Adjectives complemented by *that*- and *to*-clauses

Exploring semantico-syntactic relationships and genre variation

Henrik Kaatari

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in Ihresalen, Thunbergsvägen 3, Uppsala, Friday, 17 March 2017 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Dr Gerold Schneider (University of Zurich).

Abstract

Kaatari, H. 2017. Adjectives complemented by *that*- and *to*-clauses. Exploring semantico-syntactic relationships and genre variation. 76 pp. Uppsala: Department of English. ISBN 978-91-506-2618-6.

The present compilation thesis investigates adjectives complemented by *that*- and *to*-clauses. More specifically, the thesis is concerned with extraposed (e.g. *it is likely that she will win* and *it is important to win*) and post-predicate clauses (e.g. *I'm sure that he's alive* and *I'm glad to see you*). The thesis is most fundamentally concerned with the study of linguistic variation. Thus the aim of the thesis is to explain why a certain construction is used in a given context.

The data used in the studies comes from the British National Corpus (BNC). Study I proposes a semi-automated approach to variable patterns in corpus data. The study describes the creation of a computer program which has been designed to facilitate the extraction and coding of corpus data. In Study II, extraposed and post-predicate *that*- and *to*-clauses are contrasted in terms of their variation across genres, their lexical diversity and the meanings expressed by the adjectives most frequently found in each construction. Study III tests the applicability of the Complexity Principle and the Uniform Information Density Principle on adjectival data, by examining the variation between retaining and omitting the complementizer *that* across extraposed and post-predicate clauses. Study IV tests whether the syntactic status of *I'm sure* is similar to that of *I think*, i.e. whether it exhibits the same signs of grammaticalization.

The results show that extraposed and post-predicate *that*-clauses are associated with similar meanings but differ in most other respects. Compared to post-predicate *that*-clauses, extraposed *that*-clauses are more frequent in formal genres, they are found with fewer instances of *that*-omission, and they are found to be more frequently represented in cognitively complex environments. Similarly, the results also show that extraposed and post-predicate *to*-clauses are associated with similar meanings, but differ in terms of their genre distribution. Instead, in terms of meaning, extraposed *that*- and *to*-clauses on the one hand, and post-predicate *that*- and *to*-clauses on the other, are similar to each other. The thesis highlights the importance of studying adjectival complementation in its own right, and not to treat it as subordinate to, or part of, verbal complementation.

Henrik Kaatari, Department of English, Box 527, Uppsala University, SE-75120 Uppsala, Sweden.

ISBN 978-91-506-2618-6
urn:nbn:se:uu:diva-313799 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-313799)

List of papers

This thesis is based on the following studies, which are referred to in the text by their Roman numerals.

I Garretson, Gregory & Henrik Kaatari. 2014. ‘The computer as research assistant: A new approach to variable patterns in corpus data.’ In Vandelanotte, Lieven, Kristin Davidse, Caroline Gentens & Ditte Kimps (eds.). *Recent advances in corpus linguistics: Developing and exploiting corpora*, 55–79. Amsterdam: Rodopi.

II Kaatari, Henrik. Under review. ‘Adjectival complementation: Genre variation, lexical diversity, and meaning.’ *Journal of English Linguistics*.

III Kaatari, Henrik. 2016. ‘Variation across two dimensions: Testing the Complexity Principle and the Uniform Information Density Principle on adjectival data.’ *English Language and Linguistics* 20(3): 533–558.

IV Kaatari, Henrik. Forthcoming. ‘On the syntactic status of *I’m sure*.’ *Corpora*.

Reprints were made with kind permission from the respective publishers.

Acknowledgments

The present thesis project started out as a monograph, but halfway through, when faced with the erratic behavior of linguistic data, leading the work in all kinds of different directions, I decided that it would make more sense to try to capture these erratic structures in a looser-fitting garment. I therefore decided to opt for a compilation thesis. With the benefit of hindsight, I can now firmly say that this decision has improved the fit of the thesis as well as my mental health status. In fact, the compilation process has been somewhat organic in nature, as some of the studies included are natural extensions of findings from other studies included in the thesis. They are extensions, I might add, which were not envisioned at the planning stage, or even at a more advanced stage of the thesis. While I hesitate to start off in dramatic fashion, there is no way to get around this simple fact: the compilation format saved this thesis' life, and mine.

There are numerous persons without whom this thesis would not have seen the light of day. Most essentially, my main supervisor Professor Merja Kytö has steadily guided me through the process. Merja's calm and patient approach coupled with her razor-sharp attention to details have been immensely helpful. I am also indebted to my secondary supervisor, Docent Christer Geisler. Christer, in fact, also acted as supervisor to my undergraduate essays, which means that we have had a supervisor/supervisee relationship on and off for a decade by now. Needless to say, Christer's help has been tremendously valuable.

I am deeply grateful to the members of the English linguistics seminar; their thoughtful comments have not only improved the thesis, but served to create meaning to the project, at times when meaning seemed hard to find. Thank you Dr. Linnéa Anglemark, Erika Berglind Söderqvist, Dr. Irina Frisk, Dr. Gregory Garretson, Dr. Angela Hoffman, Dr. Christine Johansson, Dr. Ewa Jonsson, Dr. Tove Larsson, Edward Long, Dr. Susanna Lyne, Dr. Pia Norell, Göran Rönnerdal, Sarah Schwarz, Docent Erik Smitterberg, Professor Terry Walker and Dr. Ying Wang.

I am especially thankful to Dr. Gregory Garretson on both a personal and professional level. Gregory has been incredibly generous with his time, allowing for discussions of worldly matters as well as linguistic. Gregory's thought-provoking approach and his encouragement have served as an inspiration. I would also like to express my gratitude to Docent Erik Smitterberg for always carefully commenting on my texts and for his ability to give sound classificatory advice. I am also very grateful to Dr. Tove Larsson for her willingness to discuss linguistic matters as well as her patience in listening to rants about PhD-hardships, to which she seemed strangely immune.

I would like to thank *Erik Tengstrands Stipendiestiftelse* and *Kungl. Humanistiska Vetenskaps-Samfundet i Uppsala* whose grants enabled me to travel to conferences and go on a research stay. Additionally, I would like to thank *Göransson-Sandviken* for providing me with a generous grant when my faculty funding ran out.

Furthermore, I would like express my gratitude to Professor Christian Mair, who welcomed me at the Albert-Ludwigs-Universität Freiburg and facilitated my research stay at the Research training group GRK DFG 1624/1 “Frequency effects in language”. I would also like to thank Professor Mair and Professor Martin Hilpert for their feedback on my work during my time in Freiburg. I am also grateful to Dr. David Lorenz, Dr. Karin Madlener and Dr. Michael Schäfer for introducing me to Freiburg and to their work.

Finally, I am very much indebted to my family. My parents, Hans and Margot, to which I have my life to thank, and my grandmother, Britt, whose concern for my future has persuaded (and annoyed) me enough to finally finish this thesis. Last but not least I want to express my heartfelt gratitude to the two persons closest to me. Evelina, without you this would never have happened, and, trust me, I know how conventional that sounds, but it does not make it any less true, and Atle, although the same may not strictly be true of you, you are nonetheless contributory in every way imaginable.

Stockholm, January 2017

Henrik Kaatari

Contents

1. Introduction

The focus of the present thesis is on adjectives complemented by *that*- and *to*-clauses. The thesis consists of four studies, preceded by this introductory survey. More specifically, we are here mostly concerned with extraposed clauses, as in (1)–(2), and post-predicate clauses, as in (3)–(4), although pre-predicate clauses, as in (5)–(6), are briefly touched upon (see Study I and section 3 of this introductory survey).[1]

(1) It was *inevitable* that he should be nicknamed 'the Ferret', although seldom in his hearing. (CJF)
(2) It is *difficult* to test a potential cure when a disease is ill-defined. (ARF)
(3) I'm *happy* that we are married. (CEY)
(4) Yet the authorities were *unable* to silence the expression of political opposition. (FB1)
(5) That he should be nicknamed 'the Ferret' was *inevitable*.
(6) To test a potential cure when a disease is ill-defined is *difficult*.

The starting point for this project was two general observations about previous studies on complementation. First, previous studies on complementation were found to be strongly biased towards verbal complementation, thereby neglecting to distinguish between verbal and adjectival complementation. Second, previous studies on complementation were found to be biased towards including data of high frequency verbs and adjectives only, thereby excluding a large portion of data consisting of low frequency tokens. In light of these two observations, the guiding principles of the present project was first to treat adjectival complementation as distinct from verbal complementation, and second to find a methodological solution to the problems identified in previous studies by including both high frequency and low frequency tokens.

Although adjectives allow complementation by four different types of clausal complements (i.e. *that*-, *to*-, *ing*- and *wh*-clauses), a conscious decision was made to only focus on *that*- and *to*-clauses. Behind this decision lie a

[1] Throughout this introductory survey, examples followed by a three-character code (indicating the file name) are taken from the BNC. Examples without a code are constructed and/or have been manipulated, or taken from secondary literature in which case they are clearly referenced.

number of reasons. First of all, *that*- and *to*-clauses share a number of structural properties which make these two clause types subject to variation in several respects (see e.g. Mair 1990: 47 ff.; see also sections 2.4 and 3.5 of this introductory survey for examples). Second, by focusing on *that*- and *to*-clauses, we are concerned with one finite and one non-finite clausal complement. The reason why this is interesting is that finite and non-finite complements have been shown to be associated with different genre distributions (Biber et al. 1999: 754 ff.) and with different meanings (Mair 1990: 24 ff.; Herriman 2000). In other words, we are thus dealing with two clausal complements that are subject to variation, but that at the same time have been found to be associated with different uses and different meanings. Finally, *that*- and *to*-clauses are far more frequent than *ing*- and *wh*-clauses, making them suitable candidates for the quantitative analyses conducted in this thesis (see e.g. Biber et al. 1999: 754 ff. on the frequency of different clausal complements).

Each of the studies included in this thesis focuses on different aspects of adjectival complementation. However, common to all four studies is that a great deal of attention is placed on methodological aspects, and that all four studies deal with linguistic variation. First, the core methodology of this thesis is most prominently featured in Study I, which proposes a semi-automated approach for data collection and post-processing. This approach is illustrated in Study I by the collection of data which later features in Study II and Study III. Throughout the thesis, major emphasis is placed on rigorous data collection, centering on the creation of a sub-corpus of the British National Corpus (see section 4 of this introductory survey) from which the data appearing in Studies I, II and III originates. On the whole, this thesis sets out to be at the forefront of empirically driven linguistic research. This is manifested both in the types of methodologies used and in the statistical significance testing conducted throughout the thesis. Second, at the heart of this thesis is the study of linguistic variation. This is manifested throughout the thesis in a number of ways. Most fundamentally, the material used and the methodology for data collection were designed with the aim of facilitating the study of linguistic variation (see Study I). In Studies II, III and IV, variation is omnipresent in terms of variation across two (or more) linguistic features, variation of a certain linguistic feature and/or linguistic construction across genres, and across different linguistic constructions. In addition to this, the thesis also focuses on interaction of factors influencing linguistic variation. The investigations in this thesis are based on several different factors, including linguistic factors such as subject type and complexity, and extralinguistic factors such as genre. These factors are not assumed to function independently, but rather to function hierarchically and/or to interact with each other. My concern here is to pin down which the most important factors are, and to which extent these factors interact. The weighting of different factors and mapping of their interaction (not always necessarily in a strict statistical sense), is thus a common theme running throughout the thesis.

Before the aims of the thesis are presented in section 1.1, I should mention that although adjectival complementation by both *that*- and *to*-clauses features in this thesis, these two clause types are not equally represented in the studies. Due to the organic nature of this compilation thesis, with the results of one study often prompting the research questions asked in a subsequent study, *that*-clauses play a larger role than do *to*-clauses. However, in this introductory survey, both types are treated with more or less equal weight in order to give a thorough background to the discussions of the clause types in the studies (see especially section 3 of this introductory survey).

Readers may also be aware that pre-predicate clauses, as in (5) and (6), are only represented in Study I. Although they constitute a semantically equivalent rival to extraposed clauses, as shown by comparing (1) and (2) to (5) and (6), they are too infrequent to be studied quantitatively (see Study I, table 4). Pre-predicate clauses are, however, discussed and compared to extraposed clauses in this introductory survey.

1.1 Aims

Given the focus on linguistic variation in this thesis, English adjectival complementation constructions are investigated with the broader aim of explaining why a certain construction is used in a given context. Furthermore, throughout the thesis, adjectival complementation constructions are contrasted with previous research on verbal complementation constructions with the aim of answering the question of whether adjectives behave differently than verbs complemented by *that*- and *to*-clauses.

In addition to these broader aims, each study has aims of its own. The aim of Study I is to devise a principled semi-automated approach to data collection. More specifically, the study aims to test whether diverse patterns of data can be extracted by the use of part-of-speech tagging.

In Study II, the constructions under study include both extraposed and post-predicate *that*- and *to*-clauses. The aim of this study is to determine which sets of constructions exhibit the most similarities in terms of their variation across genres, their lexical diversity, and the meanings expressed by the adjectives most frequently found in each construction.

The aim of Study III is to test the applicability of the Complexity Principle (Rohdenburg 1996) and the Uniform Information Density Principle (Jaeger 2010), on adjectival data (as opposed to verbal data from which both these two principles derive), as regards the variation between retaining and omitting the complementizer *that*. Furthermore, the study aims to test other related factors found in previous research to have an effect on this particular variation.

Study IV is a direct extension of Study III, in that it was found in Study III that the subject-adjective combination *I* (*'m|am*) *sure* is disproportionately associated with *that*-omission. The aim of Study IV is to test whether the syntactic status of *I* (*'m|am*) *sure* is similar to that of *I think*, i.e. whether it exhibits the same signs of grammaticalization along two different parameters, namely syntactic mobility and the extent of the variation between omitting and retaining *that*.

1.2 Outline

The rest of this introductory survey is structured as follows: section 2 provides background on the theoretical frameworks of importance for the thesis and a short overview of previous research not covered in the four studies. Section 3 introduces and defines the constructions included in the studies and draws distinctions between complement clauses and related constructions. Following that, in section 4, the material used in the studies is introduced. The four studies included in the thesis are summarized in section 5. Section 6 provides a discussion of the main findings of the studies and their relevance.

2. Background and previous research

Complementation, due to its status as an integral component of transformational grammar, has been commented on extensively by researchers from different linguistic backgrounds. However, descriptions covering the whole spectrum of clause types and constructions have more or less been limited to verbal predicates only. When adjectival predicates have been discussed they have traditionally been treated together with their verbal counterparts or simply marginalised. Thus, although primarily focusing on verbal predicates, typological (e.g. Givón 1980; Noonan 2007), generative (Rosenbaum 1967) and general studies of complementation (e.g. Huddleston 1971; Greenbaum et al. 1996) typically include examples of adjectival predicates without distinguishing them from verbal and nominal predicates.

This section is mostly concerned with previous research directly related to the studies included in the thesis. Further, in an attempt not to reiterate what is said in the individual studies, I am here mostly addressing additional background information on complementation (readers are referred to the individual studies for more information). Note also that in section 3 of this introductory survey, I engage with previous research related to the classification schemes used in this thesis. However, before we start examining previous research on complementation in sections 2.1–2.4, a few brief comments on two important methodological frameworks are in order.

First of all, this thesis is a corpus linguistic study (see Leech 1991). In such studies the data used is drawn from large computerized collections of authentic texts (including transcriptions of speech), i.e. corpora, aimed at being representative of a given variety or genre (see section 4 for more on representativeness). Furthermore, this also means that the results presented in this thesis are meant to be replicable. This replicability, however, relies on careful and explicit categorization of the variables included in the analysis (see section 3 for some of the classification schemes used in the present thesis). Finally, this thesis is corpus-based (as opposed to corpus-driven), which means that it makes use of corpus data to test pre-existing linguistic hypotheses.

Second, a prominent feature of this thesis is the use of a quantitative corpus linguistic method called Collostruction Analysis (Stefanowitsch & Gries 2003, 2005 and Gries & Stefanowitsch 2004a), which is geared to investigating the interaction of words and constructions (see Studies II and III). Collostructional Analysis is theoretically based on the constructionist view of language as a repository of linguistic signs. Within this view, constructions, i.e.

linguistic signs, are seen as pairings of form and meaning (cf. Construction Grammar, Goldberg 1995). The term construction is thus associated with a large number of different linguistic features, ranging from specific features such as individual morphemes to abstract syntactic structures such as argument structures (see Goldberg 1995). Furthermore, in this view it is argued that constructions themselves, independent of the words that occur in them, carry meaning (cf. Goldberg 1995: 1). Fundamental to the constructionist view of language is the idea that the members of an apparent alternation are not seen as hierarchically related, i.e. with one derived from the other; instead, both members of an apparent alternation are seen as constructions in their own right (see Goldberg 2002; see also Van Bogaert 2010: 417 ff., 2011: 318 ff. and Study IV on constructional grammaticalization).

2.1 Data collection

As mentioned in section 1, a general observation about previous studies of complementation was that they typically were found to be biased towards including data on high frequency verbs and adjectives only, thereby excluding a large portion of data consisting of low frequency tokens. The case study in Study I of this thesis is an attempt to solve the problem of extracting both high frequency and low frequency tokens.

The rationale behind the approach to data collection in previous research has been that the most frequent adjectives in these complementation patterns constitute a fairly small group that represents the majority of the total number of tokens in question. The approach is thus based on the finding that the distribution of lexemes is zipfian in nature, with a few very frequent lexemes followed by a long tail consisting of many different infrequent lexemes (see e.g. Study II, figures 3–4).[2] In those studies (see e.g. Mindt 2011), a list of adjectives is typically compiled through the use of very general search terms, and an estimate of how much of the data these adjectives account for is calculated, based on these general search terms.[3] In approaches such as these, there is an obvious discrepancy between the rigor with which the tokens, derived from the lexical list, were extracted, and the generic nature of the estimate of the total tokens in the sample. The present study avoids this methodological problem by aiming to extract *all* valid tokens of adjectives complemented by *that*- and *to*-clauses, regardless of which lexemes they represent. I argue that

[2] Zipf's law, named after the American linguist George Kingsley Zipf, stipulates that the frequency of a word is relative to its rank in a frequency table, meaning that the frequency of the most frequent words will be distributed proportionally according to their frequency rank (see Zipf 1932; Li 1992).

[3] In Mindt (2011), for example, the search term ADJECTIVE+THAT is used, which discounts instances in which there is separation between the adjective and the complementizer *that*, and instances in which *that* is omitted (see Study I for a discussion of solutions to these problems).

there is a difference between retrieving 10,000 tokens at random from the entire BNC and retrieving all 10,000 tokens from a well-defined subcorpus; the latter approach allows for total accountability of the phenomenon in question, which in turn enables one to make stronger generalizations. I also argue that a non-lexical data extraction approach to complementation is preferable, since individual lexemes, however infrequent, play a part in establishing the lexical profile of a construction. For example, many of the infrequent lexemes can be synonyms and may thus form larger clusters that are missed in an analysis based on a restricted number of lexemes. Finally, by extracting *all* adjectives complemented by *that*- and *to*-clauses, it is possible to determine the extent to which different constructions differ in terms of how many lexemes they are represented by (see Study II).

2.2 Genre

Several studies have shown that adjectives complemented by *that*- and *to*-clauses are distributed differently across different genres. Biber et al. (1999: 754 ff.) examine the distribution of extraposed and post-predicate clauses across four different genres and also demonstrate that there are systematic differences in the distribution of adjectives across these clause types; extraposed clauses appear to be less frequent in less formal genres such as fiction and conversation, whereas post-predicate clauses appear to be less frequent in formal genres and more frequent in less formal genres (see Study III, figure 6 for more on the formality of different genres).[4] However, by adopting a more detailed genre division, other studies show that the variation is more complex than previously assumed. Both Peacock (2011) and Groom (2005) demonstrate that there are differences in the distribution of adjectives complemented by extraposed *that*- and *to*-clauses across different (sub-)genres and academic disciplines, indicating that the genres adopted by Biber et al. (1999: 15) are too broad to capture the variation of these clause types accurately. There are also previous studies of adjective complementation in which genre variation is neglected or given a marginal role. In her study of adjectives complemented by *that*-clauses, Mindt (2011) only treats genre variation marginally in connection with *that*-omission. Similarly, Van linden (2012) neglects the role of genre variation in explaining the distribution of adjectives and different clausal complements and constructions.

Since genre variation has proved to play an important role for the distribution of adjectives, the present thesis adopts a genre variation approach that

[4] Biber et al. (1999: 15) actually uses the term ‘register’ to refer to what in the present study is called ‘genre’. See section 4.2 of this introductory survey for how the term ‘genre’ is defined in the present study.

includes a wide variety of different genres, comparable at different levels of generality (see section 4; see also Studies II and III).

2.3 On the variation between retaining and omitting *that*

There are two apparent biases in previous research on the variation between retaining and omitting the complementizer *that* (e.g. *I'm happy* (*that*) *we are married*). First, most previous studies focus solely, or primarily, on verbal predicates (Thompson & Mulac 1991b; Rohdenburg 1996; Jaeger 2010). Second, most studies are concerned with post-predicate clauses only, thereby neglecting to examine the difference in behavior between post-predicate (e.g. *I'm happy* (*that*) *we are married*) and extraposed clauses (e.g. *It was inevitable* (*that*) *he should be nicknamed 'the Ferret'*; a notable exception here is Mindt 2011; see also Kaltenböck 2006 for a study of extraposed clauses only).

Furthermore, there are two strands of research related to the variation between retaining and omitting *that*. First, there is a large strand of research focusing on the different factors which have a bearing on this variation. These factors include cognitive complexity (Rohdenburg 1995, 1996), production preferences (Jaeger 2010), genre variation (Elsness 1984; Biber 1999; Kaltenböck 2006; Mindt 2011), matrix subject type (Thompson & Mulac 1991a, 1991b; Torres Cacoullos & Walker 2009), and coreferentiality between the matrix and complement clause subjects (Elsness 1984). Second, there is a strand of research primarily focusing on the subject-verb combination *I think* (Thompson & Mulac 1991a; Aijmer 1997; Torres Cacoullos & Walker 2009; Van Bogaert 2010, 2011; Kaltenböck 2011, 2013). These studies aim to discover whether certain uses of *I think* can be seen as instances of grammaticalization, as they are treated as reanalyzed into epistemic phrases. The fact that *I think* is found to be heavily associated with *that*-omission is a crucial part of this strand of grammaticalization research.

Studies III and IV deal with the variation between retaining and omitting *that*. Both these studies test whether the factors identified in previous research on verbal complementation are also applicable to adjectival complementation. Study III is connected to the first strand of research in that it tests different factors which have a bearing on this variation, while Study IV is connected to the second strand of research in that it tests whether *I* (*'m|am*) *sure* is similar to *I think* syntactically.

2.4 Meaning

It was mentioned in section 1 that one of the reasons behind the fact that this thesis focuses on adjectives complemented by *that*- and *to*-clauses is that these two clausal complements have been found to be associated with different meanings. In Study II, this topic is explored primarily based on the concept of modal logic (see e.g. Palmer 1979 and Perkins 1983; readers are here also referred to the discussion in Study II, section 3.3). While the classification of meaning in Study II is based on one specific classification scheme (see Herriman 2000), there are other semantic classifications that merit a short discussion.

First of all, all these semantic classifications of adjectives complemented by *that*- and/or *to*-clauses make direct or indirect reference to the concept of modality (see Kiparsky & Kiparsky 1970; Quirk et al. 1985: 1222 ff.; Mair 1990: 26; Biber et al. 1999: 672 ff.; Dixon 2005: 86–87; Mindt 2011; Van linden 2012). Modality has been defined in a number of ways (see Van linden 2012: 11 ff. for an overview), but is seen in this study as "referring to the totality of linguistic means used to express speakers' attitudes towards the truth of propositions or the likelihood or desirability of events" (Mair 1990: 84). Regarding the difference in meaning between *that*- and *to*-clauses, Mair (1990: 24) finds that extraposed *to*-clauses typically express potentiality (*e.g. it is possible to*) and ease or difficulty (e.g. *it is easy/hard to*), whereas extraposed *that*-clauses typically express truth (e.g. *it is true that*) and probability (e.g. *it is likely that*) (see also Biber et al. 1999: 672 ff. for similar categories). Similarly, Dixon (2005: 86–87) comments on the semantic class of adjectival predicates and the type of complement clause they typically occur with, concluding that 'difficulty' adjectives are predominantly complemented by *to*-clauses, whereas adjectives expressing 'qualification' or 'value' are normally complemented by *that*-clauses.

It is also clear that the type of meaning expressed by an adjectival predicate has some bearing on the type of constructions it allows complementation by. Thus we find that adjectives expressing 'truth' or 'probability' (what is seen as epistemic modality in Study II) allow complementation by extraposed *that*-clauses and post-predicate *to*-clauses (e.g. *it is likely that Sue will win* and *Sue is likely to win*). Adjectives expressing 'ease' or 'difficulty' (what is seen as dynamic modality in Study II) in contrast, allow complementation by extraposed and post-predicate *to*-clauses, but not by *that*-clauses (e.g. *it is hard to win the championship* and *the championship is hard to win*). These two types of variation contexts are referred to as 'subject-to-subject raising' and 'object-to-subject raising' (Biber et al. 1999: 716–717).[5] These two are further discussed in section 3.5 in this introductory survey, but we will note here that the variation between the different complements they exhibit can, in part, be explained by the meaning expressed by the adjectival predicate.

After this brief background, we will now further consider previous research which has been consulted in the classification schemes used in this thesis. In

[5] The term 'tough-movement' (see Mair 1990: 57 ff.) is also commonly used to refer to what is here termed object-to-subject raising.

the next section, we are concerned with defining and delimiting the class of complement clauses and the class of adjectives.

3. Complement clauses and adjectives: relevant distinctions

This section is devoted to defining and delimiting complement clauses and adjectives. In doing so, the section introduces definitions and distinctions central to the study.

The section starts by defining complement clauses more broadly (section 3.1), and proceeds by further narrowing the focus in defining extraposed, pre-predicate and post-predicate clauses in section 3.2. Moving on, complement clauses are distinguished from purpose clauses (3.3), from comparative clauses (3.4) and from a set of other similar constructions (3.5). Finally, section 3.6 defines adjectives and describes the tests adopted for distinguishing adjectives from participles.

3.1 Defining complement clauses

In this study, complement clauses are seen as dependent (subordinate) clauses functioning as an argument of a predicate (cf. Diessel & Tomasello 2001: 100; see also Noonan 2007: 52). In the function of argument, complement clauses may serve as subject (pre-predicate position) or object (post-predicate position) of the matrix clause, as illustrated in (7)–(10) (Biber et al. 1999: 659; Diessel & Tomasello 2001: 100; Huddleston & Pullum 2002: 215). A third type of complement clause can also be distinguished, in which the clausal subject is postponed to end position and replaced by non-referring *it*, as in (11) and (12).

(7) That Bill wasn't in class was *surprising*. [Pre-predicate]
(8) The teacher was *surprised* that Bill wasn't in class. [Post-predicate]
(9) To not find Bill in class was *surprising*. [Pre-predicate]
(10) The teacher was *surprised* not to find Bill in class. [Post-predicate]
(11) It was *surprising* that Bill wasn't in class. [Extraposed]
(12) It was *surprising* not to find Bill in class. [Extraposed]

This view is different from the one adopted in, for example, Quirk et al. (1985: 65), who "reserve the term COMPLEMENTATION (as distinct from *complement*) for the function of a part of a phrase or clause which follows a word, and completes the specification of a meaning relationship which that word implies." Given that this definition explicitly states that complementation involves structures that '*follow* a word,' and given that they do not include subject clauses in their chapter on complementation (see Quirk et al. 1985: Ch. 16), Quirk et al. (1985) do not seem to consider pre-predicate clauses, as in (7) and (9), to be complements. Huddleston & Pullum (2002), in contrast, explicitly state that they consider subject clauses (i.e. pre-predicate clauses) to be complements, at the same time noting that "[m]any other grammars restrict the term 'complement' to non-subject elements" (2002: 215). Similarly, Biber et al. (1999: 659) also explicitly state that "complement clauses of all structural types can occur in both **pre-predicate** (subject) position and **post-predicate** (e.g. direct object) position" (emphasis by the authors).

Another important view maintained in this study is that what is here termed a complement clause construction consists of two different parts: (i) a complement-taking adjective clause (CTA-clause), consisting of, minimally, a complement-taking adjective (CTA), and (ii) a complement clause. In this thesis, I am thus extending the terminology from the literature on verbal complementation where the terms 'complement-taking verb clauses' and 'complement-taking verbs' are used (see e.g. Diessel & Tomasello 2001; Thompson 2002; Boye & Harder 2007) to also cover adjectival complementation. The relationship between complement-taking adjective clauses and complement clauses is illustrated in figure 1.

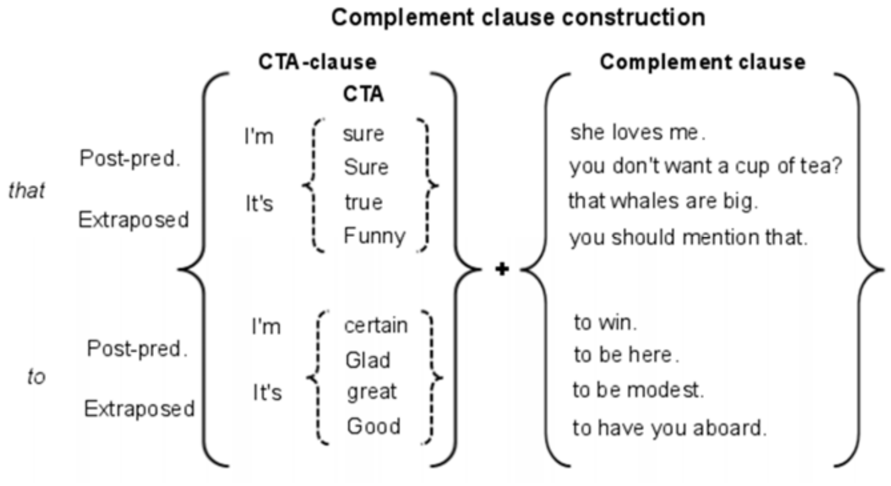

Figure 1. Schema of the relationship between complement-taking adjective clauses (CTA-clauses), complement-taking adjectives (CTA) and complement clauses.

An essential part of this view is that CTA-clauses can be realized not only by 'full clauses', i.e. comprising both a subject and a verb phrase, but also by a complement-taking adjective alone (as in *Funny you should mention that* and

Good to be here in figure 1). The term ‘clause’ is here used in a sense similar to Aarts (1992), allowing for clauses to lack overt subjects and verbs. In other words, a CTA-clause consisting of only an adjective is still considered a clause since it allows a clausal complement. We could of course also reconstruct ‘full clauses’ such as (*Are you*) *sure you don’t want a cup of tea*, from verbless and subjectless CTA-clauses. In this sense, the subject and the verb of the CTA-clause is not ‘overtly expressed’ (cf. Diessel & Tomasello 2001: 106).

Adopting this view also means that it is possible to account for pre-predicate clauses in the same way as post-predicate and extraposed clauses are accounted for. As seen in figure 2, the composition of pre-predicate clauses mirrors the composition of extraposed and post-predicate clauses, the only difference being that the complement clause precedes the CTA-clause.

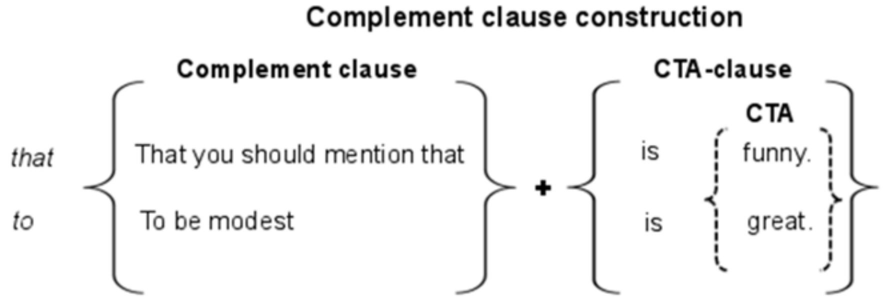

Figure 2. Schema of the relationship between complement-taking adjective clauses (CTA-clauses), complement-taking adjectives (CTA) and complement clauses: pre-predicate clauses.

The schematic relationship illustrated in figures 1 and 2 also has a bearing on the semantic relationship between CTA-clauses and complement clauses. Complement clauses are traditionally defined as “a type of dependent clause used to complete the meaning relationship of an associated verb or adjective in a higher clause” (Biber et al. 1999: 658; cf. also the definition in Quirk et al. 1985: 65 quoted above). In this type of definition, the combination of a CTA-clause and a complement clause is seen as a composite structure with the main proposition expressed by the CTA-clause. The complement clause is thus conceptually embedded in the CTA-clause (cf. Diessel & Tomasello 2001: 102). However, several studies, dealing primarily with the grammaticalization of complement-taking verbs, make a distinction between instances where a CTA-clause and a complement clause are seen as a composite structure and instances where the CTA-clause does not express a full proposition (see e.g. Thompson & Mulac 1991a; Diessel & Tomasello 2001; Boye & Harder 2007; Van Bogaert 2010, 2011; Kaltenböck 2011, 2013; see also Study IV which explicitly deals with this topic). The difference between these two types is illustrated in (13) and (14).

(13) Britain is due to have four submarines armed with the new Tridents but has been *worried* that recent votes in the Senate to cut funding for production of the missile could delay delivery of the weapons, scheduled to start next year, to the Royal Navy. (A28)

(14) 'I'm *sure* your friend will be with you shortly.' (H8H)

The difference between (13) and (14) resides in the propositional force of the CTA-clause; in (13) the central state of affairs is expressed in the proposition of the CTA-clause, whereas in (14), the CTA-clause is largely independent of the complement clause and functions more as an epistemic marker or attention-getter (cf. Thompson & Mulac 1991a ; Diessel & Tomasello 2001). The distinction between these two types has been referred to as 'assertive' vs. 'formulaic' (Diessel & Tomasello 2001), as 'propositional' vs. 'interpersonal' (Van Bogaert 2010) and as 'lexical' vs. 'grammatical' (Boye & Harder 2007).[6] Diessel & Tomasello (2001: 106–107) list seven features defining the use of formulaic CTA-clauses, based on the work of Hooper (1975), Hooper & Thompson (1973) and Thompson & Mulac (1991a):[7]

i. The CTA-clauses are always short and formulaic (suggesting that they are stored as holistic expressions).

ii. The subject of the CTA-clause is either not overtly expressed or it is a first- or second-person pronoun.

iii. The verb in the CTA-clause occurs in the present indicative active.

iv. There are no auxiliaries, modals, adverbs, or prepositional phrases in the CTA-clause.

v. The complement clause tends be much longer and more diverse [than the CTA-clause].

vi. Since the complement clause is non-embedded (both formally and conceptually) it does not include a *that*-complementizer.

vii. The order of the CTA-clause and the complement clause is variable: the CTA-clause may precede or follow the complement clause or may even be inserted into it.

[6] Diessel & Tomasello (2001) also include a 'performative' use situated between the assertive and formulaic use. Similarly, Boye & Harder (2007) distinguish between 'usage-level' phenomena and 'structure-level' phenomena of which the distinction between lexical and grammatical CTCs is characterized as a structure-level phenomenon.

[7] The features included in (i)–(vii) are quoted from Diessel & Tomasello (2001: 106–107). However, certain abbreviations have been changed since their study deals with verbal predicates and they make use of abbreviations such as CTV-clause to refer to a complement-taking verb clause. CTV-clause has been changed to CTA-clause in (i)–(vii). Note also that feature (v) has been interpreted as comparing the length of the CTA-clause to the length of the complement clause (as indicated by the addition within brackets).

It should be stressed that Diessel & Tomasello (2001) base their analysis on verbal complement-taking clauses, as do e.g. Thompson & Mulac (1991a), but it is nonetheless clear that the example in (14), with an adjectival predicate, exhibits all seven features: (i) the CTA-clause is short, (ii) the subject of the CTA-clause is a first-person pronoun, (iii) the verb in the CTA-clause is in the present indicative active, (iv) there are no auxiliaries, modals, adverbs or prepositional phrases in the CTA-clause, (v) the complement clause is longer than the CTA-clause, and it also contains a modal, (vi) it features *that*-omission, and (vii) the CTA-clause can be moved to other positions, as illustrated in (15) (see also Study IV which deals with the ability of *I* (*'m*|*am*) *sure* to occur in different syntactic positions).[8]

(15) a. I'm *sure* your friend will be with you shortly. (H8H)
b. Your friend will be with you shortly, I'm sure.
c. Your friend, I'm sure, will be with you shortly.

It has been pointed out that the formulaic use, although the construction includes both a CTA-clause and a complement clause, is not really biclausal; instead the CTA-clause has been demoted to a clausal operator, functioning as an epistemic marker or attention getter (cf. Thompson & Mulac 1991a; Diessel & Tomasello 2001). However, despite the apparent differences between the assertive and formulaic uses, the view adopted in this thesis is more structural in nature; both assertive and formulaic uses are included on the basis that they both include the minimal component required, i.e. a complement-taking adjective (CTA), which, although weakened, has some semantic relationship to the complement clause (cf. the view maintained in Van Bogaert 2010). That said, tokens in which there is no clear semantic relationship between the CTA-clause and the complement clause, as in (16), or when the relationship is ambiguous, as in (17), have been excluded.

(16) a. *Good* I'm very pleased. (KC3)
b. [*Good*] [I'm very pleased].
(17) a. Oh *sorry* I didn't notice that. (KE3)
b. [Oh] [*sorry* I didn't notice that].
c. [Oh *sorry*] [I didn't notice that].

In (16) there is no clear relationship between the CTA-clause (*good*) and the complement clause (*I'm very pleased*). This is seen as an example of a fully biclausal construction. In (17), in contrast, two different readings are possible: the reading in (b) meets the requirements posited in this study for complement clauses, whereas the reading in (c) does not. Ambiguous examples such as

[8] Van Bogaert (2010; 2011) has shown that feature (iv) is too simplistic as she has found formulaic uses with a variety of different auxiliaries and modals.

(17) have been excluded since there is no reliable way of determining whether they are complement clauses or not due to the lack of prosodic mark-up in the BNC (see Burnard 2007).

The distinction between the assertive and the formulaic uses is primarily geared towards finite complement clauses. However, as seen in figure 1, *to*-clauses are also represented by CTA-clauses consisting only of a CTA, as in (18).

(18) Always *glad* to oblige. (GUF)

There are thus examples of *to*-clauses that resemble the formulaic use of *that*-clauses, since many of the features listed in (i)–(vii) are shared. It should be noted, however, that with *to*-clauses, the CTA-clause can never be as loosely connected to the complement clause as with *that*-clauses, since *to*-clauses do not constitute 'full clauses' in the same sense as *that*-clauses, which include both a subject and finite verb, as demonstrated by the fact that a *that*-clause can function alone without the CTA-clause, as in (19), a reduced variant of (15), whereas a *to*-clause cannot, as seen in (20), a reduced variant of (18).

(19) Your friend will be with you shortly
(20) *To oblige

In summary, the view maintained in this study is that complement clauses are a type of dependent clause with the ability to occur in both subject and object position. This position is in line with the treatment in Huddleston & Pullum (2002) and Biber et al. (1999). Furthermore, a complement clause construction is defined as consisting of two different parts: (i) a complement-taking adjective clause (CTA-clause), consisting of, minimally, a complement-taking adjective (CTA), and (ii) a complement clause. This is important, since it allows the inclusion of CTA-clauses consisting only of a CTA (e.g. *Good to be here*). Furthermore, it also allows for the inclusion of formulaic CTA-clauses in which the CTA-clause is only loosely connected to the complement clause (e.g. *I'm sure he will win*).

In sections 3.3 and 3.4, complement clauses are further distinguished from two similar, but distinct, constructions. Before that, however, the three types of complement clauses included in this study are introduced.

3.2 Defining extraposed, pre-predicate and post-predicate complement clauses

As mentioned in section 3.1, there are three main types of complement clauses included in this study: extraposed, pre-predicate and post-predicate clauses. In what follows, these three types are defined and exemplified. Furthermore, two taxonomies of *that*- and *to*-clauses respectively are presented which show the taxonomic relationship and the structural properties of the three types of complement clauses. The section ends with a discussion of a special construction which does not lend itself to a straightforward classification into the three types distinguished.

Starting with extraposed clauses, extraposition is traditionally defined as a type of "[p]ostponement which involves the replacement of the postponed element by a substitute form" (Quirk et al. 1985: 1391). Following this, extraposed subject clauses, as in (21) and (23), are seen as derived from their non-extraposed counterparts, such as the pre-predicate clauses in (22) and (24).[9]

(21) It is *interesting* that he was late. [Extraposed subject clause]
(22) That he was late is *interesting*. [Pre-predicate clause]
(23) It is *interesting* to watch the game. [Extraposed subject clause]
(24) To watch the game is *interesting*. [Pre-predicate clause]

Extraposition thus involves a process in which the clausal subject of the non-extraposed variants in (22) and (24) (*that he was late*) is replaced by the anticipatory pronoun *it* (Quirk et al. 1985: 1391) in the extraposed variants in (21) and (23), thereby moving the clausal subject of the non-extraposed variants in (22) and (24) to the end of the sentence of the extraposed variants in (21) and (23). This account has been criticized on the grounds that it seems counter-intuitive to explain extraposition, a very frequent phenomenon, in terms of a less frequent variant (cf. Mair 1990: 28 ff. on structural and statistical prototypes; see also Mindt 2011: 30 ff. for a similar argument). Another problem with this approach is that it fails to account for object extraposition of *that*- and *to*-clauses. As seen in (26) and (28), there are no clear non-extraposed variants available for the extraposed object clause in (25) and (27).

(25) I made it *clear* that we will not tolerate stealing. [Extraposed object clause]
(26) ?I made that we will not tolerate stealing *clear*. [Non-extraposed]

[9] The terms extraposed and post-predicate clauses are here used to refer to the combination of a matrix clause and a complement clause. This means that the term extraposed clause is not reserved solely for the extraposed complement clause, but rather refers to a construction in which a matrix clause is complemented by an extraposed clause.

(27) Teachers find it *difficult* to prevent bullying. [Extraposed object clause]
(28) ?Teachers find to prevent bullying *difficult*. [Non-extraposed]

Quirk et al. (1985: 1393) note that "[w]hen the object is an *ing*-clause […] it can undergo extraposition; when it is a *to*-infinitive clause or a *that*-clause, it must do so" (see also Huddleston & Pullum 2002: 1255 and Mair 1990: 91 for similar observations). In light of these examples, it seems problematic to define extraposition in terms of a process of movement or replacement from a non-extraposed variant. Following this, in this study, extraposition is defined as CTA-clauses which include non-referring *it* in subject or object position. In other words, extraposition is not necessarily seen as the result of a movement or replacement process, but rather treated as a construction with one particular feature, i.e. including non-referring *it*.

Pre-predicate clauses, on the other hand, are distinguished from extraposed clauses by always including a *that*- or *to*-clause in subject position of the matrix clause (cf. (22) and (24); see Biber et al. 1999: 659). It should also be noted that pre-predicate *to*-clauses allow a *for*-subject, as illustrated in (29).

(29) For Bill not to be in class was *surprising*.

Post-predicate clauses (a term adopted from Biber et al. 1999) are defined in this study as clauses in which the subject, or, very rarely, the object, of the CTA-class is realized by a referring pronoun or a noun phrase. Note, however, that post-predicate clauses never include non-referring *it* in subject or object position of the CTA-clause (cf. the extraposed object clauses in (25) and (27) which have a referring pronoun (25) and a noun phrase (27) as subjects of the CTA-clause, but are nonetheless treated as extraposed clauses since they have non-referring *it* in object position). Post-predicate subject clauses match Quirk et al.'s (1985: 1223) 'experiencer as subject' class of complement clauses, in which experiencer subjects are defined as "[a]nimate subjects of copular verbs followed by an emotive complement" (Quirk et al 1985: 745). Post-predicate clauses are almost always found with a referring pronoun, as in (30), or a noun phrase, as in (31), as subject of the CTA-clause. Very rarely, referring pronouns, as in (32), or noun phrases, as in (33), are found as object of the CTA-clause (cf. Mindt 2011: 111).

(30) They are *aware* that the sun is lethal. [Post-pred. subject]
(31) Suntan lotion is *likely* to save your life. [Post-pred. subject]
(32) She made them *aware* that the sun is lethal. [Post-pred. object]
(33) They found suntan lotion *easy* to use. [Post-pred. object]

Based on the definitions outlined above, the taxonomic relationship of pre-predicate, extraposed and post-predicate clauses is illustrated in figure 3 (*that*-

clauses) and figure 4 (*to*-clauses) together with corpus examples. Note that the taxonomic relationship of pre-predicate, extraposed and post-predicate clauses is identical for both clause types; both *that*- and *to*-clauses allow subject and object extraposed clauses as well as subject and object post-predicate clauses. Variants which are not attested in the data, such as object post-predicate clauses with *that*-omission (42), are given within brackets in figures 3 and 4.

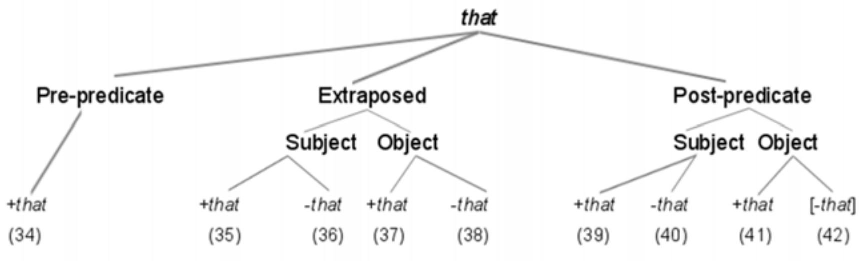

Figure 3. Taxonomy of *that*-clauses.

(34) That James would attempt to recover his lost throne seemed *inevitable*, especially since he had taken refuge in France. (BNB)

(35) And it was *odd* that the boy was so resistant to being driven home. (CJF)

(36) With all their bathtime favourites to play with it's not *surprising* it quickly becomes a real highlight for even the youngest toddler in the family. (H07)

(37) I made it *clear* that our policy was to use that money to reduce the level of the council tax over that period of time. (KGX)

(38) I made it quite *plain* he would go before she did, and I rather think he has accepted the status quo. (J19)

(39) He seemed *disappointed* that she hadn't been the one on the music track, but not too much. (GW0)

(40) In any case, we er and we want er no opposition to it, and in fact I I was *surprised* we owned a pack a pack of goodies (J9B)

(41) You've then got to go below that and say, this was achieved through actual and it is the advertising making the public *aware* that we are prepared to employ females providing they're good enough. (J9D)

(42) He made her *aware* they were coming soon. [Not attested]

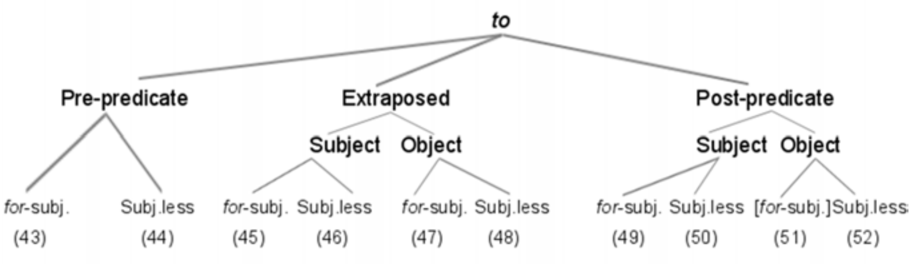

Figure 4. Taxonomy of *to*-clauses.

(43) For investors to expect battle to commence given these circumstances is *ambitious*. (A2V)

(44) To give it to a child would seem unnaturally *insensitive*; to assume it oneself would argue a capacity for self-inflicted wounds beyond even Onyx's nature. (H8Y)

(45) It was *hard* for the donkey to hold a fiddlestick in one hoof, but he managed. (ADM)

(46) It is *hard* to explain my feelings once I did finally set off. (AR3)

(47) These advantages make it *possible* for the miners to produce a high output, of about 3 tonnes per man per shift. (B1H)

(48) Companies find it increasingly *difficult* to attract the right people in the £14,000 to £18,000 range unless they offer a car. (A1J)

(49) For a moment she was speechless, then she said, ‘You mean you're *happy* for Kirsty to come and live with me?’ (JXS)

(50) The company is not *likely* to make a profit either in the second half of this year or next year. (A55)

(51) We need to make the company *easy* for investors to like. [Not attested]

(52) They may know, intellectually, that their partner is less adaptable, but may not find the reality *easy* to deal with. (B3G)

There is an additional class of tokens that do not readily fit into the taxonomies presented in figures 3 and 4. This class of tokens is represented by verb + adjective combinations such as MAKE *clear,* MAKE *sure* and SEE *fit*, as seen in (53)–(55).

(53) Mr Fitton yesterday made *clear* that he had always been acting in an independent capacity in the offer for Eagle, and that it had no connection with Braithwaite. (A26)

(54) He's made *sure* there's nothing to break and lever with, in any case. (G07)

(55) Having stumped up £250 to help sponsor a general practitioners' management meeting in Nottingham, the Royal Bank of Scotland saw *fit* to leave it at that. (A55)

These tokens are clearly not examples of extraposition since they do not include a non-referring *it*. They are, however, very similar to extraposed object clauses, such as (37)–(38) and (47)–(48). These tokens also infringe on the definition of post-predicate clauses as they are represented with a referring pronoun or a noun phrase as subject of the CTA-clause, without being represented by non-referring *it* in object position. Examples such as (53)–(55) are nonetheless treated as a separate class in this study, and are thus seen as exceptional in that they do not fit into taxonomies presented in figures 3 and 4.

Quirk et al. (1985: 1198) note that "[t]he collocations *make sure* and *make certain* are peculiar in that the object is a *that*-clause and always follows the adjectival complement." They also note that with MAKE *clear*, an extraposed object variant is possible (1985: 1198), as illustrated in (56), a variant of (53).

(56) Mr Fitton yesterday made it *clear* that he had always been acting in an independent capacity in the offer for Eagle, and that it had no connection with Braithwaite.

Similar observations on MAKE *sure*, MAKE *certain* and MAKE *clear* are made by Aarts (1992: 150 ff.) noting that insertion of *it* is optional with MAKE *clear*, but not with MAKE *sure* and MAKE *certain*. Aarts (1992: 152) also notes that SEE *fit* and THINK *fit*, followed by a *to*-infinitive, does not allow the insertion of non-referring *it* (cf. (55)). Mindt (2011: 77 ff.), on the other hand, actually finds examples of MAKE *sure* and MAKE *certain* with object extraposition, although these constructions appear to be extremely infrequent.

(57) Apply Sorbie Deep Conditioning Treatment, making it *sure* it is worked into every part of her hair shaft. (Mindt 2011: 97)

(58) The deal makes it almost *certain* that Mr Sam Nujoma, president of Swapo, will become the country's first president. (Mindt 2011: 90)

As alluded to in the quote from Quirk et al. (1985) above, there is also the question whether or not these verb + adjective combinations should be considered as complex verbs (see Mindt 2011: 77 ff. for a comprehensive survey of the literature on *that*-clauses). Huddleston & Pullum (2002) label MAKE *sure* and MAKE *certain* 'verbal idioms,' "whose major element is a verb" (2002: 273). Aarts (1992: 153) considers THINK *fit* as a complex verb, whereas he places examples with MAKE *sure* in a different class based on the fact that MAKE *sure* allows modification, while THINK *fit* does not (e.g. *make quite sure* but not **think very fit*; see Aarts 1992: 152). Mindt (2011: 77 ff.), in a fairly extensive analysis of MAKE *sure*, MAKE *certain* and MAKE *clear* considers MAKE *sure* and MAKE *certain* verb-adjective combinations and not free combinations, whereas she treats MAKE *clear* differently concluding that it "also exhibits some verbal characteristics on the one hand but reveals similarities to MAKE Od *clear* on the other" (2011: 103).

As seen from the discussion in previous studies, the status of these constructions is by no means clear. Some of the combinations allow the variation between inserting or omitting a non-referring *it* in object position, such as MAKE *clear* in (53) and (56), whereas some very infrequently (or possibly never) do so, such as MAKE *sure* and SEE *fit* in (54) and (55). These constructions also seem to be indeterminate as to whether they should be seen as complex verbs or not. Due to this indeterminacy, these constructions have been excluded from the study.

3.3 Purpose clauses vs. complement clauses

Purpose clauses (such as *food is necessary in order to survive*) are distinguished from complement clauses on a semantic and a syntactic basis. Semantically, purpose clauses express a desired result or intention, and are as such oriented towards the future (Quirk et al. 1985: 1108; Huddleston & Pullum 2002: 726–727). Syntactically, purpose clauses are adjuncts rather than complements (Quirk et al. 1985: 1107; see also Jones 1991). Due to these differences, purpose clauses have been excluded from the present study. Infinitival purpose clauses are more frequent than finite purpose clauses. Explicit subordinators of infinitival purpose clauses include *in order* (*to*) (as in (59)) and *so as* (*to*) (as in (60)).

(59) But there are targets, and and I think targets are *important* <u>in order to</u> shape a culture. (J9D)

(60) The dye colour chosen should be *complementary* to the rock colour <u>so as to</u> provide the greatest degree of contrast. (H9S)

There are also examples of ungoverned purpose clauses, i.e. clauses without explicit subordinators (see Huddleston & Pullum 2002: 728). A test for identifying ungoverned purpose clauses is to see whether the explicit infinitival subordinators can be inserted, as illustrated in (61) and (62).

(61) a. THE Government last night signalled its determination to ride out the political storm surrounding its compulsory repatriation of Vietnamese refugees and said it would deport as many as *necessary* to convince thousands more would-be boat people not to seek refuge in Hong Kong next spring. (A9W)

b. THE Government last night signalled its determination to ride out the political storm surrounding its compulsory repatriation of Vietnamese refugees and said it would deport as many as *necessary* <u>so as to</u> convince thousands more would-be boat people not to seek refuge in Hong Kong next spring.

(62) a. Whoever had called him must have been someone very *powerful* to make the head of the KGB jump like that. (CML)
b. Whoever had called him must have been someone very *powerful* in order to make the head of the KGB jump like that.

Examples such as (61)–(62) are excluded from the study since they constitute clear examples of purpose clauses. Infinitival purpose clauses are thus distinguished from infinitival complement clauses in that complement clauses do not express purpose, which is demonstrated by the fact that complement clauses do not allow the insertion of the explicit infinitival subordinators of purpose, as seen in (63).

(63) a. Relevant data are therefore *difficult* to identify and scenario building is one of the few approaches available. (B1G)
b.*Relevant data are therefore *difficult* in order to identify and scenario building is one of the few approaches available.

Finite *that*-clauses of purpose are less frequent than their infinitival counterparts. Unlike *to*-clauses, they cannot be ungoverned but have to be introduced by the explicit subordinators *so* (*that*), as in (64), or *in order* (*that*), as in (65) (see Quirk et al. 1985: 1107–1108; Huddleston & Pullum 2002: 727).

(64) a. The categories in each case are artificial and social rather than self-evident, but they are *necessary* so that we can conduct our affairs in an orderly manner. (H10)
b.*The categories in each case are artificial and social rather than self-evident, but they are *necessary* that we can conduct our affairs in an orderly manner.

(65) a. Such a model is *central* to the management of air quality in order that alternative regulations of emissions can be examined for compliance with air quality standards. (GU5)
b.*Such a model is *central* to the management of air quality that alternative regulations of emissions can be examined for com pliance with air quality standards.

That-clauses of purpose are identified through the fact that they cannot be complemented by a *that*-clause without the explicit subordinators *so that* and *in order that*, as seen in the (b)-versions of (64) and (65).

3.4 Comparative clauses vs. complement clauses

Comparative clauses also constitute a clause type that is similar to, but distinct from, complement clauses. Comparative *to*-clauses are most frequently introduced by *enough* and *too*. Comparative *to*-clauses are termed 'comparative clauses of sufficiency and excess' by Quirk et al. as they "combine the notions of sufficiency or excess with the notions of purpose […] or result" (1985: 1139). More formal synonyms of *enough* and *too* include *sufficient*(*ly*) and *excessive*(*ly*) (Quirk et al. 1985: 1140). Lasnik & Fiengo (1974: 536) show that in infinitival comparative clauses the complements are not complements of the adjectives but rather of *too* and *enough*, as illustrated in (66) and (67). This means that comparative clauses are excluded from this study since they do not meet the requirement of being a complement to the adjectival predicate.

(66) a. The mattress is thin
b.*The mattress is thin to sleep on
c. The mattress is too thin to sleep on

(67) a. The football is soft
b.*The football is soft to kick
c. The football is soft enough to kick
(Lasnik & Fiengo 1974: 536)

In (68)–(70), examples of comparative clauses excluded from this study are given.

(68) a. Then he looked up towards the back, and Shelley knew, with a sudden shock, that he must be looking for her, but she was too *nervous* to stand up. (JYA)
b.*Then he looked up towards the back, and Shelley knew, with a sudden shock, that he must be looking for her, but she was *nervous* to stand up.

(69) a. This poor self-image is clearly not just about looks but also about identity and worth — whether you are 'worth anything', whether your problems are serious or *important* enough to trouble anyone with. (ADG)
b.*This poor self-image is clearly not just about looks but also about identity and worth — whether you are 'worth anything', whether your problems are serious or *important* to trouble anyone with.

(70) a. 'I'm flattered that you think I can read your mind,' she told him sarcastically, 'but I can promise you I wasn't sufficiently *interested* in the outcome to try. (HGM)
b.*'I'm flattered that you think I can read your mind,' she told him sarcastically, 'but I can promise you I wasn't *interested* in the outcome to try.

Examples such as those in (68)–(70) are excluded from this study since they cannot be complemented without *too* (68), *enough* (69) and *sufficiently* (70) expressed in the present context. There are, however, examples of tokens with *too* and *enough* that are included in the study. Consider (71) and (72), with *too* and *enough*, respectively, which are considered to be complement clauses, despite the fact that they look very similar to the comparative clauses in (68) and (69).

(71) a. Yet it is not too *hard* to imagine how a computer parsing program might achieve some such effect. (CM2)
b. Yet it is not very *hard* to imagine how a computer parsing program might achieve some such effect.

(72) a. I was *lucky* enough to be offered this house to rent, so I jumped at the chance. (JXV)
b. I was very *lucky* to be offered this house to rent, so I jumped at the chance.

In both (71) and (72), *too* and *enough* can be removed without causing ungrammaticality. Furthermore, *too* and *enough* have a different meaning in these examples, as illustrated by the fact that they can be replaced by the adverb *very*. In examples such as (71) and (72), the *to*-clause is considered to be the complement of the adjective and not of *too* and *enough*, and examples such as these are thus included in the study.

Unlike *to*-clauses, comparative *that*-clauses are most frequently introduced by *so*. However, similar to *to*-clauses, in comparative *that*-clauses, we find that they are complements of *so* rather than of the adjective, again illustrated by the fact that *so* cannot be removed without causing ungrammaticality:

(73) a. In local authority work today committee work has become so *important* that a great measure of power is given to committees. (B0S)
b.*In local authority work today committee work has become *important* that a great measure of power is given to committees.

There are also examples of valid tokens in which the adjective is preceded by *so*, as illustrated in (74).

(74) a. That's how we can be so *sure* that no slip-ups occur. (HR7)
b. That's how we can be extremely *sure* that no slip-ups occur.

Similar to the examples in (71) and (72), in (74) the *that*-clause is considered a complement of the adjective and not of *so*. This is again illustrated by the fact that *so* can be omitted without causing ungrammaticality, and by the fact that *so* can be replaced by an adverb such as *very* or *extremely*. The distinction between examples such as (73) and (74) is referred to by Mindt (2011: 127 ff.) as a distinction between the resultative, (73), and the explanative construction, (74). In (73) the *that*-clause expresses a result or a consequence of the proposition expressed in the matrix clause, whereas in (74) the *that*-clause provides an explanation of the proposition brought forward in the matrix clause. In this study, instances of the resultative construction are excluded since they do not meet the requirement for adjectival complementation, whereas instances of the explanative construction are treated as ordinary valid tokens.

Although *so* is, by far, the most common word associated with the resultative construction, there are also additional adverbs found in this function, of which only *that* has been attested in the dataset (exemplified in (75)) (see Quirk et al. 1985: 1143 and Mindt 2011: 138).[10]

(75) a. er, it got that *bad* that I thought she was taking the piss out of me. (KCG)
b.*er, it got *bad* that I thought she was taking the piss out of me.

Finally, it should also be noted that there is a close correspondence between infinitival comparative clauses with *enough* and *that*-clauses with *so* (see Quirk et al. 1985: 1142 and Meier 2003: 97), as illustrated in (76).

(76) a. The jet flies *fast* enough to beat the speed record.
b. The jet flies so *fast* that it can beat the speed record.
(Meier 2003: 97)

3.5 Further delimiting the class of complement clauses

Pseudo-cleft clauses (Huddleston & Pullum 2002: 958) have also been excluded from this study. Although they are very similar to extraposed and predicate clauses, they are not included in this study since, structurally, they do not conform to the schema of CTA-clauses and complement clauses outlined in section 3.1. An example of a pseudo-cleft clause construction is given

[10] Mindt (2011: 138) reports that in 99% of the cases she has found of the resultative construction, the adjective is preceded by *so*.

in (77); the corresponding extraposed and pre-predicate versions of (77) are given in (78) and (79).

(77) What is *interesting* is that this capacity is not developed, or its exercise rewarded in Buid society. (CJ1)
(78) It is *interesting* that this capacity is not developed, or its exercise rewarded in Buid society. [Extraposed]
(79) That this capacity is not developed, or its exercise rewarded in Buid society is *interesting*. [Pre-predicate]

Coordination is also a topic that merits a short discussion. In constructions with coordinated adjectives, each adjective is counted as a separate token provided that each coordinated adjective allows complementation in isolation. In (80) and (81), two examples are given in which all coordinated adjectives are counted as separate tokens and thus included in the study (the (b) and (c) variants of (80) and (81) illustrate that the adjectives can be complemented in isolation).

(80) a. It is, however, *right* and *proper* that we do report to this sub-committee on the activities of the staff commission, and that I do in the paper before you. (J9D)
b. It is, however, *right* that we do report to this sub-committee on the activities of the staff commission, and that I do in the paper before you.
c. It is, however, *proper* that we do report to this sub-committee on the activities of the staff commission, and that I do in the paper before you.

(81) a. Nevertheless, it is *dangerous* and indeed *impossible* to push too far the analogies between the heads of the two outer demoiselles and various particular kinds of tribal art. (GUJ)
b. Nevertheless, it is *dangerous* to push too far the analogies between the heads of the two outer demoiselles and various particular kinds of tribal art.
c. Nevertheless, it is *impossible* to push too far the analogies between the heads of the two outer demoiselles and various particular kinds of tribal art.

There are also constructions with coordinated adjectives in which only one adjective allows complementation. As seen in (82), *ambiguous* does not allow complementation by a *to*-clause, and is thus excluded.

(82) a. Although this approach may be more *ambiguous* and more *difficult* to implement than more common practices, it seems to re-

sult in services that are more fluid, creative, and egalitarian and potentially more responsive to client and community needs. (ALN)

b.*Although this approach may be more *ambiguous* to implement than more common practices…

c. Although this approach may be more *difficult* to implement than more common practices…

Finally, constructions which are referred to by Mair (1990: 72 ff.) as 'pseudo tough-movement' are also excluded from this study. They are excluded because technically the predicate of the complement-taking clause is a noun or a pronoun, not an adjective, as illustrated in (83).[11] Note, however, that pseudo tough-movement constructions are very similar to pre-predicate, extraposed and post-predicate clauses with adjectival complement-taking predicates, as illustrated in (84)–(85). The post-predicate clause in (86) is an example of what is typically referred to as 'tough-movement' (Mair 1990: 57 ff.) or 'object-to-subject raising' (Biber et al. 1999: 717), in which the object of the *to*-clause in the pre-predicate clause and the extraposed clause (*the effect*) in (84) and (85) 'raises' to become the subject of the CTA-clause in (86).

(83) The effect is, therefore, an infernally *hard* <u>one</u> to spot…and yet Thom found more alignments with standstills than could realistically be explained by chance. (CET)

(84) To spot the effect is *hard*. [Pre-predicate]

(85) It is *hard* to spot the effect. [Extraposed]

(86) The effect is *hard* to spot. [Post-predicate]

Mair (1990) notes that the pseudo tough-movement construction "is best regarded as a cross between Tough-Movement structures, infinitival relative clauses and elliptical infinitival clauses of purpose" (1990: 72). The affinity of pseudo tough-movement constructions with infinitival relative clauses is illustrated in (87) and (88). As seen, the pseudo tough-movement construction in (87) can be paraphrased with a relative clause construction, as in (88), in which the adjective, not the noun, is being complemented.

(87) He has a *hard* <u>master</u> to satisfy, and nowhere else to go.' (G0M)

(88) He has a master who is *hard* to satisfy.

Again, pseudo tough-movement constructions, such as (83) and (87), are excluded from this study as it is formally the noun, not the adjective, being complemented.

[11] Note that all examples of pseudo tough-movement, such as (83) and (87), was filtered out in data collection process, since the adjective is directly followed by a noun and thus meets one of the filter specifications (see Study I, section 4.2).

3.6 Defining adjectives

As explained in Study I, section 4.3, ambiguity tags were used in extracting the data, in an effort to make sure that as many valid tokens as possible could be accounted for. However, this also meant that there were many examples of participles included in the raw output, which had to be manually removed. In this section, the principles for distinguishing between adjectives and participles will be explained.

There are two tests applied to distinguish between participles and adjectives. If a potentially ambiguous word meets one of the two tests, it is considered to be an adjective. The two tests applied are as follows:

i. Insertion of a degree adverb (Quirk et al. 1985: 414; Huddleston & Pullum 2002: 79)
ii. Replacing BE with SEEM, BECOME, APPEAR, LOOK or REMAIN (Quirk et al. 1985: 414; Huddleston & Pullum 2002: 79)

It should be pointed out that there are several other tests or indicators associated with the distinction between participles and adjectives (see e.g. Gleby 2002: 128 ff.; Quirk et al. 1985: 413 ff.; Huddleston & Pullum 2002: 540 ff.). However, many of these tests cannot readily be applied to the data in this study. For example, one of the central characteristics of adjectival status is for adjectives to be found in attributive and predicative position. However, since all tokens of adjectival complementation by default occur in predicative position, this indicator is not applicable to the data used in this study. Furthermore, indicators such as coordination with a central or near-central adjective (e.g. *clear and unexpected*) are only applicable to a very small proportion of the data in context (cf. Gleby 2002: 134). The approach opted for in this study includes two tests that can be applied to all tokens in context. Test (i) is an insertion test, and test (ii) is a substitution test.

Starting with test (i) (insertion of a degree adverb), Quirk et al. (1985: 415) point out that we have a good indication of adjectival status if the insertion of *very much* is allowed while the participle form disallows *very*. Huddleston & Pullum (2002: 79) add that the insertion test only works in one direction: “if the word in question can be modified by *very* or *too* it must be an adjective, not a verb, but if it can’t be so modified it could be either.” Test (i) is thus an insertion test, in which *very* and *very much* are inserted before a participle-like word. If a given word is found premodified by *very* in the data, the test is redundant and the word is considered to be an adjective.

Test (ii), on the other hand, is concerned with the ability of words to occur with a linking verb other than BE (see Quirk et al. 1985: 414; Huddleston & Pullum 2002: 79). This test is thus a substitution test used to see whether a participle-like word can occur with a linking verb other than BE. Of course, if a participle-like word is found with a different linking verb in the data, the test does not need to be applied.

An illustration of the tests is given in (89)–(92); (89) and (90) illustrate the distinction between past participles (*-ed*) and adjectives, whereas (91) and (92) are examples of the distinction between present participles (*-ing*) and adjectives. *Expected* and *revealing* in (89) and (91) have been classified as participles, and *prepared* and *alarming* in (90) and (92) have been classified as adjectives.

(89) a. But Ron Todd, the union's general secretary, is *expected* to concede defeat gracefully. (A1F)
b.*But Ron Todd, the union's general secretary, is very *expected* to concede defeat gracefully. (i) -> *indeterminate*
c. But Ron Todd, the union's general secretary, is very much *expected* to concede defeat gracefully. (i) -> *indeterminate, indication of participle status*
d.*But Ron Todd, the union's general secretary, seems *expected* to concede defeat gracefully. (ii) -> *confirmed as participle*

(90) a. There is PCB-burning capacity in Sweden, Finland, Germany and France, of which only the last is, like Britain, *prepared* to import such waste. (A1U)
b. ?There is PCB-burning capacity in Sweden, Finland, Germany and France, of which only the last is, like Britain, very *prepared* to import such waste. (i) -> *indeterminate*
c. ?There is PCB-burning capacity in Sweden, Finland, Germany and France, of which only the last is, like Britain, very much *prepared* to import such waste. (i) -> *indeterminate*
d. There is PCB-burning capacity in Sweden, Finland, Germany and France, of which only the last seems, like Britain, *prepared* to import such waste. (ii) ->*confirmed as adjective*

(91) a. They associated it particularly with fertility, and modern research is *revealing* that the moon does influence the reproductive cycles of at least some creatures. (FEV)
b.*modern research is very *revealing* that the moon does influence the reproductive cycles of at least some creatures (i) -> *indeterminate*
c. modern research is very much *revealing* that the moon does influence the reproductive cycles of at least some creatures (i) ->*indeterminate, indication of participle status*
d.*modern research seems *revealing* that the moon does influence the reproductive cycles of at least some creatures (ii) -> *confirmed as participle*

(92) a. It is *alarming* that families with children who accept a social fund loan have to try to make ends meet on 85 per cent of their normal income support entitlement. (G20)
b. It is very *alarming* that families with children who accept a social fund loan have to try to make ends meet on 85 per cent of their normal income support entitlement. (i) -> *confirmed as adjective*
c. ?It is <u>very much</u> *alarming* that families with children who accept a social fund loan have to try to make ends meet on 85 per cent of their normal income support entitlement. (i) -> *already confirmed as adjective*
d. It <u>remains</u> *alarming* that families with children who accept a social fund loan have to try to make ends meet on 85 per cent of their normal income support entitlement. (ii) -> *again, confirmed as adjective*

As seen in (89), when applying test (i), *expected* is indeterminate as regards its adjectival status since it does not readily allow the insertion of *very* (?*very expected* [b]), although we have an indication that it is in fact a participle since *very much* is allowed (*very much expected* [c]). This indication is confirmed when applying test (ii), as we see that BE cannot be substituted by SEEM (**seem expected* [d]). Applying test (i) to *prepared* in (90) we also find that *prepared* is indeterminate (?*very prepared* [b]) (?*very much prepared* [c]). The adjectival status of *prepared* is instead confirmed when applying test (ii), as it does allow substitution of BE with SEEM (*seems prepared* [d]). Applying the same two tests to *revealing* and *alarming* in (91) and (92) we see that *revealing* does not meet any of the two requirements (?*very revealing* [b]) (**seems revealing* [d]), whereas the adjectival status of *alarming* is confirmed after applying the first test (*very alarming* [b]).

Now that the constructions included in the thesis have been introduced and defined, we turn to introducing the material used, which is presented in section 4.

4. Customizing the BNC

All four articles included in this thesis draw on data from the British National Corpus (henceforth BNC; see Burnard 2007). The corpus used in Studies I, II and III consists of a sub-sample of the BNC; Study IV includes data from the spoken component of BNC (see Study IV, section 5). The reason behind this approach is that, for the purposes of these studies, the full BNC was considered to constitute too large a collection of texts with too many different genre distinctions.[12] Restricting the size and possible genre distinctions of corpora is typically achieved through the creation of sub-corpora. A sub-corpus can be created to represent a specified part of a corpus, or it can be created to represent a sample of the corpus as a whole. The two most widely circulated sub-corpora of the BNC, BNC Baby (see Burnard 2008) and the BNC Sampler (see Burnard 1999), are examples of these two different strategies. BNC Baby is a four-part corpus comprising four million words, with one million words sampled from each of four genres. The genres included in BNC Baby (academic texts, fiction, newspaper texts and spoken language) are the same genres adopted by Biber et al. (1999) and described by them as 'core registers' as they "cover much of the range of variation in English, while being restricted to a manageable number of distinctions" (Biber et al. 1999: 24–25).[13] By contrast, the BNC Sampler is a corpus divided into a spoken and a written part, each consisting of one million words, both of which are designed to "mirror the composition of the full BNC as far as possible" (BNC Sampler 2008: 1). The difference between these two sub-corpora is that the BNC Sampler only reduces the size of the corpus, whereas BNC Baby both reduces the size and restricts the number of possible genre distinctions.[14]

There are, however, additional problems connected to the BNC. First, the different genres included in the BNC are represented by widely varying amounts of text, and the genres are represented by a highly variable number

[12] The full BNC consists of 100 million words distributed across 70 different genres (see Lee 2001).

[13] Biber et al. (1999: 15) uses the term 'register' to refer to what in the present study is called 'genre'. See section 4.2 of this introductory survey for how the term 'genre' is defined in the present study.

[14] The BNC Sampler does also reduce the number of possible genre distinctions, as some of the genres will inevitably be left out when downsizing in such a drastic way, but it does so by necessity, and not by a principled sampling decision. See also Lee (2001: 53–54) for a discussion on how successful the compilers of the BNC Sampler actually were in creating a sub-corpus representative of the full BNC.

of text files (see Lee 2002). To give some examples, the genre ‘written fiction drama’ is represented by two text files consisting of a total of 45,757 words, whereas the ‘written miscellaneous’ genre is represented by 500 text files consisting of a total of 9,140,957 words (by way of comparison, the full spoken component of the BNC is represented by 909 text files consisting of 10,334,947 words; see Lee 2002 for details). The same problem exists at the text file level: the smallest text-file (H4L) contains 25 words, whereas the largest (HHV) contains as many as 428,248 words. From a sampling point of view, when sampling whole text files, this is problematic, as different genres will be represented by a different number of text files. The composition of BNC Baby in terms of the number of text files sampled is thus far from ideal, as the genres are represented by different numbers of text files, and the number of text files sampled from most of the genres is very low (see table 1). Consider, for example, that each of the Brown family corpora, with a total size of one million words, includes samples from 500 different text files.

Table 1. Number of text files included in BNC Baby across genres.

Genre	Text-files	Words
Spoken	30	1,017,025
Fiction	25	1,010,279
Academic	30	1,013,256
Newspapers	97	1,001,821
Total	182	4,042,381

Second, an inherent problem with the genres in the BNC is that the genres operate at different levels of abstraction, and Lee (2001: 59) himself admits that his genre scheme is in fact a mixture of genres and sub-genres. To give an example, with the help of the genre classification of the BNC (Lee 2001), the genres included in the BNC can be subdivided into sub-genres. The ‘academic’ genre can be subdivided into six different subject fields (e.g. social sciences, natural sciences and humanities, etc.). Similarly, the genre including newspapers can be categorized as ‘broadsheets’, ‘tabloids’ and ‘regional/local newspapers’ as well as according to section (i.e. ‘commerce’, ‘reportage’, ‘sports,’ etc.) with as many as 15 different genres. However, the ‘fiction’ genre only includes three different sub-genres − ‘fiction prose’, ‘fiction poetry’ and ‘fiction drama’ − making distinctions between traditional sub-genres of fiction such as ‘romance fiction’ and ‘mystery fiction’ impossible. This discrepancy between different genres makes it difficult to compile a sub-corpus in which comparisons across sub-genres are feasible, since the available sub-genres operate at different levels of generality (see section 4.2 for a discussion on the distinction between genres and sub-genres).

The corpus compiled for the purposes of the present study provides a solution to the problems outlined above. In line with Evert (2006), I argue that it is preferable to sample a fixed number of words from each of a large number of text files, rather than to take whole text files of varying sizes (cf. also Biber 1993). This approach has two obvious advantages: (i) increasing the number of different text files sampled, and (ii) making sure that all genres and sub-genres are represented by an equal number of words as well as by an equal number of text files. In genre variation studies that are not concerned with the study of discourse, it is not necessary to sample whole text files, since the unit of analysis is typically found at the sentence or word level. It is in fact preferable not to do so, since the sampling of whole text files restricts the number of different samples (cf. Evert 2006). Furthermore, I argue that, in genre variation studies, it is crucial that the genres included operate at the same level of generality in order for comparisons to be possible at all different levels. It is thus crucial that the sampling frame of each genre (i.e. the population of samples to sample from) is clearly defined, in order to maximize the representativeness of each genre and to maximize comparability at the different levels of generality (cf. Biber 1993; see also Hunston 2008).

4.1 Introducing BNC-15

The corpus used in the present study consists of a three-million-word sample from the XML edition of the BNC. The corpus is named BNC-15 (Kaatari 2012), as it contains 15 sub-genres (see table 2). These 15 sub-genres are distributed across five genres, with three sub-genres representing each genre. The corpus has been compiled to facilitate comparisons across genres at various levels of generality. As such, the corpus allows for comparisons across two different media (speech and writing), two different super-genre distinctions (fiction vs. non-fiction and academic vs. non-academic), five different genres (conversation, novels, academic prose, popular science and news), and three different academic disciplines (humanities, natural sciences and social sciences). In total, the corpus consists of 15 different sub-genres, three from each genre.

Table 2. The different levels of generality of BNC-15.

Medium	Super-genre(s)		Genre	Sub-genres
Speech (600,000 words)	Dialogue (600,000)		Conversation (600,000)	Casual conv. Interview Meeting
Writing (2,400,000 words)	Fiction (600,000)		Novels (600,000)	Adventure Mystery Romance
	Non-fiction (1,800,000)	Non-academic (1,200,000)	News (600,000)	Commerce Reportage Sports
			Popular science (600,000)	Humanities Nat. sciences Social sciences
		Academic (600,000)	Academic prose (600,000)	Humanities Nat. sciences Social sciences

The five genres included in BNC-15 match the genres included in BNC Baby and the genres used by Biber et al. (1999) apart from the addition in BNC-15 of ‘popular science’. In table 2, the genres are ordered (from top to bottom) based on their situational characteristics (see. Biber et al. 1999: 16 for an overview of the situational characteristics of different genres). ‘Conversation’ differs from the written genres by being interactive, although the degree of interactivity differs across sub-genres. ‘Fiction’ is closest to ‘conversation’, as it includes written dialogue. The remaining three written genres all share an informational focus but differ on other grounds. ‘News’ has a wide public audience, whereas ‘academic prose’ targets a specialist audience. ‘Popular science’ is situated between ‘news’ and ‘academic prose’ as it both targets specialist audience and a wider audience, albeit not as wide as ‘news’ (see section 4.2 for a discussion of the distinction between genres and sub-genres based on their differing characteristics). The inclusion of ‘popular science’ in BNC-15 is also motivated by the fact that it creates the possibility of studying the same discipline across different genres, as the same disciplines are represented in academic prose and popular science.

In the compilation process, a stratified sampling procedure was adopted, in which each sub-genre (i.e. each stratum) was sampled individually from a predefined sampling frame (see Biber 1993: 246). Twenty samples of 10,000 words each were drawn from each sub-genre, making sure that each sub-genre and genre was represented by an equal number of words as well as by an equal number of different text files. Following Evert (2006), whole text files were not sampled, in order to increase the number of different samples in the corpus. Furthermore, this sampling approach is preferred since the 'unit of sampling' coincides with the 'unit of measurement', thus minimizing potential clustering effects (cf. Evert 2006: 182–183). Each sub-genre thus contains 200,000 words and 20 different samples, and each genre contains 600,000 words and 60 different samples. The drawback of this approach is that it disqualifies text files that contain fewer than 10,000 words, as these text files cannot be sampled. As seen in table 3, when comparing BNC-15 to BNC Baby, BNC-15 is represented by a larger number of text files despite the fact that it is smaller than BNC Baby in terms of the total number of words.

Table 3. Comparison between BNC Baby and BNC-15.

	BNC Baby		**BNC-15**	
Genre	**Text files**	**Words**	**Text files**	**Words**
Conversation	30	1,017,025	60	600,049
Novels	25	1,010,279	60	600,334
Academic prose	30	1,013,256	60	600,117
Popular science	-	-	60	600,122
News	97	1,001,821	60	600,326
Total	182	4,042,381	300	3,000,948

Note also that the total size of BNC-15 is much closer to its target size of three million words than BNC Baby, which is off by more than 40,000 words.

The sampling process for BNC-15 was conducted through the use of a program that was written specifically for the task at hand.[15] The program was instructed to randomly draw samples from a predefined sampling pool of files (see sections 4.2.1–4.2.5 on the individual genres) and to draw an equal number of samples from the beginning, the middle and the end of text files to minimize any skewing due to textual colligation and clustering (see Hoey 2005; Evert 2006). This consideration was taken based on the fact that different words and linguistic features have a tendency to display a non-random distribution within texts (see also Biber 1993 on linguistic representativeness). The s-unit was taken as the smallest unit sampled in order to ensure, as far as possible, that the corpus would only include full sentences.[16] Furthermore, a cut-off point was established at 10,020 words; the cut-off point works in such a way that if 10,020 words cannot be drawn from a given text-file because the last s-unit is too long, the program rejects the last s-unit and produces a sample with a word count slightly below the cut-off point. In this way, the corpus matches its target size very well at every level of generality. The spread of sample lengths is remarkably limited, as illustrated in table 4.

[15] The program was designed and implemented by Gregory Garretson, Uppsala University.

[16] The s-unit, marked by the <s> tag in the BNC XML edition "contains a sentence-like division of a text" (Burnard 2007: section 2.4) and should only be seen as a proxy for sentences.

Table 4. Smallest and largest sample lengths in words.

	Smallest	Largest	Difference
Text file	9,953	10,020	64
Sub-genre	199,925	200,139	214
Genre	600,049	600,334	285

The fact that BNC-15 matches its target size very closely at all levels is a useful feature, since it means that normalization of figures is not necessary except for in studies of high-frequency items.

4.2 Genres and sub-genres in BNC-15

In this section, the genres and sub-genres of BNC-15 will be introduced. A guiding principle in the process of defining and sampling genres and sub-genres has been to maximize comparability at the different levels of generality. Most crucially, this is achieved by making sure that the genres and sub-genres consist of homogeneous sets of texts that are directly comparable to each other. A list of all the texts included in BNC-15 can be found in the appendix.

Before the individual genres are introduced, a few notes on how the term 'genre' is defined and a few notes on related concepts are in order. The term 'genre' is defined here in accordance with Lee (2001: 46):[17]

> *Genre* is used when we view the text as a member of a category; a culturally recognised artifact, a grouping of texts according to some conventionally recognised criteria, a grouping according to purposive goals, culturally defined.

An essential part of 'viewing the text as a member of a category' is to recognize the fact that categories have different levels of generality. As mentioned in section 4.1, an inherent problem with the BNC is that the genres included operate at different levels of generality. This problem is primarily manifested in the distinction between genres and sub-genres (see Lee 2001: 59). The question then, that needs to be addressed is how we differentiate a genre from a sub-genre. Although Lee (2001, 2002) largely leaves this question open in his classification scheme of the BNC, he discusses the distinctions between the two, drawing on Steen (1999). Table 5 (adapted from Lee 2001: 48) shows a tentative approach based on prototype theory to solve the issue.

[17] See Lee (2001) for a comprehensive discussion of the difference between the terms 'register', 'genre' and 'text-type'.

Table 5. The nesting of different levels of generality.

SUPERORDINATE [SUPER-GENRE]	Newspaper texts	Fiction	Speech
BASIC-LEVEL [GENRE]	News, obituaries, editorials	Novel, Poem, short-story	Conversation, Speeches, Broadcasts
SUBORDINATE [SUB-GENRE]	Commerce, Reportage, Sports	Adventure, Mystery, Romance	Casual conversation, Interview, Meeting

In this approach, genres, as the basic-level category, are seen as maximally distinct from one another in terms of their characteristics, whereas sub-genres represent a level with fewer distinctions (cf. Lee 2001 and Steen 1999). Essentially, this means that genres are easier to distinguish from each other, while sub-genres, at the subordinate level, represent a level with fuzzy boundaries. In BNC-15 this concept is used to distinguish genres from sub-genres, in that genres are seen as clearly distinct from one another in terms of how they are 'culturally recognized and defined,' whereas sub-genres are members of the genres in which they are nested, some better or more prototypical members than others. It should be mentioned that sub-genres are not seen as indivisible in this approach; sub-genres can themselves be further subdivided. It is also important to note that the distinction between genres and sub-genres depends on different types of attributes. For example, novels are distinguished from short-stories on the basis of *form* whereas 'adventure' is distinguished from 'romance' on the basis of *content*. This is also true for the sub-genre level, as 'casual conversation' represents a category that is based on *content* whereas interviews and meetings are based on *form* (see section 4.2.1 for the types of texts included in these sub-genres). Other attributes important for distinguishing between genres and sub-genres include *function* (e.g. 'informational' (news), 'narrative' (novels)) and *audience level* (cf. section 4.1). On the basis of this approach, I saw the need to include a sub-genre level of the 'fiction prose' genre of the BNC in order to make different genres and sub-genres comparable at different levels of abstraction (see section 4.2.2 for details).

A question that is related to the definition of genre is how we can best approach the labelling of different genres. Steen (1999: 111) has the following to say about this:

> A generally valid taxonomy of discourse should not project our expert scientific view of discourse types onto the range of discourse but instead begin with an examination of discourse concepts as they are valid for ordinary language users.

Drawing on this, I argue that a genre label such as ‘popular lore’ (included in both the BNC and the Brown family corpora although with an apparently different definition in the two) is not meaningful as a genre label, since it is highly unlikely to be valid, or indeed meaningful, to ordinary language users, and I suspect the same is true in this case for expert scientists (see also Kessler et al. 1997: 33 for similar criticism of the popular lore genre in the Brown family corpora). Having said this, while the genre labels used in BNC-15 largely follow the labelling introduced by Lee (2001, 2002), they differ on a few grounds, all of which will be discussed in the introductions to the individual genres and sub-genres that now follow.

4.2.1 Conversation

The medium of speech is represented by one genre in BNC-15, with a share of 20 per cent (600,049 words) of the total corpus size. The three sub-genres included in this genre represent three different types of interaction that can be differentiated by their differing degree of formality and by their level of interactiveness. The sampling pool has been carefully restricted for each sub-genre to maximize the homogeneity of each set. The three sub-genres included in ‘conversation’ are ‘casual conversation’, ‘interview’ and ‘meeting’.

‘Casual conversation’ comprises text files from the ‘conversation’ genre (s_conv) of the BNC (see Lee 2001). However, in BNC-15, only text files marked as highly spontaneous are included in this sub-genre. In order to determine the spontaneity of the text files, use was made of the spontaneity (<spont>) tag included in the metadata of the BNC XML edition.[18] Text files that did not include a spontaneity marking or were marked for ‘medium’ or ‘low’ spontaneity were excluded. ‘Casual conversation’ is thus considered to be situated at the informal end of ‘conversation’ due to its high spontaneity and level of interactiveness. It should be noted that text files included in the ‘conversation’ genre of the BNC consist of several different conversations. The spontaneity tag of all individual conversations was checked and text files that included one or more conversations that did not meet the spontaneity criterion were excluded (see Burnard 2007 on the mark-up of the BNC XML edition). Furthermore, only text files that are marked as produced in the 1985–1994 time-period are included.

The text files included in ‘**interview**’ are all from the ‘oral history interview’ genre (s_interview_oral_history) of the BNC. There is an ‘interview’ genre (s_interview) in the BNC, but it only consists of 13 text files of which only seven include 10,000 words or more. The text files in the ‘oral history interview’ genre form a homogeneous set, as the subjects interviewed are all

[18] The spontaneity (<spont>) tag is nested within the locale (<locale>) and activity (<activity>) tags in the BNC XML edition (see Burnard 2007, section 5.3.4).

asked to speak about themselves or about events or themes as they have experienced them (Lee 2001). ‘Interview’ differs most notably from ‘casual conversation’ in that there is an obvious asymmetry in the amount of speech produced by the participants; interviewers typically speak far less than their interviewees. In BNC-15, all text files included in ‘interview’ except for three are from the 1985–1994 time-period. Three text files from the 1975–1984 time-period had to be included due to the limited number of text files available for this sub-genre. Two of these text files are from 1980 and the third is from 1982. ‘Interview’ is considered to be situated at an intermediate position on the formal-informal scale of conversation.

‘Meeting’ comprises text files from the ‘meeting’ genre (s_meeting) of the BNC. The sampling pool was further restricted by only including text files that represent county and council meetings. This was achieved by restricting the sampling pool to those text files marked as belonging to the domain of public institutions (s_cg_public_instit), and by manually scanning the bibliographical details of the text files, restricting the selection to text files that clearly represent county and council meetings. The ‘meeting’ sub-genre of BNC-15 was restricted to county and council meetings because the ‘meeting’ genre of the BNC was found to be too diverse, as it represented a wide variety of different meetings, centering around a number of diverse topics. Only text files from the 1985–1994 time-period are included. ‘Meeting’ is considered to be situated at the formal end of ‘conversation,’ due to the fact that the meetings included are taking place in an institutional context.

4.2.2 Novels

David Lee’s (2001, 2002) genre classification of the BNC has many merits, perhaps the greatest being that some genres are hierarchically nested to allow for super- and sub-genre distinctions to be made. Furthermore, in his BNC World Index, Lee provides an abundance of bibliographical information on the texts included in the BNC. Unfortunately, the ‘fiction prose’ genre (w_fict_prose) constitutes an anomaly in Lee’s rigid classification scheme on two different grounds. First, the ‘fiction prose’ genre is not subcategorized into sub-genres such as ‘mystery fiction’, ‘adventure fiction’ etc. Second, Lee provides very little information (such as keywords and subject headings) on the texts included in the ‘fiction prose’ genre, making further large-scale sub-genre classifications impossible. As a consequence, the ‘fiction prose’ genre is very broad compared to more narrow genres such as ‘newspapers broadsheets editorial’, making comparisons across these genres problematic since they operate at different levels of generality.

For the compilation of BNC-15, it was seen as necessary to somehow work around this imbalance in Lee’s genre scheme, in order to create a more balanced sub-corpus of the BNC. I thus decided to look more closely at what is included in the ‘fiction prose’ genre and try to arrive at a sub-genre classification that included the necessary distinctions. In what follows, the classification process is described and the new sub-genres are introduced.

For most of the written texts included in the BNC, Lee (2001, 2002), in his BNC Index, provides additional information such as keywords, alternative genre headings and general comments about the texts. Unfortunately, the ‘fiction prose’ genre is an exception, since there is very little additional subject and genre information available on the titles included. To supplement the information in the BNC Index, all titles in the ‘fiction prose’ genre that are marked as targeting an adult audience (i.e. excluding children’s stories and titles targeting teenagers, based on the coding of the texts in the BNC World Index) were manually searched for in three different online catalogues, in order to extract as much information about each title as possible. The catalogues consulted include COPAC (an online library catalogue which merges the catalogues of major British and Irish academic libraries), the Library of Congress catalogue (LOC) and the London Library Consortium catalogue (LLC).[19] To varying degrees, all three of these catalogues include subject headings and genre headings for the titles listed. However, not all titles are listed in all three catalogues, and not all catalogue entries include subject and genre headings. COPAC is the most extensive catalogue of the three; it surpasses the other two catalogues in terms of the subject information included, and all titles searched for were found in COPAC. LOC, for its part, has better coverage of genre headings. The London Library Consortium catalogue was useful since it also includes shelfmark and category information, which proved helpful for the genre classification.

The genre classification of the fiction prose genre of the BNC was conducted in the following way: all available information on each title from the three catalogues was copied into a spreadsheet. The spreadsheet follows the same structure as the BNC World Index but has two additional columns labelled ‘GSAFD’ and ‘sub-genre’. The GSAFD[20] column stores the genre heading(s) supplied in COPAC and the LOC catalogue. The sub-genre column is the final classification of a title, based on the subject and genre headings available in the catalogues. The most relevant columns are illustrated in table 6 (changes made to the original BNC Index are marked in bold).

[19] www.copac.ac.uk; www.loc.gov; https://llc.overdrive.com/

[20] Guidelines on Subject Access to individual works of Fiction, Drama, etc. (see the American Library Association 2000).

Table 6. Illustration of the keyword and genre coding.

File	Domain	Genre	**Sub-genre**	COPAC Keywords
AB9	W_imaginative	W_fict_prose	**W_fict_prose_ mystery**	**Crime & mystery FICTION / Mystery & Detective**

Keywords	**GSAFD**	Word Total	Bibliographical Details
---	**Detective and mystery stories, LOC**	40,510	Death of a partner. Neel, Janet. London: Constable & Company Ltd, 1991, pp. ??. 2,674 s-units.

The COPAC keywords are the subject headings available in COPAC. Not all titles are represented with subject headings in COPAC, and there is also a considerable discrepancy in terms of how specific different subject headings are. Where applicable, the GSAFD column lists the genre/form heading(s) assigned and the source of the information is given. The text in table 6 is quite straightforward, as both the subject and the genre information clearly indicate that this particular title can be classified as 'mystery/detective fiction'. Far from all titles include a genre heading, however, and the classification is then dependent on the COPAC subject terms and/or other information such as shelfmark or category information from LLC. In classifying the material, I have only relied on the information illustrated in table 6, with the aim of arriving at an objective categorisation. I have thus steered away from trying to elicit information about the titles from other sources in a manner which would involve reading extracts or summaries and categorising the titles on that basis. Such an approach is not only very time-consuming, but it is also subject to a great deal of subjectivity on the part of the person carrying out the classification. Since many texts have several different subject and genre headings, the classification process is far from straightforward. When introducing the subgenres, problems in distinguishing between them will be discussed. Note that the three sub-genres included in BNC-15 ('adventure', 'mystery' and 'romance') are not the only sub-genres that were identified in the classification process; additional sub-genres such as 'historical fiction' and 'science fiction' were also identified, but these are not included in BNC-15. The three subgenres included in BNC-15 were chosen since they are well-established literary genres, and they all fulfilled the criterion of being represented by a sufficient number of text files of sufficient size (cf. the sampling criteria in section 4.1). Finally, it should be noted that several texts in the BNC were written by the same author (including the use of pseudonyms). The sampling pool was thus manually checked to avoid any author being represented twice in the final corpus.

‘Adventure’ is defined in the following way in the Guidelines on Subject Access to individual works of Fiction, Drama, etc. (GSAFD):

> Use for works characterized by an emphasis on psychical and often violent action, exotic locales, and danger, generally with little character development. (American Library Association 2000: 6)

I have followed the convention used in the Brown family corpora and grouped ‘adventure fiction’ together with ‘western fiction’.[21] ‘Adventure fiction’ is slightly problematic, since it is sometimes hard to clearly distinguish from ‘mystery fiction’. There are titles that include subject headings and/or genre headings that point in the direction of both ‘adventure fiction’ and ‘mystery fiction’. Following this ambiguity, in the ‘adventure fiction’ sub-genre I included titles with the genre heading ‘suspense fiction’, which includes thrillers and spy-stories (see American Library Association 2000: 32). The example given in table 7 is thus classified as ‘adventure fiction’ on the basis that this particular entry includes the GSAFD heading ‘suspense fiction’, which means that the GSAFD genre heading is considered more important than the COPAC keywords for the classification (again, changes made to the original BNC World Index are marked in bold).[22]

Table 7. Classification, adventure.

File	Domain	Genre	**Sub-genre**	COPAC Keywords
BP9	W_imaginative	W_fict_prose	**W_fict_prose_ adventure**	Fiction in English — 1945 — Texts Adventure / thriller Crime & mystery FICTION / Action & Adventure FICTION / Mystery & Detective Horse racing Fiction

Keywords	**GSAFD**	Word Total	Bibliographical Details
---	**Suspense fiction, COPAC**	36,942	The edge. Francis, Dick. London: Pan Books Ltd, 1989, pp. 82-190. 2,650 s-units.

[21] Text category ‘N’ in the Brown family corpora is labelled ‘adventure and western fiction’ (see Francis & Kučera 1964; Johansson et al. 1986; Hundt et al. 1998; Hundt et al. 1999).

[22] Note that, in the example in table 7, the COPAC keywords were already supplied in the BNC Index and are thus not marked in bold.

‘Mystery’ is defined in the following way in the GSAFD: “Use for novels and stories dealing with the detection and solution of crimes” (American Library Association 2000: 24). In the guidelines, there are a number of genre headings subsumed under ‘mystery fiction’, the most common one being ‘detective and mystery stories’ (see table 6). As mentioned, ‘mystery fiction’ is sometimes hard to disentangle from ‘adventure fiction’, but apart from that, there are no clear conflicts with other potential sub-genre headings.

‘Romance’ is here used as a cover term for what is termed ‘love stories and romantic fiction’, with the following definition in the GSAFD: “Use for works dealing primarily with romantic love” (American Library Association 2000: 22). For several of the texts in the ‘fiction prose’ genre, Lee (2001) includes notes such as ‘romance fiction’. It is unclear why romance texts are overrepresented in the inclusion of these types of notes, but most likely it has to do with the fact that there are many different texts in the BNC from one particular publisher, named Mills and Boon. These are also the texts that Lee marked as ‘romance fiction’. The problem, however, is that there is very little information available about these texts; none of them include COPAC keywords, and none of them are represented with any GSAFD genre headings. The only available information about these texts comes from the LLC catalogue in which some of these titles include shelfmark information that indicates that they are indeed ‘romance fiction’. Following this, I have only classified texts from the publisher Mills and Boon as ‘romance fiction’ if the texts include shelfmark information indicating that the title has been classified as ‘romance fiction’. This means that some of the texts that Lee has marked as ‘romance fiction’ have simply been classified as ‘miscellaneous’ in my classification scheme and thus excluded from BNC-15, due to lack of reliable information about these titles.

4.2.3 Academic prose

The ‘academic prose’ genre, along with the ‘popular science’ genre (see section 4.2.4), has been included to allow for distinctions between academic disciplines. Both of these genres are represented by three different subject fields. The subject fields – humanities, natural sciences, and social sciences – were chosen since they are well-known within academia. Having said that, there is still some controversy about in which subject field a particular discipline should be included. For the sake of simplicity, I have followed the categorisation in the BNC Index; the disciplines included within each field are given in the presentation of each sub-genre.

Several studies on adjective complementation suggest that the inclusion of different subject fields is motivated. Peacock (2011), in looking at research articles from eight different disciplines, finds that science writers (biology, chemistry, physics, and environmental science) used extraposed *that*- and *to*-

clauses significantly less than non-science writers (business, language and linguistics, law, and public and social administration). Similarly, Groom (2005) examines extraposed *that* and *to*-clauses across two genres (research articles and book reviews) and across two disciplines (history and literary criticism). Groom finds systematic differences in the distribution of these clause types across genres: research articles are dominated by *that*-clauses, whereas *to*-clauses are more frequent in books reviews. More general observations on the linguistic difference between different disciplines are found in Cao & Fang (2009), who show that, by investigating the type similarity and token similarity of adjectives in the academic genres in the BNC, there is a basic distinction between arts (humanities, social science and politics/law) and sciences (medicine, natural science and technical/engineering).

All texts included in the 'academic prose' genre are from the time-period of 1985–1994. Further, all texts that include an alternative genre label were excluded from the sampling pool, in order to keep the sub-genres as coherent as possible.[23]

The **'humanities'** sub-genre (w_ac_humanities_arts) consists of texts from the disciplines of philosophy, history, literature, art, and music (see Lee 2001: 58). All texts included in this sub-genre are from books except for one text from a periodical. Ideally, the mix between books and periodicals should be more balanced, but there are very few periodicals available for this sub-genre in the BNC. Furthermore, all texts included in this sub-genre are marked for audience level 'high', one of the criteria used in the BNC Index to distinguish between academic and non-academic texts (cf. Lee 2001: 59). The idea behind only including texts that are marked for audience level 'high' is to create as much separation as possible between the academic and the popular science genres.

The disciplines represented in the '**natural sciences**' sub-genre (w_ac_nat_science) include psychics, chemistry, and biology (see Lee 2001: 58). As in the 'humanities' sub-genre, all texts except for one are from books. Unfortunately, the sampling pool for this sub-genre was not as extensive as that for the other two academic sub-genres and, as a consequence, three texts marked for audience level 'medium' were sampled. The remaining 17 texts are all marked for audience level 'high'.

'Social sciences' includes texts from psychology, sociology, linguistics, and social work (see Lee 2001: 58).[24] All texts are marked for a 'high' audience level, and all texts except for one are from books.

[23] Lee (2001) adopts a system with alternative genre labels for texts that are represented by two different disciplines, and/or are ambiguous between two established genres.

[24] Some would perhaps argue that linguistics is better placed in the humanities. For a brief justification of the present categorisation, and a telling example of the problems with the original classification of the texts in the BNC, see Lee (2001: 60).

4.2.4 Popular science

‘Popular science’ comprises the same three subject fields as ‘academic prose’, thus enabling comparisons both across subject fields and across academic vs. non-academic discourse. ‘Popular science’ consists of texts marked as non-academic (non_ac) in the BNC Index. In most instances, the coding of a text as non-academic coincides with audience level ‘low’ or ‘medium’. Another important criterion for distinguishing between academic and non-academic texts in the BNC Index is the name of the publisher (based on the fact that academic publishers form a fairly closed and easily identifiable set) (see Lee 2001: 59). As far as possible, ‘popular science’ includes texts marked for a ‘low’ or ‘medium’ audience level. However, in the subject field of ‘natural sciences’, due to the low number of available texts, six texts were included that are marked for audience level ‘high’ (cf. section 4.2.3). The principle of not including two texts from the same author or publication also had to be abandoned, since two texts from the journal *New Scientist* were included in the ‘natural sciences’ sub-genre. Again, this was necessary due to the shortage of texts in this particular sub-genre.

4.2.5 News

The term ‘news’ is used here, instead of the commonly used term ‘newspaper texts’, since all three sub-genres are in fact sub-genres of news (cf. Ljung 2000). In this sense, ‘newspaper texts’ is seen as a superordinate category to ‘news’ in that it includes newspaper texts that do not strictly qualify as news, such as obituaries, classified advertisements, etc. The three sub-genres included in ‘news’ (‘commerce’, ‘reportage’ and ‘sports’) are all examples of well-established newspaper sections. More broadly, ‘reportage’ can be seen as general news, while ‘sports’ and ‘commerce’ are examples of topic specific news (Ljung 2000: 136).

In previous research, there are arguments supporting the division of ‘news’ into different sub-genres such as ‘reportage’, ‘commerce’ and ‘sports’. Wallace (1977, 1981) provides evidence to differentiate between ‘news’ (i.e. hard news reportage) and ‘sports’, and, in doing so, he finds lexical as well as grammatical differences between these two genres. Jucker (1992) differentiates between ‘home’ and ‘foreign news’ as well as between ‘sports’ and ‘business’. In BNC-15, ‘reportage’ consists of both ‘home’ and ‘foreign news’. It has not been seen as necessary to disentangle these two forms of ‘reportage’, although this has been done in some previous studies (cf. Jucker 1992).

Another aspect of ‘news’ concerns the distinction between hard and soft news. Both Bell (1991) and Ljung (2000) differentiate between hard and soft news, by which the former has to be concerned with reports on recent past events (Ljung 2000: 141) whereas the latter is seen as features (i.e. human interest stories; see Ljung 2000: 137 and Bell 1991: 14). Obviously, the distinction between hard news and soft news is fairly loose. The BNC documentation gives very little information about the news texts included, but the name 'reportage' suggests that we are dealing with hard news rather than soft, although, apart from reading the texts, there is no way of knowing whether some soft news is included. The sub-genres of 'sports' and 'commerce' could also potentially include soft news.

All texts included in 'news' are from national broadsheets, represented by the *Independent*, the *Guardian* and the *Daily Telegraph*. Regional/local newspapers and tabloids have been excluded from the corpus on two different grounds. First, the selection of regional/local newspapers and tabloids in the BNC is severely restricted, meaning that it is impossible to find enough texts spread across different sections. Second, for the present study, the variation across sections is considered to be more important than the variation across newspaper categories (i.e. broadsheet vs. tabloid). Jucker (1992), in his study of premodification and postmodification of NPs across different sections and categories of newspapers, has the following to say about this:

> But they [the results of a cluster analysis] suggest, among other things, that the newspaper section is just as important, if not more important, than the socio-economic categories of the papers to account for the stylistic variation across newspaper language, and also that a simple matrix of socio-economic category and newspaper section will not be enough to account for the interaction. (1992: 136).

One problem with the news texts in the BNC is that each text file consists of several different articles. Moreover, all text files are sampled from the electronically archived editions of the newspapers. One problem concerning this approach is the presence of duplicate text in the sampling. Hoffmann & Evert (2007) have identified 'newspaper text' as the genre in which the problem of duplicates is greatest in the BNC. I have used their list of identified duplicate texts to avoid including these duplicates in the corpus. However, I have observed that there are still duplicate stretches of texts included in my corpus, not previously identified by Hoffmann & Evert (2007).

> The duplicate stretches of text listed below were found by comparing all s-units with 10 or 20 tokens in BNC-XML and printing out any repeated sequences to a file. The list below is the result of checking the more obvious cases. We suspect that there may be quite a few more duplicate stretches of text that could be detected with the help of a more sophisticated method (and more consistent checking). However, we would be surprised if any further texts were fully duplicated. (Hoffmann & Evert 2007: 1)

Unfortunately, these duplicates were identified too late, at the data collection stage, and have thus not been removed from the corpus. What this shows is that the problem of duplicate stretches of text in the 'newspaper' genre is more extensive than previously supposed, and something that researchers have to be aware of when using the BNC.

The sub-genre termed **'commerce'** refers to what is known as the 'commerce section' in the *Daily Telegraph*, the 'business section' in the *Independent* and the 'city section' in the *Guardian*.[25] Although this section has different names in the three newspapers, it forms a homogeneous set since all these sections center around a business/commerce/financial aspect.

'Reportage' consists of both home and foreign material. Unfortunately, the texts included here are from only two different newspapers, the *Independent* and the *Guardian*. The quantities of home and foreign material are more or less evenly balanced, although this was not taken into consideration in restricting the sampling pool.

The texts included in **'sports'** are likewise from the *Independent* and the *Guardian*. Again, due to the sampling of the BNC, only two newspapers are represented in this sub-genre. The only available information about the texts included in 'sports' is that they represent sport material; there is thus no way of knowing which types of sports are represented in the different text files, apart from reading the texts.

Before the relevance of the results are contextualized and discussed in section 6, we now turn to section 5, in which the individual studies are summarized.

[25] The name 'city section' refers to the City of London which functions as the financial district.

5. Summaries

In this section, the four studies included in this compilation thesis are summarized.

5.1 Study I: "The computer as research assistant: A new approach to variable patterns in corpus data" (Garretson & Kaatari, 2014)

This study started out from a desire to collect data for the present thesis in a principled way, but expanded into a study proposing a semi-automated approach to variable patterns in corpus data in general (exemplified in the study with a case study of adjectival complementation). The study centers around SVEP, a computer program designed to implement this semi-automated approach termed "shared evaluation". The term shared evaluation refers to the fact that the computer takes on a more significant role in the evaluation of the data. SVEP is designed to rely on part-of-speech tagging in its identifying of different patterns. Gregory Garretson wrote the program and I provided the linguistic input on which the program operates. In the study, the shared evaluation approach is described and then implemented through SVEP. The study also includes a case study in which data on adjective complementation is presented and discussed.

The evaluation of SVEP in terms of recall and precision shows that SVEP performs very well in terms of recall, but fairly poorly in terms of precision. This means that the program misses very few valid tokens, but it performs more poorly in terms of assigning a valid token to the correct pattern. However, the underlying idea of SVEP, and of the shared evaluation approach, is that tokens are scored based on how well they fit a specified pattern, thus making manual inspection less time-consuming. In fact, a manual evaluation of the recall and the precision of the coding showed that the rate of recall for the automated and manual rounds together was 96.3% (due to tagger and human error) and the precision of the manual coding was 98.7%.

5.2 Study II: "Adjectival complementation: Genre variation, lexical diversity, and meaning" (Kaatari, under review)

In this study, the focus of investigation is on adjectival complementation by extraposed and post-predicate *that*- and *to*-clauses. The study thus includes four different constructions which can be compared across a vertical dimension (i.e. extraposed *that* vs. post-predicate *that* and extraposed *to* vs. post-predicate *to*) as well as across a horizontal dimension (i.e. extraposed *that* vs. extraposed *to* and post-predicate *that* vs. post-predicate *to*). These four constructions are compared in terms of their variation across genres, their lexical diversity, and the meanings expressed by the adjectives most frequently found in each construction, with the aim of establishing which of these two dimensions is dominant.

The results show that while extraposed *that*- and *to*-clauses are remarkably similar in terms of their genre distribution and in terms of their lexical diversity, they are associated with different meanings. Instead, *that*-clauses on the one hand, and *to*-clauses on the other, are similar to each other in terms of meaning, although they show very little similarity in terms of their genre distribution or in terms of their lexical diversity. Further, the results of a correspondence analysis suggest that the relationship between genre variation and meaning is multifaceted, as some semantic groupings can be distinguished by their affinity for different genres while there is also considerable variation regarding the distance between genres, and between the distance of adjectives from the same group of meaning.

5.3 Study III: "Variation across two dimensions: Testing the Complexity Principle and the Uniform Information Density Principle on adjectival data" (Kaatari 2016)

This study tests the applicability of the Complexity Principle (Rohdenburg 1996) and the Uniform Information Density Principle (Jaeger 2010) on adjectival data. These two principles have previously been tested on verbal data only. The design of the study facilitates testing the influence of different factors on the variation between retaining and omitting the complementizer *that* in adjectival constructions across extraposed and post-predicate clauses. The analysis includes five factors, two of which are directly related to the Complexity Principle and the Uniform Information Density Principle, and three of which have been shown in previous research to have an effect on this particular variation. Furthermore, the study includes a multivariate analysis, which examines the combined explanatory power of all of the factors.

The results show that while the factors related to the Complexity Principle and the Uniform Information Density Principle have a clear effect on post-predicate clauses, less clear effects are found for extraposed clauses. I mainly attribute the difference between extraposed and post-predicate clauses to the fact that adjectival post-predicate clauses are very similar to verbal post-predicate clauses, on which the majority of previous research has been conducted. Furthermore, the results show that adjectives complemented by post-predicate clauses are found to appear in entrenched formulaic expressions, such as *I'm sure*, thus rendering the omission of *that* more likely.

5.4 Study IV: "On the syntactic status of *I'm sure*" (Kaatari, forthcoming)

Study IV is a direct extension of one of the findings in Study III, namely that the adjective SURE in general, and the subject-adjective combination *I* ('*m*|*am*) *sure* in particular, is a major trigger for *that*-omission. It thus became apparent that the adjective SURE and the subject-adjective combination *I* ('*m*|*am*) *sure* resemble the verb THINK and the subject-verb combination *I think*, which has also been found to be heavily associated with *that*-omission. The aim of this study was to test whether the syntactic status of *I* ('*m*|*am*) *sure* is similar to that of *I think*, i.e. whether it exhibits the same signs of grammaticalization along two different parameters. The study was designed to closely mirror previous studies on *I think* by testing the ability of *I* ('*m*|*am*) *sure* to (i) occur in clause-medial and clause-final position, and (ii) in its preference for *that*-omission, by comparing the behavior of *I* ('*m*|*am*) *sure* to the results reported for *I think* in previous studies.

First, the results show that the distribution of *I* ('*m*|*am*) *sure* across clause-medial and clause-final position closely mirrors that of *I think*. Second, *I* ('*m*|*am*) *sure* is similar to *I think* in being highly associated with *that*-omission, although the proportion of *that*-omission is slightly lower with *I* ('*m*|*am*) *sure* than it is with *I think*. Due to the fact that SURE is much less frequent than THINK in general, and is also proportionally less dominant among the class of adjectival predicates followed by *that*-clauses than THINK is among verbal predicates, I argue that THINK and SURE are part of the same constructional grammaticalization schema. In this view, each instantiation of *I think* serves not only to entrench the combination *I* + *think*, but also the higher order schematic category of '*that*-complementation construction', thus considering the frequency of *I think* as an important factor for the development of *I* ('*m*|*am*) *sure*.

6. Overview of the main findings

In this section, the main findings of the thesis are discussed. These findings are also contextualized in terms of how they are linked to the aims of the thesis (see section 1.1). The section then ends with some suggestions of possible future extensions of the project.

The broader aim of this thesis was to explain why a certain construction is used in a given context. Several factors have been identified as influencing the propensity of a particular construction to be used. To begin with *that*-clauses, extraposed and post-predicate *that*-clauses are found to be associated with similar meanings, but differ in many other respects: extraposed *that*-clauses are most frequently found in formal genres, they are represented by a higher number of different adjectives (in relation to the total number of tokens; i.e. type-token ratio (TTR)), they are found with fewer instances of *that*-omission, and they are found to be more frequently represented in cognitively complex environments. Post-predicate *that*-clauses, on the other hand, are most frequently found in informal genres, they are represented by fewer different adjectives, they are more frequently found with *that*-omission, and they are found to be infrequently represented in cognitively complex environments. These findings are summarized in table 8.

Table 8. Summary of the results: *that*-clauses.

Constr.	**Genre**	**Meaning**	**TTR**	***−that***	**Comp.**
Extrap.	Formal	Epistemic	High(er)	Low(er)	High(er)
Post-pred.	Informal	Epistemic	Low(er)	High(er)	Low(er)

The difference between extraposed and post-predicate *that*-clauses is further illustrated in (93) and (94).

(93) It is *clear* from the context of such pairs that in every case where the second speaker responds to a proposition of the first speaker with “you know what I mean!” or “you know!”, the second speaker is in fact agreeing with the first. (HXY)

(94) I’m *sure* you’re capable of taking the bull by the horns. (KE6)

The example in (93) is from academic prose (i.e. a formal genre); the adjective *clear* expresses epistemic modality (see Study II); *clear* is furthermore the adjective most frequently complemented by extraposed *that*-clauses (see

Studies II and III). Note also that the complementizer *that* is retained (see Study III), and as regards complexity, the adjective and the complementizer *that* are separated by a prepositional phrase, and the subject of the complement clause (*the second speaker*) is realized by a noun phrase headed by a noun (as opposed to headed by a pronoun; see Study III). In this example, there is also added complexity in that the complementizer *that* and the subject of the complement clause are separated by an additional prepositional phrase.[26]

The post-predicate clause in (94) is from conversation; the adjective *sure* expresses epistemic modality (see Study II); *sure* is the adjective most frequently complemented by post-predicate *that*-clauses (see Studies II and III). In (94) we also find that *that* is omitted, and in conjunction with this we might note that *sure* is also the adjective found to have the highest proportion of *that*-omission (see Studies III and IV). Furthermore, the matrix subject is realized by the pronoun *I*. This is significant in two different ways: first, the context in which the matrix subject is realized by *I* (or *you*) has been found to be significant in triggering *that*-omission (see Study III). Second, the subject-adjective combination *I'm sure* has been found to frequently co-occur with *that*-omission (see Studies III and IV).[27] Finally, we may note that none of the complexity factors found in (93) are found in (94); there is no separation between the adjective and the complementizer *that*, and the subject of the complement clause is realized by a noun phrase headed by pronoun (as opposed to headed by a noun; see Study III).

There are fewer factors concerned with *to*-clauses included in the thesis. The results for *to*-clauses are summarized in table 9.

Table 9. Summary of the results: *to*-clauses.

Construction	**Genre**	**Meaning**	**TTR**
Extraposed	Formal	Dynamic	High(er)
Post-predicate	Less formal	Dynamic	Low(er)

As with *that*-clauses, extraposed and post-predicate *to*-clauses are associated with the same meaning, but differ in terms of the number of adjectives they are represented by (in relation to the total number of tokens; i.e. type-token ratio (TTR)), as well as their genre distribution; extraposed *to*-clauses are represented by a higher number of adjectives, and are found in more formal genres than are post-predicate *to*-clauses. It is thus clear that even though they typically express similar meanings, these two types are used in different contexts.

[26] This type of complexity falls outside of the scope of inquiry in Study II, but may be noted in this particular example.

[27] Note also that *I'm sure* in (94) can be moved to clause-medial (*You're capable, I'm sure, of taking the bull by the horns*) and clause-final position (*You're capable of taking the bull by the horns, I'm sure*) (see section 3.1 of this introductory survey and Study III).

Prototypical examples of extraposed and post-predicate *to*-clauses are given in (95) and (96).

(95) What is particularly interesting about the agreement-marking "you know what I mean" is that it is *possible* — anecdotally, and rather impressionistically — to trace its recent history in London. (HXY)

(96) Mr Cocks said: 'Some four or five-year-olds are perfectly *able* to give a coherent account of themselves, while some 16-year-olds cannot.' (A2P)

In (95), the extraposed *to*-clause is from academic prose. The adjective *possible* is the adjective most frequently complemented by extraposed *to*-clauses, and it expresses dynamic modality (see Study II). We may also note that, although complexity is not addressed in this thesis in connection with *to*-clauses, there is separation between the adjective and the infinitival marker *to* (see also the end of this section for suggestions for future research).

The example in (96) is from news, the genre in which most instances of post-predicate *to*-clauses are found (see Study II). In this respect, post-predicate *to*-clauses are different from the other constructions included in the study in that they are fairly evenly distributed across genres. The adjective *able* is, by far, the adjective frequently complemented by post-predicate *to*-clauses, and it also expresses dynamic modality (see Study II).

A second aim of this thesis was to try to answer the question of whether adjectives behave differently than verbs complemented by *that*- and *to*-clauses. It was shown in Study III, that the predictions of the Complexity Principle (Rohdenburg 1996) and the Uniform Information Density Principle (Jaeger 2010), both based on verbal complementation, were only fully applicable to post-predicate clauses. In Study III, I attribute these findings to the fact that post-predicate *that*-clauses are found in recurring subject–adjective combinations, forming a highly recognizable unit more susceptible to *that*-omission. Further, in Study IV, which is concerned with comparing the subject-adjective combination *I'm sure* to the behavior of *I think*, the results show that SURE is both much less frequent than THINK in general, and is also proportionally less dominant among the class of adjectival predicates followed by *that*-clauses, than THINK is among verbal predicates. Despite this, it was found that *I'm sure* and *I think* behave very similarly in most other respects. In Study IV, I argue that we should treat THINK and SURE as part of the same schematic category, in which the frequency of *I think* not only serves to entrench the combination *I* + *think*, but also the higher order schematic category, and thus the lower order category combination *I* + *sure*.

Finally, there are two possible avenues of future research that I would like to address briefly. First, in his work on the Complexity Principle, Rohdenburg (1995, 1996) shows that verbal *to*-clauses are found in cognitively less complex environments than are verbal *that*-clauses. The reason behind this is that

that-clauses constitute a more explicit grammatical option, due to the fact that "the relatively autonomous finite clause suggests a less immediate relationship between the superordinate and subordinate actions involved" (as compared to *to*-clauses; Rohdenburg 1995: 166). Given the results of Study III, which showed that the complexity factors were largely applicable to post-predicate clauses only, it would be interesting to test (i) whether Rohdenburg's claims are applicable to adjectival data, and (ii) whether this might explain differences between extraposed *that*- and *to*-clauses and between post-predicate *that*- and *to*-clauses.

Second, the results of Study IV suggest that frequency is not as important as previously assumed in grammaticalization studies. In light of this, it might be interesting to investigate whether other subject-adjective combinations are candidates for grammaticalization. Furthermore, such lines of inquiry might also test the claim that only epistemic predicates are candidates for syntactic mobility, and thus grammaticalization (cf. e.g. Hooper 1975; Dor 2005; Boye & Harder 2007).

References

Aarts, Bas. 1992. *Small clauses in English: The nonverbal types*. Berlin: Mouton de Gruyter.

Aijmer, Karin. 1997. '*I think* – an English modal particle.' In Swan, Toril & Olaf Jansen Westwik (eds.). *Modality in Germanic languages: Historical and comparative perspectives*, 1–47. Berlin: Mouton de Gruyter.

American Library Association. 2000. *Guidelines on subject access to individual works of fiction, drama, etc*. Chicago: American Library Association.

Bell, Allan. 1991. *The language of news media.* Oxford: Blackwell.

Biber, Douglas. 1993. 'Representativeness in corpus design.' *Literary and Linguistic Computing* 8: 243–257.

Biber, Douglas. 1999. A register perspective on grammar and discourse: Variability in the form and use of English complement clauses. *Discourse Studies* 1: 131–150.

Biber, Douglas, Stig Johansson, Susan Conrad, Geoffrey Leech & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Boye, Kasper & Peter Harder. 2007. 'Complement-taking predicates: Usage and linguistic structure.' *Studies in Language* 31(3): 569–606.

BNC Sampler: XML edition. 2008. Available at http://www.natcorp.ox.ac.uk/corpus/sampler/sampler.pdf

Burnard, Lou. 1999. *Users reference guide for the BNC Sampler*. Available at http://www.natcorp.ox.ac.uk/corpus/sampler/

Burnard, Lou. 2007. *Reference guide for the British National Corpus (XML edition)*. Available at http://www.natcorp.ox.ac.uk/docs/URG/

Burnard, Lou. 2008. *Reference guide to BNC Baby (second edition)*. Available at http://www.natcorp.ox.ac.uk/corpus/baby/manual.pdf

Cao, Jing & Alex C. Fang. 2009. 'Investigating variations in adjective use across different text categories.' *Research in Computing Science* 41: 207–216.

Diessel, Holger & Michael Tomasello. 2001. 'The acquisition of finite complement clauses in English: A corpus-based analysis.' *Cognitive Linguistics* 12(2): 97–141.

Dixon, Robert M.W. 2005. *A semantic approach to English grammar*. Oxford: Oxford University Press.

Dor, Daniel. 2005. 'Toward a semantic account of *that*-deletion in English.' *Linguistics* 43(2): 345–382.

Elsness, Johan. 1984. 'That or zero? A Look at the choice of object clause connective in a corpus of American English.' *English Studies* 65(6): 519–533.

Evert, Stefan. 2006. 'How random is a corpus? The library metaphor.' *Zeitschrift für Anglistik und Amerikanistik* 54(2): 177–190.

Francis, Nelson W. & Henry Kučera. 1964. *Brown Corpus manual: Manual of information to accompany a standard corpus of present-day edited American English for use with digital computers.* Providence: Brown University. Available at http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM

Givón, Talmy. 1980. 'The binding hierarchy and the typology of complements.' *Studies in Language* 4(3): 333–377.

Gleby, Arne. 2002. *English deverbal adjectives: Their identification and use in a corpus of modern written English*. Gothenburg: University of Gothenburg.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, Adele E. 2002. 'Surface generalizations: An alternative to alternations.' *Cognitive Linguistics* 13(4): 327–356.

Greenbaum, Sidney, Gerald Nelson & Michael Weitzman. 1996. 'Complement clauses in English.' In Thomas, Jenny & Mick Short (eds.). *Using corpora for language research: Studies in honour of Geoffrey Leech*, 76–92. London: Longman.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004. 'Extending collostructional analysis: A corpus-based perspective on alternations.' *International Journal of Corpus Linguistics* 9(1): 97–129.

Groom, Nicholas. 2005. 'Pattern and meaning across genres and disciplines: An exploratory study.' *Journal of English for Academic Purposes* 4: 252–277.

Herriman, Jennifer. 2000. 'Extraposition in English: A Study of the interaction between the matrix predicate and the type of extraposed clause.' *English Studies* 81(6): 582–599.

Hoey, Michael. 2005. *Lexical priming: A new theory of words and language*. London: Routledge

Hoffmann, Sebastian & Stefan Evert. 2007. *Errors, omissions and inconsistencies in the XML-version of the BNC*. Available at http://corpora.lancs.ac.uk/BNCweb/BNC-errors_and_inconsistencies.pdf

Hooper, Joan B. 1975. 'On assertive predicates.' In Kimball, John (ed.). *Syntax and Semantics*, vol. 4, 91–124. New York: Academic Press.

Hooper, Joan B. & Sandra A. Thompson. 1973. 'On the applicability of root transformations.' *Linguistic Inquiry* 4: 465–497.

Huddleston, Rodney. 1971. *The sentence in written English: A syntactic study based on an analysis of scientific texts*. Cambridge: Cambridge University Press.

Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Hundt, Marianne, Andrea Sand & Rainer Siemund. 1998. *Manual of information to accompany the Freiburg-LOB Corpus of British English (FLOB)*. Freiburg: Albert-Ludwigs-Universität Freiburg. Available at http://clu.uni.no/icame/manuals/FLOB/INDEX.HTM

Hundt, Marianne, Andrea Sand & Paul Skandera. 1999. *Manual of information to accompany the Freiburg-Brown Corpus of American English (Frown)*. Freiburg: Albert-Ludwigs-Universität Freiburg. Available at http://clu.uni.no/icame/manuals/FROWN/INDEX.HTM

Hunston, Susan. 2008. 'Corpus compilation and corpus types: Collection strategies and design decisions.' In Lüdeling, Anke & Merja Kytö (eds.). *Corpus linguistics: An international handbook*, vol. I, 154–168. Berlin: Walter de Gruyter.

Jaeger, Florian. 2010. 'Redundancy and reduction: Speakers manage syntactic information density.' *Cognitive Psychology* 61: 23–62.

Johansson, Stig, Geoffrey Leech & Helen Goodluck. 1978. *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English for use with digital computers*. Oslo: University of Oslo. Available at http://clu.uni.no/icame/manuals/LOB/INDEX.HTM

Jones, Charles. 1991. *Purpose clauses. Syntax, thematics, and semantics of English purpose constructions*. Kluwer Academic Publishers: Boston.

Jucker, Andreas H. 1992. *Social stylistics: Syntactic variation in British newspapers*. Berlin: Mouton de Gruyter.

Kaatari, Henrik. 2012. Sampling the BNC – creating a randomly sampled subcorpus for comparing multiple genres. Poster presented at *ICAME* 33, Leuven, Belgium, May 30–June 3, 2012.

Kaltenböck, Gunther. 2011. 'Explaining diverging evidence: The case of clause-initial *I think*.' In. Schönefeld, Doris (ed.). *Converging evidence. Methodological and theoretical issues for linguistic research*, 81–112. Amsterdam: John Benjamins.

Kaltenböck, Gunther. 2013. 'Development of comment clauses.' In Aarts, Bas, Joanne Close, Geoffrey Leech & Sean Wallis (eds.). *The English verb phrase: Investigating recent change with corpora*, 286–317. Cambridge: Cambridge University Press.

Kessler, Brett, Geoffrey Nunberg & Hinrich Schütze. 1997. 'Automatic detection of text genre.' In *EACL '97 Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 32–38.

Kiparsky, Paul & Carol Kiparsky. 1970. 'Fact.' In Bierwisch, Manfred & Karl E. Heidolph (eds.). *Progress in linguistics*, 143–173. The Hague: Mouton.

Lasnik, Howard & Robert Fiengo. 1974. 'Complement object deletion.' *Linguistic Inquiry* 5(4): 535–571.

Lee, David. 2001. 'Genres, registers, text-types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle.' *Language Learning and Technology* 5(3): 37–72.

Lee, David. 2002. *Notes to accompany the BNC World (Bibliographical) Index*. Available at http://www.uow.edu.au/~dlee/home/BNCWIndexNotes.pdf

Leech, Geoffrey. 1991. 'The state of the art in corpus linguistics.' In Aijmer, Karin & Bengt Altenberg (eds.). *English corpus linguistics*, 8–29. London: Longman.

Li, Wentian. 1992. 'Random texts exhibit Zipf's-law-like word frequency distribution.' *IEEE Transactions on Information Theory* 38(6): 1842–1845.

Ljung, Magnus. 2000. 'Newspaper genres and newspaper English.' In Ungerer, Friedrich (ed.). *English media texts – past and present*, 131–149. Philadelphia: John Benjamins.

Mair, Christian. 1990. *Infinitival complement clauses: A study of syntax in discourse*. Cambridge: Cambridge University Press.

Meier, Cécile. 2003. 'The meaning of *too*, *enough*, and *so… that*.' *Natural Language Semantics* 11: 69–107.

Mindt, Ilka. 2011. *Adjective complementation: An empirical analysis of adjectives followed by* that-*clauses*. Amsterdam: John Benjamins.

Noonan, Michael. 2007. 'Complementation.' In Shopen, Timothy (ed.). *Language typology and syntactic description*, second edition, vol. II 52–150. Cambridge: Cambridge University Press.

Palmer, Frank R. 1979. *Modality and the English modals*. London: Longman.

Peacock, Matthew. 2011. 'A comparative study of introductory *it* in research articles across eight disciplines.' *International Journal of Corpus Linguistics* 16(1): 72–100.

Perkins, Michael R. 1983. *Modal expressions in English*. London: Frances Pinter.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. Harlow: Longman.

Rohdenburg, Günter. 1995. 'On the replacement of finite complement clauses by infinitives in English.' *English Studies* 76(4): 367–388.

Rohdenburg, Günter. 1996. 'Cognitive complexity and increased grammatical explicitness in English.' *Cognitive Linguistics* 7(2): 149–182.

Rosenbaum, Peter S. 1967. *The grammar of English predicate complement constructions*. Cambridge, MA: Massachusetts Institute of Technology Press.

Steen, Gerard. 1999. 'Genres of discourse and the definition of literature.' *Discourse Processes* 28(2): 109–120.

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. 'Collostructions: Investigating the interaction of words and constructions.' *International Journal of Corpus Linguistics* 8(2): 209–243.

Stefanowitsch, Anatol & Stefan Th. Gries. 2005. 'Co-varying collexemes.' *Corpus Linguistics and Linguistic Theory* 1(1): 1–43.

Thompson, Sandra A. 2002. '"Object complements" and conversation: Towards a realistic account.' *Studies in Language* 26(1): 125–163.

Thompson, Sandra A. & Anthony Mulac. 1991a. 'A quantitative perspective on the grammaticization of epistemic parentheticals in English.' In Closs Traugott, Elizabeth & Bernd Heine (eds.). *Approaches to grammaticalization*, vol. II, 313–339. Amsterdam: John Benjamins.

Thompson, Sandra A. & Anthony Mulac. 1991b. 'The discourse conditions for the use of the complementizer *that* in conversational English.' *Journal of Pragmatics* 15(3): 237–251.

Torres Cacoullos, Rena and James A. Walker. 2009. 'On the persistence of grammar in discourse formulas: A variationist study of *that*.' *Linguistics* 47(1), 1–43.

Van Bogaert, Julie. 2010. 'A constructional taxonomy of *I think* and related expressions: Acconting for the variability of complement-taking mental predicates.' *English Language and Linguistics* 14(3): 399–427.

Van Bogaert, Julie. 2011. '*I think* and other complement-taking predicates: A case of and for constructional grammaticalization.' *Linguistics* 49(2): 295–332.

Van linden, An. 2012. *Modal adjectives: English deontic and evaluative constructions in synchrony and diachrony*. Berlin: Mouton de Gruyter.

Wallace, William D. 1977. 'How registers register: A study in the language of news and sports.' *Studies in Linguistic Sciences* 7: 46–78.

Wallace, William D. 1981. 'How registers register: Toward the analysis of language use.' *International Review of Applied Linguistics in Language Teaching* 14(4): 267–286.

Zipf, George K. 1932. *Selective studies and the principle of relative frequency in language*. Cambridge, MA: MIT Press.

Appendix

Files included in BNC-15

Academic prose

ID	**Genre**	**Sampling**	**Start**	**End**	**Length**
GUJ	Acad_humanities	MIDDLE	399	709	9956
HY6	Acad_humanities	END	980	1240	9983
A05	Acad_humanities	BEGINNING	1	407	9989
G1N	Acad_humanities	MIDDLE	777	1158	9993
CM2	Acad_humanities	MIDDLE	567	939	9994
HY9	Acad_humanities	MIDDLE	554	853	9996
ANT	Acad_humanities	BEGINNING	1	446	9997
CFX	Acad_humanities	BEGINNING	1	339	9998
CK1	Acad_humanities	END	1230	1732	10003
A6D	Acad_humanities	MIDDLE	684	1119	10004
A1B	Acad_humanities	MIDDLE	917	1350	10006
CM9	Acad_humanities	END	1097	1498	10011
FYT	Acad_humanities	END	884	1185	10012
A6B	Acad_humanities	BEGINNING	1	474	10014
GWM	Acad_humanities	MIDDLE	447	839	10014
HY8	Acad_humanities	BEGINNING	1	403	10014
CS2	Acad_humanities	BEGINNING	1	386	10015
A1A	Acad_humanities	BEGINNING	1	388	10016
BN8	Acad_humanities	END	715	1139	10016
H7S	Acad_humanities	BEGINNING	1	214	10020
FEF	Acad_natural_sci	MIDDLE	477	1021	9975
H9R	Acad_natural_sci	BEGINNING	1	429	9981
H79	Acad_natural_sci	END	1587	2172	9985
GV0	Acad_natural_sci	BEGINNING	1	359	9990
CMH	Acad_natural_sci	BEGINNING	1	409	9995
K5N	Acad_natural_sci	BEGINNING	1	453	9999
FU0	Acad_natural_sci	END	328	815	10000

AML	Acad_natural_sci	MIDDLE	402	959	10003
B2K	Acad_natural_sci	MIDDLE	341	844	10004
HRG	Acad_natural_sci	END	1424	1853	10004
J12	Acad_natural_sci	END	1431	1914	10006
GU5	Acad_natural_sci	MIDDLE	560	909	10007
GU8	Acad_natural_sci	END	1650	2159	10009
H9S	Acad_natural_sci	MIDDLE	727	1353	10012
JXN	Acad_natural_sci	MIDDLE	679	1200	10012
J18	Acad_natural_sci	MIDDLE	591	952	10013
GW6	Acad_natural_sci	BEGINNING	1	452	10017
G1E	Acad_natural_sci	BEGINNING	1	463	10018
H8K	Acad_natural_sci	BEGINNING	1	450	10019
EV6	Acad_natural_sci	BEGINNING	1	514	10020
FR2	Acad_social_sci	END	386	693	9974
FS7	Acad_social_sci	MIDDLE	630	1067	9980
B1G	Acad_social_sci	MIDDLE	677	1074	9982
HTP	Acad_social_sci	MIDDLE	416	718	9983
FR4	Acad_social_sci	BEGINNING	1	380	9987
HXY	Acad_social_sci	MIDDLE	686	1115	9988
CGY	Acad_social_sci	MIDDLE	503	929	9990
APH	Acad_social_sci	MIDDLE	627	997	10000
CKP	Acad_social_sci	END	1021	1397	10003
CLH	Acad_social_sci	BEGINNING	1	398	10004
FAC	Acad_social_sci	MIDDLE	858	1434	10007
HPY	Acad_social_sci	END	713	922	10007
CMN	Acad_social_sci	END	1119	1471	10008
ALN	Acad_social_sci	END	539	908	10009
CBR	Acad_social_sci	END	1394	1828	10009
FED	Acad_social_sci	BEGINNING	1	348	10010
J14	Acad_social_sci	BEGINNING	1	414	10010
A0K	Acad_social_sci	END	918	1216	10012
CJ1	Acad_social_sci	END	1067	1498	10015
G08	Acad_social_sci	END	971	1365	10019

Conversation

ID	**Genre**	**Sampling**	**Start**	**End**	**Length**
KPG	Conv_casual	END	5508	6804	10000

KBR	Conv_casual	END	428	2220	10001
KDB	Conv_casual	BEGINNING	1	1512	10001
KE6	Conv_casual	MIDDLE	4777	6440	10001
KD9	Conv_casual	END	611	2788	10003
KE3	Conv_casual	BEGINNING	1	1442	10003
KP5	Conv_casual	MIDDLE	1365	3131	10003
KBH	Conv_casual	MIDDLE	2734	4191	10004
KE2	Conv_casual	END	8775	10081	10004
KD8	Conv_casual	MIDDLE	4753	5985	10005
KCG	Conv_casual	MIDDLE	988	1979	10006
KCT	Conv_casual	END	12818	14214	10006
KCB	Conv_casual	END	988	2686	10007
KC3	Conv_casual	MIDDLE	1127	2439	10011
KNY	Conv_casual	BEGINNING	1	1704	10011
KR0	Conv_casual	BEGINNING	1	1205	10011
KD0	Conv_casual	BEGINNING	1	1176	10012
KCD	Conv_casual	BEGINNING	1	1649	10014
KDJ	Conv_casual	MIDDLE	220	1149	10014
KBU	Conv_casual	BEGINNING	1	1530	10019
K61	Conv_interview	MIDDLE	22	305	9959
GYV	Conv_interview	BEGINNING	1	339	9963
FYJ	Conv_interview	MIDDLE	52	687	9975
FY5	Conv_interview	MIDDLE	259	754	9988
H4B	Conv_interview	BEGINNING	1	710	10000
HEM	Conv_interview	END	11	576	10000
HYC	Conv_interview	END	386	1455	10000
J8G	Conv_interview	END	82	662	10000
FY1	Conv_interview	BEGINNING	1	635	10001
G62	Conv_interview	BEGINNING	1	656	10003
FY8	Conv_interview	END	220	953	10004
H4C	Conv_interview	MIDDLE	253	846	10004
H5H	Conv_interview	BEGINNING	1	526	10004
HDM	Conv_interview	MIDDLE	26	537	10005
K65	Conv_interview	END	101	1205	10006
FYH	Conv_interview	BEGINNING	1	516	10007
GYW	Conv_interview	MIDDLE	87	545	10010
HDL	Conv_interview	BEGINNING	1	294	10019
GYU	Conv_interview	END	150	753	10020

H5G	Conv_interview	END	411	986	10020
F7V	Conv_meeting	MIDDLE	214	617	9953
KS0	Conv_meeting	MIDDLE	55	432	9970
KGX	Conv_meeting	BEGINNING	1	352	9982
KGM	Conv_meeting	END	559	891	9984
H49	Conv_meeting	END	424	1066	9986
J3T	Conv_meeting	BEGINNING	1	512	9995
KN3	Conv_meeting	END	353	875	9995
J3R	Conv_meeting	END	145	633	9997
JNB	Conv_meeting	END	380	749	9998
JT7	Conv_meeting	MIDDLE	38	424	9998
J41	Conv_meeting	MIDDLE	21	328	10000
JWA	Conv_meeting	BEGINNING	1	415	10001
J9D	Conv_meeting	MIDDLE	147	695	10002
KS1	Conv_meeting	END	513	903	10003
KM8	Conv_meeting	MIDDLE	94	1097	10004
JS8	Conv_meeting	BEGINNING	1	514	10006
J9B	Conv_meeting	BEGINNING	1	506	10007
J3P	Conv_meeting	MIDDLE	195	566	10009
HYX	Conv_meeting	BEGINNING	1	386	10016
JT8	Conv_meeting	END	91	489	10019

Novels

ID	**Genre**	**Sampling**	**Start**	**End**	**Length**
ARK	Fict_adventure	END	1753	2738	10000
G07	Fict_adventure	MIDDLE	1683	2719	10000
GW0	Fict_adventure	END	2634	3361	10000
H8T	Fict_adventure	END	2751	3616	10000
CKC	Fict_adventure	MIDDLE	1509	2367	10001
CR6	Fict_adventure	BEGINNING	1	948	10002
CAM	Fict_adventure	END	1805	2636	10004
BP7	Fict_adventure	END	2169	3155	10005
ECK	Fict_adventure	BEGINNING	1	629	10005
HR7	Fict_adventure	BEGINNING	1	944	10005
HWA	Fict_adventure	MIDDLE	1583	2417	10005
EVG	Fict_adventure	MIDDLE	939	1703	10010
H8M	Fict_adventure	MIDDLE	1460	2333	10011
CEC	Fict_adventure	END	2371	3309	10015

J2G	Fict_adventure	BEGINNING	1	831	10015
GV6	Fict_adventure	END	2336	3048	10016
HTW	Fict_adventure	BEGINNING	1	957	10016
BP9	Fict_adventure	BEGINNING	1	677	10017
HWM	Fict_mystery	BEGINNING	1	833	9998
GVP	Fict_mystery	BEGINNING	1	726	10000
G0M	Fict_mystery	MIDDLE	823	1470	10001
CEB	Fict_mystery	MIDDLE	1138	1966	10002
G01	Fict_mystery	BEGINNING	1	884	10002
J10	Fict_mystery	MIDDLE	1880	3122	10002
FP6	Fict_mystery	BEGINNING	1	630	10003
GUU	Fict_mystery	END	3220	4340	10003
HWL	Fict_mystery	END	2588	3345	10003
GV2	Fict_mystery	MIDDLE	1461	2326	10004
HA2	Fict_mystery	END	2646	3532	10005
CJF	Fict_mystery	MIDDLE	967	1835	10006
K8V	Fict_mystery	BEGINNING	1	1014	10006
CS4	Fict_mystery	MIDDLE	755	1339	10007
BMN	Fict_mystery	MIDDLE	969	1684	10008
GUF	Fict_mystery	BEGINNING	1	1151	10008
HNK	Fict_mystery	MIDDLE	878	1441	10008
HTT	Fict_mystery	END	2334	3100	10014
H8Y	Fict_mystery	BEGINNING	1	704	10015
H98	Fict_mystery	END	2783	3510	10018
APR	Fict_romance	END	1508	2360	9980
AT7	Fict_romance	MIDDLE	1025	1766	9989
H8H	Fict_romance	END	2808	3587	9995
J19	Fict_romance	END	2426	3179	9996
JXU	Fict_romance	BEGINNING	1	785	10000
FPB	Fict_romance	MIDDLE	1174	1935	10002
HGY	Fict_romance	END	3454	4267	10002
JXW	Fict_romance	END	3634	4553	10003
JXV	Fict_romance	MIDDLE	1405	2074	10004
CEY	Fict_romance	END	2557	3390	10005
H94	Fict_romance	MIDDLE	2071	3005	10006
AR3	Fict_romance	BEGINNING	1	350	10008
HGM	Fict_romance	BEGINNING	1	679	10009
JXS	Fict_romance	END	3342	4282	10009

CCM	Fict_romance	BEGINNING	1	914	10013
HGV	Fict_romance	MIDDLE	2575	3446	10015
JYA	Fict_romance	MIDDLE	2072	2999	10017
JYB	Fict_romance	END	3427	4473	10017
HH1	Fict_romance	BEGINNING	1	652	10018
JY1	Fict_romance	BEGINNING	1	496	10018

News

ID	**Genre**	**Sampling**	**Start**	**End**	**Length**
A1S	News_commerce	MIDDLE	35	526	9986
A55	News_commerce	MIDDLE	18	516	9991
A5G	News_commerce	BEGINNING	1	505	9991
A9D	News_commerce	MIDDLE	303	763	9995
AJ9	News_commerce	END	200	700	9996
AHT	News_commerce	BEGINNING	1	520	9998
AJH	News_commerce	BEGINNING	1	475	9998
A26	News_commerce	BEGINNING	1	492	10000
AKU	News_commerce	MIDDLE	116	610	10002
AKL	News_commerce	BEGINNING	1	480	10003
AHJ	News_commerce	END	557	1102	10008
AKD	News_commerce	END	497	980	10011
A94	News_commerce	MIDDLE	7	485	10012
AHB	News_commerce	MIDDLE	156	637	10012
A2H	News_commerce	MIDDLE	53	574	10014
A4F	News_commerce	END	115	620	10014
A2V	News_commerce	MIDDLE	27	531	10015
AJ2	News_commerce	MIDDLE	94	577	10017
A3S	News_commerce	END	18	458	10020
AJX	News_commerce	MIDDLE	117	591	10020
A87	News_report	MIDDLE	28	489	9985
AAU	News_report	END	10	510	9985
A95	News_report	END	15	528	9987
A8K	News_report	MIDDLE	340	842	9988
AA5	News_report	BEGINNING	1	479	9989
A28	News_report	MIDDLE	25	523	9996
A9N	News_report	END	111	613	10001
A1Y	News_report	END	243	742	10002
A4K	News_report	BEGINNING	1	501	10002

A5R	News_report	MIDDLE	132	598	10004
A2A	News_report	END	240	731	10008
A4X	News_report	END	97	554	10010
A9M	News_report	MIDDLE	35	489	10010
A2P	News_report	END	260	764	10011
A1J	News_report	END	152	654	10013
A88	News_report	END	155	637	10017
A9W	News_report	MIDDLE	241	707	10018
A30	News_report	END	242	748	10019
A9E	News_report	END	417	888	10019
A96	News_report	BEGINNING	1	489	10020
A8N	News_sport	END	90	612	9988
A1N	News_sport	BEGINNING	1	479	9993
A5C	News_sport	END	93	579	9995
A90	News_sport	BEGINNING	1	543	10000
A4P	News_sport	MIDDLE	28	519	10002
AAW	News_sport	BEGINNING	1	512	10005
A2E	News_sport	BEGINNING	1	488	10006
AAE	News_sport	MIDDLE	39	507	10006
A3L	News_sport	MIDDLE	97	597	10007
A9R	News_sport	BEGINNING	1	482	10007
AA0	News_sport	END	10	478	10008
A2S	News_sport	BEGINNING	1	493	10010
A4B	News_sport	MIDDLE	23	482	10010
A9H	News_sport	BEGINNING	1	536	10011
A40	News_sport	BEGINNING	1	456	10012
A5U	News_sport	END	242	812	10013
A8C	News_sport	BEGINNING	1	505	10015
A33	News_sport	BEGINNING	1	481	10016
A52	News_sport	BEGINNING	1	499	10017
A80	News_sport	BEGINNING	1	480	10018

Popular science

ID	Genre	Sampling	Start	End	Length
CE7	Pop_humanities	BEGINNING	1	418	9980
H0B	Pop_humanities	BEGINNING	1	390	9984
A06	Pop_humanities	MIDDLE	818	1469	9986
EW9	Pop_humanities	BEGINNING	1	382	9988

CCN	Pop_humanities	MIDDLE	681	1149	9989
FE5	Pop_humanities	END	658	994	9998
EDP	Pop_humanities	MIDDLE	597	1002	10001
HRE	Pop_humanities	MIDDLE	577	976	10002
BNB	Pop_humanities	BEGINNING	1	305	10003
BNN	Pop_humanities	MIDDLE	737	1271	10003
FB1	Pop_humanities	BEGINNING	1	437	10003
B1P	Pop_humanities	BEGINNING	1	357	10004
B2S	Pop_humanities	END	1292	1728	10004
B0G	Pop_humanities	BEGINNING	1	438	10005
C93	Pop_humanities	END	1397	1888	10006
CAC	Pop_humanities	MIDDLE	891	1444	10007
CB9	Pop_humanities	BEGINNING	1	520	10010
CRY	Pop_humanities	END	1540	2061	10017
B09	Pop_humanities	END	1082	1623	10019
B0S	Pop_humanities	END	238	687	10020
ABC	Pop_natural_sci	MIDDLE	676	1110	9979
F9F	Pop_natural_sci	END	1549	2013	9982
B7M	Pop_natural_sci	END	1595	2065	9989
EAW	Pop_natural_sci	END	1482	2003	9997
CJ3	Pop_natural_sci	BEGINNING	1	492	9998
CET	Pop_natural_sci	MIDDLE	670	1120	10000
CRM	Pop_natural_sci	MIDDLE	5748	6156	10000
FEV	Pop_natural_sci	END	1459	1961	10000
CER	Pop_natural_sci	BEGINNING	1	383	10004
H78	Pop_natural_sci	MIDDLE	667	1041	10004
ALW	Pop_natural_sci	BEGINNING	1	501	10006
BLX	Pop_natural_sci	BEGINNING	1	494	10006
H7X	Pop_natural_sci	MIDDLE	671	1121	10006
B72	Pop_natural_sci	MIDDLE	829	1332	10008
ASL	Pop_natural_sci	BEGINNING	1	491	10010
EW6	Pop_natural_sci	END	1057	1539	10011
C9A	Pop_natural_sci	BEGINNING	1	442	10013
ARF	Pop_natural_sci	MIDDLE	590	1038	10017
AMS	Pop_natural_sci	END	1337	1819	10019
CK2	Pop_natural_sci	BEGINNING	1	567	10019
H10	Pop_social_sci	MIDDLE	653	1045	9957
CBJ	Pop_social_sci	MIDDLE	428	779	9958

FAK	Pop_social_sci	END	990	1322	9999
B1H	Pop_social_sci	END	2013	2619	10000
BPK	Pop_social_sci	MIDDLE	659	1074	10000
GVY	Pop_social_sci	MIDDLE	643	1112	10000
B0N	Pop_social_sci	MIDDLE	618	1048	10001
B0W	Pop_social_sci	BEGINNING	1	315	10002
B3G	Pop_social_sci	BEGINNING	1	425	10002
G09	Pop_social_sci	BEGINNING	1	500	10002
HRM	Pop_social_sci	BEGINNING	1	456	10002
ADG	Pop_social_sci	END	1213	1671	10004
G0T	Pop_social_sci	END	1466	1933	10004
H07	Pop_social_sci	END	1439	2089	10006
A77	Pop_social_sci	END	1436	1904	10010
ADM	Pop_social_sci	BEGINNING	1	597	10012
G20	Pop_social_sci	END	1355	1814	10014
FBD	Pop_social_sci	MIDDLE	455	903	10016
EE1	Pop_social_sci	END	619	1062	10017
FA6	Pop_social_sci	BEGINNING	1	399	10019