant ook invloed hebben op zinsstructuur. Dit betekent dat teksttype, doel van de tekst en de beoogde lezer van de tekst op verschillende wijze van invloed zijn op de zinsstructuur.

Een van de hoofdkenmerken van het korte verhalengene is dat het uit twee zeer verschillende tekstsoorten bestaat: beschrijvende passages en nagebootste dialogen. Met name de nagebootste dialogen hebben een sterke invloed op de structuur van de zinnen in dit genre, die gekenmerkt worden door hun korte en fragmentarische karakter. In vergelijking met de andere genres bevat dit genre aanmerkelijk kortere zinnen en een aanzienlijk lagere frequentie van zinnen met een lange aanloop of veel onderbrekingen – aspecten die bij een hoge frequentie geassocieerd worden met zinscomplexiteit. In dit opzicht kan het korte verhalengene dan ook gekenmerkt worden als een genre dat wat betreft zinsstructuur minder complex is dan de andere genres. Verschillen tussen de talen worden wederom zichtbaar wanneer gekeken wordt naar de frequentie van zinnen met een lange aanloop en de frequentie van zinnen met onderbrekingen, maar ook als er gekeken wordt naar zinnen met bepaalde grammaticale constructies, zoals niet-finiete bijzinnen, die alle in het Engels vaker voorkomen dan in het Nederlands.

Bij het laatste genre, de informatiefolders, is heel duidelijk zichtbaar dat het doel van de tekst, namelijk informatie op toegankelijke wijze presenteren aan een brede doelgroep, tot uiting komt in de zinsstructuur. Kenmerkend voor dit genre in

beide talen is de hoge frequentie van korte zinnen en de informatie die in de vorm van opsommingen wordt gepresenteerd. Ondanks het overeenkomstige doel van dit teksttype in beide talen, zijn er toch verschillen tussen de talen te vinden, wederom op het niveau van zinslengte, zinnen met een lange aanloop en onderbrekingen, die allemaal frequenter in het Engels voorkomen. Aangezien dezelfde verschillen tussen de talen zich in meer of mindere mate voordoen in de verschillende genres en aangezien de aspecten waarop de talen van elkaar verschillen ook precies de zaken zijn die aan de orde komen in de Nederlandse schrijfgidsen, lijkt in ieder geval een aantal verschillen in verband gebracht te kunnen worden met het prescriptieve advies zoals geformuleerd in Nederlandse schrijfgidsen. Een meer nauwkeurige analyse van de combinatie van discoursestructuur, grammaticale realisatie en interpunctie laat echter ook voor dit genre weer zien dat niet alleen de frequentie, maar ook de functie van bepaalde aspecten van zinsstructuur verschillen tussen de talen blootleggen.

### **Conclusie**

Het hoofddoel van dit onderzoek is om de meest frequente zinspatronen in het Engels en het Nederlands in kaart te brengen en om te bepalen in hoeverre de talen van elkaar verschillen op dit analyseniveau. Een ander doel is vast te stellen of verschillen tussen de talen toe te schrijven zijn aan verschillen tussen de taalsystemen of dat deze zijn toe te kennen aan andere factoren, zoals het genre waarbinnen een zin geschreven is of verschillen tussen de Nederlandse en Engelse schrijfcultuur. Zoals beschreven in de korte bespreking van de belangrijkste resultaten, vertonen de talen op het meest globale analyseniveau – de analyse van de discoursepatronen van zinnen – veel overlap, maar worden de verschillen pas zichtbaar op het moment dat het volledige plaatje – de interactie tussen discourse, grammatica en interpunctie – in de analyse wordt betrokken. Naast een aantal verschillen tussen de talen die te verklaren zijn door ofwel het taalsysteem ofwel de schrijfcultuur van de talen, zijn de voornaamste verschillen tussen de talen te vinden op het niveau van de genres, die in de twee talen regelmatig een verschillende invloed hebben op de zinsstructuur.

Op verschillende gronden zijn er echter ook bezwaren aan te voeren tegen de in dit onderzoek gehanteerde aanpak. Vanuit methodologisch perspectief kan men zich bijvoorbeeld afvragen of het analysemodel dat voor deze studie werd ontwikkeld geschikt is voor de analyse van twee verschillende talen. Aangezien het gebruik van interpunctie bijvoorbeeld een belangrijke rol speelde in het



discoursesegmentatieproces en er tussen het Engels en het Nederlands wel bepaalde verschillen bestaan op dit gebied, leidde dit soms tot bepaalde adhoc annotatiebeslissingen. Hetzelfde geldt voor de keuze van een Engelse grammatica (Quirk et al. 1985) om niet alleen de analyse van de grammaticale structuur van zinnen in zowel het Engels als het Nederlands te verrichten, maar ook de zinnen afkomstig uit vier verschillende genres binnen deze talen. Hoewel allemaal uitvoerig bediscussieerd en weloverwogen, zouden ook de keuzes voor juist deze vier genres en de selectie van grammaticale categorieën waarop het onderzoek zich met name heeft gericht op basis van voortschrijdende inzicht kunnen worden heroverwogen. Ook al hebben deze keuzes de consistentie van de analyses niet in gevaar gebracht, toch betreft het hier aspecten die verfijnd zouden kunnen worden in vervolgonderzoek.

In overeenstemming met het hoofddoel van het onderzoek – het in kaart brengen van zinspatronen – is de nadruk komen te liggen op een kwantitatieve analyse van zinspatronen in beide talen. Door deze nadruk is de kwalitatieve analyse van verschillende aspecten van de kwantitatieve bevindingen wat onderbelicht gebleven. In een vervolgstudie zou het begrip van schrijfcultuur bijvoorbeeld verder uitgediept kunnen worden, door dit niet alleen te baseren op een analyse van schrijfgidsen en handboeken, maar door bijvoorbeeld ook te kijken naar verschillen in het schrijfonderwijs. In de kwalitatieve analyse zou dan ook gezocht kunnen worden naar een verklaring voor bepaalde frequentieverschillen tussen de talen van bepaalde grammaticale categorieën of constructies en zou er extra aandacht besteed kunnen worden aan verklaringen op een beneden of boven de zin gelegen analyseniveau.

De grondige analyse van de interactie tussen discoursestructuur, grammaticale realisatie en interpunctiegebruik heeft inzicht gegeven in de verschillen in zinspatronen tussen het Engels en het Nederlands en tussen de vier verschillende genres binnen deze talen. Het is juist de interactie tussen deze verschillende aspecten van zinsbouw die inzicht geeft in de subtiele verschillen tussen deze nauw verwante talen en daardoor ook inzicht biedt in de verschillende retorische middelen die een schrijver tot zijn beschikking heeft om zijn communicatieve doel te bereiken.



## Curriculum Vitae

Lotte Tavecchio was born in Amsterdam on 23 May 1978. After finishing secondary school (Montessori Lyceum Amsterdam) in 1996, she spent one year in America as an exchange student. She started a degree in English Language and Culture at the Rijksuniversiteit Groningen and continued her studies at the Vrije Universiteit Amsterdam, where she graduated (cum laude) in 2002. After graduation she started her PhD project at the English Department and the Department of Language and Communication of the Vrije Unversiteit Amsterdam. Lotte Tavecchio currently holds a position as lecturer at Amsterdam University College.