ACTA UNIVERSITATIS UPSALIENSIS
Studia Anglistica Upsaliensia
126

Editores
Rolf Lundén, Merja Kytö & Monica Correa Fryckstedt

Ylva Berglund

Expressions of Future in Present-day English

A Corpus-based Approach

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in room 16-0043, Humanistiskt centrum, Uppsala, Friday, June 10, 2005 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

Berglund, Y. 2005. Expressions of Future in Present-day English: A Corpus-based Approach. Acta Universitatis Upsaliensis. *Studia Anglistica Upsaliensia* 126. 194 pp. Uppsala. ISBN 91-554-6248-0.

This corpus-based study of the use of expressions of future in English has two aims: to examine how certain expressions of future are used in Present-day English, and to explore how electronic corpora can be exploited for linguistic study.

The expressions focused on in this thesis are five auxiliary or semi-auxiliary verb phrases frequently discussed in studies of future reference in English: *will*, *'ll*, *shall*, *going to* and *gonna*. The study examines the patterned ways in which the expressions are used in association with various linguistic and non-linguistic (or extra-linguistic) factors. The linguistic factors investigated are co-occurrence with particular words and co-occurrence with items of particular grammatical classes. The non-linguistic factors examined are medium (written vs. spoken), text category, speaker characteristics (age, sex, social class, etc.), region and time. The data for the study are exclusively drawn from computer-readable corpora of Present-day English. Corpus analyses are performed with automatic and interactive methods, and exploit both quantitative and qualitative analytical techniques.

The study finds that the use of these expressions of future varies with a number of factors. Differences between spoken and written language are particularly prominent and usage also varies between different types of text, both within spoken and written corpora. Variation between groups of speakers is also attested. Although the linguistic co-occurrence patterns are similar to some degree, there are nonetheless differences in the collocational patterns in which the expressions are used.

Methodological issues related to corpus-based studies in general are discussed in the light of the insights gained from this study of expressions of future.

Keywords: association patterns, corpus linguistics, English, expressions of future, tense, variation

Ylva Berglund, Department of English, Box 527, Uppsala University, SE-75120 Uppsala, Sweden

ISSN 0562-2719
ISBN 91-554-6248-0
urn:nbn:se:uu:diva-5794 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-5794)

To my parents

List of papers

Study I
Future in Present-day English: corpus-based evidence on the rivalry of expressions. *ICAME Journal* 21:7-20 (1997)

Study II
Exploiting a large spoken corpus: an end-user's way to the BNC. *International Journal of Corpus Linguistics* 4(1): 29-52 (1999)

Study III
"You're gonna, you're not going to": a corpus-based study of colligation and collocation patterns of the (BE) going to construction in Present-day spoken British English. In Lewandowska-Tomaszczyk, B. and P. J. Melia (eds.) *PALC'99: Practical Applications in Language Corpora. Papers from the International Conference at the University of Lódz, 15-18 April 1999* Frankfurt am Main: Peter Lang. 161-192 (2000)

Study IV
Utilising Present-day English corpora: a case study concerning expressions of future. *ICAME Journal* 24:25–63 (2000)

Study V
Gonna and *going to* in the spoken component of the British National Corpus. In Mair, C. and M. Hundt (eds.) *Corpus linguistics and linguistic theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20) Freiburg im Breisgau 1999.* Amsterdam: Rodopi. 35-49 (2000)

Contents

List of tables and figures

Tables

The following tables are given in Appendix B

Figures

Abbreviations

BNC	The British National Corpus
Brown	The Brown Corpus of Standard American English
CG	The Context-governed component of BNC and BNC Sampler
DS	The Demographically Sampled component of BNC and BNC Sampler
FLOB	The Freiburg - LOB Corpus of British English
Frown	The Freiburg - Brown Corpus
FUT	Expression(s) of future
Kolhapur	The Kolhapur Corpus of Indian English
LLC	The London-Lund Corpus
LOB	The Lancaster-Oslo/Bergen Corpus of British English
pmw	per million words
Sampler	The BNC Sampler corpus

Acknowledgements

Writing a thesis is supposedly a lonely pursuit. That makes me realise how fortunate I have been to be surrounded by people ready to encourage and support me. Consequently, in addition to the papers for this ‘cumulative’ thesis I have also accumulated a great debt of gratitude, which I would like to express here.

My first thanks must go to my supervisor, Professor Merja Kytö. Her enthusiasm and devotion have been truly inspirational, and her encouragement and support invaluable.

The papers that form the basis of this thesis all stem from material I have presented at international conferences. I am immensely grateful for having been given the opportunity to present my work in this context, and I wish to express my gratitude to participants at these events for the comments, discussions, exchange of ideas and inspiration they have provided. My work would not have been the same without them. In this context I would like to acknowledge the generous financial support from Professor Erik Tengstrands Fond, Rektorsfonden, Ograduerade Forskares Fond at the Department of English that made it possible for me to attend these conferences. The bulk of the work on this thesis was funded through a grant from the Faculty of Languages, Uppsala University. The importance of this support is unmeasurable.

Colleagues at the Department of English have offered a pleasant and stimulating environment in which to work, for which I am grateful. I am indebted to past and current members of the English Linguistics seminar for taking the time to read and comment on my work in various stages. My thanks also go to my fellow corpus linguists and former PhD students Sebastian Hoffmann (Zürich), Oliver Mason (Birmingham) and Patrik Svensson (Umeå), for stimulating collaboration and for offering an outlet for thoughts and ideas. I am indebted to Sebastian and the rest of the Zürich team for letting me work with the BNCweb tool. You know what a difference it has made. Jari Appelgren (Umeå) has offered valuable advice on the use of statistics. Dr. Christopher Williams (Bari) has contributed to the completion of this thesis in more than one way. Working with him has re-kindled my fascination for the English future and opened new ways to approach the topic. Chris has also painstakingly and patiently gone through this manuscript and checked the language.

A number of people have contributed to making my time as a PhD student a pleasant one, both in the office and outside. Dr. Margareta Westergren Axelsson has meant a lot to me, both as a colleague and as a friend. Astrid Sandberg has been a great room mate and a true friend, always there to share good and bad, and any biscuits she had. My colleagues in Oxford have made work here a pleasant and motivating experience. Fellow choir singers have provided welcome relaxation and Anna Swärdh, Kerstin Lindmark and Jimmy Gordon have offered recreational distraction through their letters, emails and Chinese lunches. Thank you all!

I have been privileged to be surrounded by my extended family whose support, encouragement and love have been unfailing. Thank you Barbro, Mårten and the boys for always being there, Britta and Börje Prütz for everything you do for me, and Karin, Calle, Hampus, Herman and Gusten for making me part of the family. Last but in no way least, to my partner Klas Prütz – thank you for just being there.

This book is dedicated to my parents, not only for their abiding support for as long as I can remember but also for their habit of always looking things up. I am convinced that is what sparked my interested in research. My brother's suggestion that this is for both of us has spurred me on.

Oxford April 2005

Ylva Berglund

1. Introduction

1.1 Aim and scope

The present thesis is a corpus-based study of the use of expressions of future in English. The thesis has two aims: to examine how certain expressions of future are used in Present-day English, and to explore how electronic corpora can be exploited for a linguistic study such as this one. Methodological issues related to corpus-based studies in general will be discussed in the light of the insights gained from the study of the expressions of future.

The expressions that are the focus of this thesis are five auxiliary or semi-auxiliary verb phrases frequently discussed in treatments of future reference in English: *will, 'll, shall, going to* and *gonna*, henceforth 'the expressions of future' or FUT (the shorter forms *going to* and *gonna* are used for what in some studies are referred to as *be going to* and *be gonna,* respectively). The thesis examines the patterned ways in which the expressions are used in association with various linguistic and non-linguistic (or extra-linguistic) factors. The non-linguistic factors that have been examined in the present context are medium (written vs. spoken), text category, speaker characteristics (age, sex, social class, etc.), region, and time. The linguistic factors investigated are co-occurrence with particular words and co-occurrence with items of particular grammatical classes.

The thesis comprises the present Comprehensive Summary (henceforth Summary) and five articles. The articles present a number of case-studies and the Summary provides an introduction to the topic, binds together the results presented in the articles, and sheds further light on the significance of the results by turning to additional sets of data. The articles, which are abstracted in Appendix A, are referred to as Studies I-V in the text.

The data for the study were exclusively drawn from computer-readable corpora of Present-day English. The term 'present-day' is used here for, roughly, the period from 1960 to 1995. The starting-point of the time-span was chosen to coincide with the date of publication of the texts included in the Brown corpus (1961), while 1995 is the year of the first public release of the British National Corpus. The corpora that have been used are all principled collections of natural text (see Chapter 3). The corpus analyses were performed with automatic and interactive methods, and exploit both quantitative and qualitative analytical techniques. As such, the present thesis follows the tradition developed in a number of previous corpus-based studies, the essential characteristics of which are summarised by Biber (1996:172) and discussed further in Chapter 3 of this Summary.

1.2 Expressions of future dealt with in the present study

As stated above, the five expressions that are the focus of the present thesis are *will, 'll, shall, going to* and *gonna.* One difference between the modal auxiliary verbs *will, 'll,* and *shall* on the one hand and the semi-auxiliary verbs *going to* and *gonna* on the other, is that the semi-auxiliaries are formed with forms of the auxiliary *be.* This means that the linguistic co-occurrence patterns vary by default (see Chapter 9 for further discussion). The expression *going to* differs from the other FUT in that it is formed with the infinitival marker *to.*

Despite these differences, *will, 'll, shall, going to* and *gonna* can all occur in a similar syntactic construction: with a subject (SUBJ), and an infinitival verb (INF). The subject and infinitival verb can be overt or implied. The overt subject can be found preceding or following the FUT, while the infinitive (if overt) always comes after the FUT, sometimes after the subject. I will refer to this pattern as the FUT paradigm. Some examples illustrate the paradigm:

I *will* sing. SUBJ FUT INF
I *am going to* sing. SUBJ FUT INF
Shall I sing? FUT SUBJ INF
I *won't.* SUBJ FUT ØINF
Gonna sing? ØSUBJ FUT INF

The expressions *be about to* and *be to* are not considered in the present study even though they are paradigmatically similar to the five forms above. One reason for not including these two and other similar expressions in the study is that they are very infrequent. The scarcity of data is likely to make investigations of the usage patterns difficult to interpret and compare to those of the other expressions (in the BNC there are about 0.15 instances per million words of *be about to* in the present tense, compared to about 245 instances per million words of the modal auxiliary *will*). A further reason for not including the *be about to* expression and other similar constructions in this survey is that these expressions are not often included in studies of expressions of future. There is thus little potential for comparison of results between different studies.

Other constructions used for expressing future time reference, such as the simple present ('I *go* to London tomorrow') and present progressive ('I *am going* to London tomorrow'), are at times dealt with in connection with treatments of the future in English (see Chapter 2). They are, however, not included in the present study. The motivation for excluding these constructions is that they differ from the expressions I examine in a number of ways. One syntactic dissimilarity is related to the paradigm. Even though the simple present and present progressive forms can be used to refer to the future in some cases, they are not auxiliary or semi-auxiliary verbs that can be used with infinitival verbs like the expressions in the FUT paradigm presented above. A further argument for not including the simple present and present progressive constructions is that it can be claimed that the future reference in those constructions primarily lies in what Biber et al. (1999:455) refer to as "grammatical contexts". Occurrences of simple present tense referring to future time usually co-occur with time adverbials that refer to the future or are found in adverbial clauses with future time reference (Biber et al. 1999:455). It can thus be argued that it is not primarily the simple present form as such that expresses reference to the future, but the context where the construction is found. The following constructed examples may illustrate this point:

(1a) I *go* to London tomorrow. *future time reference*
(1b) I *go* to London regularly. *habitual action, present time reference*
(1c) I *go* to London. *unknown temporal reference*

(2a) I *will* go to London tomorrow. *future time reference*
(2b) I *will* go to London regularly. *habitual action, future time reference*
(2c) I *will* go to London. *future time reference*

Examples (1a) and (1b) contain a verb in the simple present form, *go.* The sentences have different time references depending on the adverbial that is found in the sentence. When the adverbial refers to the future (*tomorrow*) the sentence has future reference as in (1a). When the adverbial refers to a habitual action (*regularly*), the sentence has a habitual interpretation, as in (1b). Although this habitual action can stretch into the future as well as into the past, I claim that the temporal reference, if any, is primarily present time (see the discussion of speaker's point of primary concern (SPPC) in Chapter 2). When there is no adverbial or further context it is not possible to determine whether the sentence refers to the future or not, as in (1c). In examples (2a-c), the sentences with the auxiliary verb *will* all express that the action is to take place at a time after the present.

Similar reasoning can hold in connection with the use of the present progressive form. The progressive can be used to express future reference but, like the simple present, the future reference suggested in a sentence with a progressive form can be attributed to contextual elements (see, for example, Visser 1973:1922, Williams 2002:198).

Further motivation for not including the simple present and present progressive forms in this study lies in the practical limitations brought about by the search techniques and the corpora available. There is no method available where the instances of the simple present or present progressive forms used to refer to the future can be automatically or semi-automatically identified. It is possible to identify the instances of the simple present or present progressive forms at times (in some tagged corpora). To isolate the instances used for future reference, however, it would be necessary to go through all instances of the forms manually, which is a task that cannot be performed within the scope and time restrictions of the present thesis. Excluding the simple present and present progressive forms from the study illustrates the kind of choices that have to be made when working on large corpora. This is an important point to make in relation to one of the aims of my studies: to explore how electronic corpora can be used for linguistic study.

1.3 Identifying the expressions of future

The expressions of future studied in this thesis are thus the constructions with the five auxiliary or semi-auxiliary verbs in the FUT paradigm: *will*, *'ll*, *shall*, *going to*, and *gonna*. My studies I and IV treat all five expressions. Studies II and V present closer comparisons of two of the variants, *gonna* and *going to,* while Study III focuses on *gonna* and *going to* but also deals with the other expressions to some extent.

The instances included in my studies have been retrieved from electronic corpora with the help of retrieval software, as further presented below (see also Studies I-V). The retrieved instances have been automatically or manually disambiguated to identify the relevant examples. This section describes the basis on which the examples have been identified, and presents the kind of instances included and excluded from the studies. The section is concluded with some examples of included and excluded instances (Table 1.1).

All instances of *will* (including *won't*) where the verb is a modal auxiliary (as opposed to a noun or main verb) were included in my studies (Studies I, III, IV). The instances were classified as modal auxiliaries either through the part-of-speech tagging found in the corpora (Study III), by manual inspection of the instances (Study I), or by a combination of the two methods (Study IV), as further described in the articles.

In the literature, the *'ll* form is sometimes taken to represent both *will* and *shall.* There is, however, convincing argumentation, based on historical and semantic grounds, that the underlying form of *'ll* is *will* (see, for example, *Oxford English Dictionary*, and Quirk et al. 1985:228). I have therefore in some contexts considered the *'ll* and *will* forms in combination, in particular in relation to transcribed, spoken data (for example Study IV). Generally, however, the *'ll* form is treated here as an expression in its own right, in paradigmatic variation with the other four expressions in this study. My studies have shown that the use of *'ll* is different from that of both *will* and *shall,* which further warrants treating the expression individually (see Studies I, III, IV).

All instances of *shall* and *shan't* were included in my studies. It has, at times, been argued that *shall* does not necessarily have a future meaning. It has been suggested that a future interpretation can only be found when the expression is used with first person subjects. 'He shall do it' has been said to imply primarily that the person in question is obliged to perform a certain action. The fact that this action is to take place at a time after the present should then be secondary, and possibly merit less attention. It has, however, been shown that *shall,* as well as other modals, usually has more than one meaning at the same time so that, for example, a future and modal interpretation is possible simultaneously. In the present study, no distinction has been made between instances that are primarily future or primarily modal but all instances of *shall/shan't* have been included. There are at least two reasons for adopting this approach. One is related to semantic interpretation. Even if an instance of the expression can be interpreted as also expressing obligation, it is nevertheless the case that the time when this obligation is to be fulfilled lies in the future, so the temporal reference is thus futural. The second reason is methodological. A study where all instances of the expression are included in the analysis can be easily replicated and critically examined as no counts are based on subjective assessment. As is further noted in Chapter 9, the number of instances of the expression that is not used with first person pronouns is very low, and including these has not seriously affected the overall patterns identified.

As regards *going to,* all instances where *to* is not an infinitival marker have been identified (by manual inspection or through the use of word class annotation) and excluded from the studies. Instances where the expression is used in a past tense context, i.e. with an overt or implied past tense form of the auxiliary *be,* have been excluded from studies where the expression is compared to *will/'ll/shall* (Studies I and IV). Both past and present tense instances have been included in the studies where the focus is on a comparison between *going to* and *gonna* (II, III, V). Instances where the auxiliary *be* is missing have been excluded from Study I.

It seems to be generally understood that *gonna* is a variant form of *going to* (see for example *Longman dictionary of contemporary English* 2000, Quirk et al. 1985). Further support for this is given by the results presented in Study III. The study shows that the two expressions are very similar with regard to their linguistic association patterns (see also Chapter 9). This would motivate pooling together the results obtained for the two constructions. A further motivation is that in corpora of transcribed spoken language, differences in transcription practices may affect the extent to which either expression is found in the text (see Chapter 3). Therefore, in some contexts I conflate the statistics for *gonna* with those of *going to,* for example in comparisons with *will+'ll* (+*shall*). Generally, however, the two expressions are treated separately, in parallel to the treatment of *will* and *'ll.* This is also motivated by the emerging interest in the variation in the use of the expressions (see, for example, Facchinetti 1998, Krug 2000, Poplack and Tagliamonte 2000). Further motivation is provided through the results of my analyses which show that there are certain differences between the *gonna* and *going to* expressions, for example with regard to where and by whom they are used. These differences are examined further in some of my studies (in particular in Studies III and V).

In Studies I and IV, as pointed out above, instances of *gonna* are included when used in a present tense context, i.e. with a present tense form of the auxiliary *be* (explicit or implicit). Past tense instances are examined in some of the studies where the expression is compared to *going to* only (Studies II, III, V). With the exception of Study I, all instances of the expressions have been included whether or not they are used with an overt or implied subject and infinitival verb. Study I only deals with instances with overt subjects and infinitival verbs.

To summarise, Table 1.1 presents the expressions included and excluded in the present study.

Table 1.1. *Examples of expressions included or excluded in Studies I–V*

Expression	**Example**	**Included/excluded**
will	I/you/he/she/X/etc. *will* do it	Included (I, IV, III)
	In my/your/his/her/X's/ etc. *will*	Excluded (not a modal auxiliary)
	I/you/etc. do it at *will* He/she/X does it at *will*	Excluded (not a modal auxiliary)
	I/you/he/she/X/etc. *won't* (do it)	Included (I, IV, III)
'll	I/you/he/she/X/etc. *'ll* do it	Included (I, IV, III)
shall	I/you/he/she/X/etc. *shall* do it	Included (I, III, IV)
going to	I am/you are/he is/etc. *going to* do it	Included (all studies)
	I was/you were/he was/etc. *going to* do it	Included in Studies II, III, V Excluded fr. Studies I, IV (past tense)
	I am/you are/he is/ etc. *going to* London	Excluded (*to* is not the infinitival marker)
gonna	I'm/ you are/he is/ etc. *gonna* do it	Included (all studies)
	I was/you were/he was/etc. *gonna* go it	Included in Studies II, III, V Excluded fr. Studies I, IV (past tense)

1.4 Corpora examined

The selection of the corpora to be used for this thesis was based on a number of criteria. The corpora had to contain Present-day English (as defined in Chapter 3) and be available in computer-readable format. To enable comparisons between my results and those obtained previously by others, I aimed at using only generally available and widely used corpora. That has the further benefit of making it relatively simple for anyone using the same corpora to compare their results to mine. As far as possible, the corpora were also selected to be comparable to each other. The issue of corpus comparability is discussed further in Chapter 3. The corpora I have used are listed below (Table 1.2) and presented further in Chapter 3. Issues related to the choice and use of the corpora are treated in all my studies, but are dealt with more extensively in Studies II and IV.

Table 1.2. *Corpora used*

Corpus		Study					
Short name	Full name	I	II	III	IV	V	Summary
Brown	The Standard Corpus of Present-Day Edited American English	X					X
LOB	The Lancaster-Oslo/Bergen Corpus of British English	X			X		X
Kolhapur	The Kolhapur Corpus of Indian English	X					X
Frown	The Freiburg Brown Corpus of American English						X
FLOB	The Freiburg - LOB Corpus of British English				X		X
LLC	The London-Lund Corpus of Spoken English	X					X
BNC	The British National Corpus		X	X		X	X
Sampler	The BNC Sampler				X		X

1.5 Tools and methods

For this thesis I have thus analysed the occurrences of the five expressions of future in the corpora listed above. A number of corpus tools were used to identify and analyse the data, but the choice of tools and the methods used varied slightly between different studies, as further described below. For all studies, concordances of the expressions were first retrieved automatically. These were then analysed and irrelevant instances (for example *will* used as a noun) were discarded. The remaining instances were then counted and examined further.

The data were primarily analysed quantitatively as frequencies and proportions, and, as far as possible, automatic methods were used in this process. One reason for using automatic methods is that it has not been possible to go through and manually analyse the large quantities of data that are examined in most of the studies. As an example it can be mentioned that the expression *will* alone occurs about a quarter of a million times in the British National Corpus. Manual counting and analysis were used for one of the published studies (Study I). That study showed that although there are benefits with the manual approach (such as making the researcher well acquainted with the material), the difficulties of using such a method for the present purposes were considerable. Not only was it necessary to go through the material several times to make sure the counts were correct, but it was also a laborious and time-consuming task to make alternative calculations in order to compare the results across different variables, something which is easily done with most corpus tools.

The corpus tools that I have used in my studies and this Summary are listed below (Table 1.3). A more detailed presentation can be found in Chapter 3. The use of the tools has also been discussed in Study IV.

Table 1.3. *Tools used*

	Study					
Tools	I	II	III	IV	V	Summary
WordSmith Tools				X		X
Qwick				X		
SARA		X	X	X	X	X
BNCweb		X	X	X	X	X
WordCruncher	X					
MonoConc	general reference					
MicroConcord	general reference					

In addition to the corpus tools, a number of word processing and database programs have been used in the course of the study. Diagrams and some of the tables in the case-studies and the Summary have been created with different versions of Microsoft Excel. Statistical calculations have been performed with Microsoft Excel, GoldVarb, and SPSS. Whenever possible, results in the studies obtained through the use of one program have been verified through comparisons with the results obtained with other programs.

1.6 Plan of the thesis

As mentioned above (Section 1.1), the present thesis consists of a Summary introducing the project and pulling together the results obtained, and five articles (Studies I-V). The articles are abstracted in Appendix A. The Summary is organised as follows:

Chapter 2 contains a brief discussion of previous research on the expressions of future in English. Chapter 3 concerns corpus linguistics and the use of corpora as a methodological framework, pointing to the advantages and disadvantages of this approach. Chapters 4-9 discuss the results of my studies on the use of the expressions of future; variation with non-linguistic factors (Chapters 4-8) and variation with linguistic factors (Chapter 9). Chapter 10 comprises the conclusion where my results are discussed in relation to the methodological framework.

Some raw figures and figures for statistical calculations are presented in Appendix B. The abstracts of the five articles are given in Appendix A. The bibliography contains the references for the present Summary. Additional references are given in the five published studies.

2. Previous research on expressions of future in English

> whereas the past is a chronicle of fact, the future is a tale untold, a mirage that each interprets in his own fashion (Close 1977:146)

2.1 Introduction

Expressions of future have attracted the attention of a number of scholars, and a considerable number of studies deal with different aspects of the topic. Where the expressions of future in English are concerned, much of the discussion has evolved around four main questions:

Is there a future tense?
By what linguistic means is futurity expressed?
How have the expressions of future developed?
What is the meaning of the different expressions of future?

The present thesis differs from much previous work on the English expressions of future (henceforth also 'English future') in that it does not concentrate on these main questions. Instead it examines how certain expressions of future are used in English today, and whether the use can be found to correlate with selected linguistic and non-linguistic factors. Such an empirical study is undoubtedly much indebted to those investigating other issues relating to the English future. To place the present work in the context of previous research, in this chapter I present a brief overview of the central issues that have been dealt with in the literature, summarised under headings based on the four questions listed above.[1] My aim in doing so is neither to agree with nor argue against any particular idea or study. Instead I want to show that my approach was not conceived in isolation but fills a gap in the exist-

[1] The selection of previous research treated here is by no means comprehensive or all-inclusive. The works dealt with have been chosen either because they are frequently referred to in the literature on the subject, or because they offer information that is interesting or relevant in connection to the present study. Previous research of particular relevance to my case studies is also discussed in Chapters 4-9 as well as in the published articles (Studies I-V).

ing body of research on the English future. I hope that the findings about the use of the expressions I present in this thesis will open new vistas to those interested in studying other aspects of the English future in further detail.

2.2 Is there a future tense in English?

There is considerable disagreement among scholars whether English has a future tense. As this study does not aim to take a stand in the question, the term 'expressions of future' is preferred to 'future tense' except in reference to other scholars' work dealing with the topic.

Three main arguments have been put forth against English having a future tense. These arguments can all be said to refer to the unspecific nature of the means available to express future reference in English, or to the fact that no future tense marker can be identified and isolated as having the one and only function of showing future reference.

A first argument against a future tense in English is that this tense is not morphologically marked. Quirk et al. (1985:176), for example, state that they "... prefer to follow those grammarians who have treated tense strictly as a category realized by verb inflection" and thus choose not to use the term 'future tense'. Instead they say that "... certain grammatical constructions are capable of expressing the semantic category of FUTURE TIME". Quirk et al. are not alone in taking this view. Crystal (2003:196) also discusses the issue and claims that English "... has only one inflectional form to express time: the past tense marker (typically *-ed*), ... There is therefore a two-way tense contrast in English: ... present tense vs past tense". Joos (1968:120) suggests that "...tense is our category in which a finite verb ... is either marked with -D or lacks that marker. Then by definition there can be only two tenses". He claims this is a common standpoint: "This is not my invention; for over a century grammarians have been saying that English (like the other Germanic languages and Russian and many others) has only two tenses: past and non-past" (Joos 1968:121). Jespersen (1933:231) states that "[t]he English verb has only two tenses proper, the Present and the Preterite".

Another argument raised against the existence of an English future tense is that there is no single, distinct way to refer to the future in English. Crystal (2003:196) notes that "English has no future tense ending, but uses a wide range of other techniques to express future time". Huddleston and Pullum argues that "...while there are numerous ways of indicating future **time**, there is no grammatical category that can properly be analysed as a future **tense**" (2002:209). As shown in Chapter 1, *will, shall,* and *going to* are all examples of expressions that can be used to refer to the future, and there are also other means available that are at times used in references to the future (such as the simple present form or the present progressive, and adverbial constructions). The fact that future reference can be expressed in a number of ways is taken by some as another argument against a future tense in English. Some refer to the various ways used to express futurity in English as ‘the future tenses’. Joos (1968:120), for example, suggests that “[i]n the folklore, an English verb has a good many tenses”. See Close (1977) for an overview.

A third argument against accepting the notion of an English future tense is that the constructions used for future reference in English do not exclusively express futurity. Palmer, for example, argues against treating *will* and *shall* as the markers of future tense in English not only because they are not the only means available but also because *will* “often does not refer to the future at all” (1974:37). The fact that *will/’ll/shall* also has a modal meaning is a factor brought up against treating them as future tense markers.

There are, however, those who choose to adopt the term ‘future tense’ in their discussion of expressing future reference in English. Close (1977:126) suggests that “[i]n trying to explain the meaning and function of verb phrases having future reference, one is handicapped without the notion of tense or the word ‘tense’ itself”.

Wekker (1976) is perhaps one of the most tenacious proponents for a future tense in English. He argues, for example, that “*will* and *shall* can be used to make purely neutral and factual statements about future events” and that these two expressions thereby fulfil some of the criteria for a future tense (1976:18). He further points out that it is not only the expressions of future that can contain modal overtones, but that “past and present tenses may to a certain extent also be coloured by modality” (1976:18). He maintains that not accepting the notion of a future tense in English on the grounds that the tense is also modal, would then, consequently, suggest that there could be neither present nor past tense either. The consequence of not accepting the future tense because the notion of futurity can be expressed in several ways would, again according to Wekker, “lead us to reject the future tenses of, for example, the Romance languages, which possess a variety of verbal expressions to refer to future time” (1976:19). Other proponents for the existence of an English future tense put forward similar arguments. For an overview of the discussion, see for example Close (1977).

2.3 By what linguistic means is futurity expressed in English?

In the extensive literature on the English future, there is considerable variation in how different authors choose to approach the question and what expressions they include in their discussions. As the following will illustrate, there is a certain degree of agreement with regard to the expressions dealt with, but there is also disagreement. The expressions I have chosen to study (*will, 'll, shall, going to,* and *gonna*) are found in most descriptions of the future in English that I have come across (see also Table 2.1).

In their well-known grammar, Quirk et al. (1985:217) list five constructions which they state are "the most important methods of referring to future time". Those are (with examples from Leech 1971:51):

> *Will/shall* + infinitive ("The parcel *will arrive* tomorrow")
> *Be going to* + infinitive ("The parcel *is going to arrive* tomorrow")
> Present progressive ("The parcel *is arriving* tomorrow")
> Simple present ("The parcel *arrives* tomorrow")
> *Will/shall* + progressive infinitive ("The parcel *will be arriving* tomorrow").

The authors also mention two quasi-auxiliary constructions which are used for referring to the future: *be to* + infinitive ("The parcel *is to arrive* tomorrow") and *about to* + infinitive ("The parcel *is about to* arrive"). In his extensive treatment of the expression of future in English, Wekker (1976) also chooses to focus on the five constructions listed above, and so does Leech (1971, 2004)

Biber et al. (1999) do not deal with the marking of future time at any length in their corpus-based grammar. They note that "there is no formal future tense in English" (1999:456) and state that "future time is typically marked in the verb phrase by modal or semi-modal verbs such as *will, shall,* and *be going to*". Palmer (1965, 1974) deals primarily with four constructions in his discussion of future reference: *will/shall* (+infinitive), *be going to* (+infinitive), simple present, and present progressive, choosing not to treat *will/shall* + progressive infinitive separately in this context. Close considers "five forms that the finite verb phrase can take in referring to the future" (1977:128), adding the *be to* construction to the list of constructions dealt with by Palmer. Crystal includes modal verbs "which also convey a future implication" in his list of the six main ways of referring to the future (2003:224). Huddleston and Pullum claim that "[a]lthough English has no future tense it has a range of constructions which select or permit a future time interpretation" (2002:210), and they add imperative and mandative clauses to the list of constructions.[2] Table 2.1 illustrates what expressions different authors deal with in their treatments of the English future.

[2] The terminology used by Huddleston and Pullum differs slightly from that used by many other authors. Their examples and terms are (2002:210):

i. *Give* her my regards. [imperative]
ii. It is essential [that she *tell* the truth]. [mandative]
iii. The match *starts* tomorrow. [main clause present futurate]
iv. If [she *goes*], I'll go too. [subordinate present]
v. I may/will [*see* her tomorrow]. [bare infinitival]
vi. I intent/want [to *see* her tomorrow]. [*to*-infinitival]
vii. I intend/am [*seeing* her tomorrow]. [gerund-participal]

Table 2.1. *Expressions of future treated by different authors*

	will/ 'll/ shall + inf.	*going to/ gonna* + inf.	simple pre-sent	present pro-gressive	*will/shall* + progressive inf	*be to* + inf.	*be about to* + inf.	modals	other
Examples (see below)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Biber et al. (1999)	x	x						(x)	
Close (1977)	x	x	x	x		x			
Crystal (2003)	x	x	x	x		x		x	
Huddleston & Pullum (2002)	x		x	x				x	x
Leech (1971, 2004)	x	x	x	x	x				
Palmer (1965, 1974)	x	x	x	x					
Quirk et al. (1985)	x	x	x	x	x	x	x		
Wekker (1976)	x	x	x	x	x				

Examples:
1. The parcel *will arrive* tomorrow
2. The parcel *is going to arrive* tomorrow
3. The parcel *arrives* tomorrow
4. The parcel *is arriving* tomorrow
5. The parcel *will be arriving* tomorrow
6. The parcel *is to arrive* tomorrow
7. The parcel *is about to arrive*
8. The parcel *can/may/should arrive* tomorrow
9. *Give* her my regards
It is essential that she *tell* the truth

The expressions *will* and *going to* have received most attention in the literature, and those, together with the variants *'ll, shall,* and *gonna*, are also the expressions I have chosen to study. The reasons why I have decided not to deal with other means of expressing future reference (such as the simple present and present progressive forms) are discussed in Chapter 1.

2.4 How have the expressions of future and their use developed over time?

When the historical development of expressions of future is dealt with in the literature, it is primarily *will, shall,* and *going to* that are discussed. It is generally agreed that *will* stems from the Old English verb *willan* expressing volition. *Shall* comes from the verb of obligation *sculan* and *going to* has developed from a construction consisting of a verb denoting movement (*go*) and the preposition *to.* It has been shown that the use of the *going to* construction to express future reference is more recent than that of the other expressions (the first recorded, albeit debated, instance in the *Oxford English Dictionary* is from 1482). It is argued that the meaning of the different expressions of future can be explained with reference to the origin of the constructions, so that *will* denotes desire/volition, *shall* carries a note of obligation while *going to* indicates movement towards a (real or figurative) goal (see, for example, Bybee and Pagliuca 1987). The fact that the expressions can often be seen to carry a double meaning, both future and modal, can then be explained with reference to their origin. Bybee and Pagliuca (1987:118) argue that "... these modal flavors do not develop from the futurate meaning, but rather, when present, must be interpreted as retentions...". In my studies, all instances of the expressions of future are included in the analysis and no discrimination is made between instances of an expression of future that can be interpreted as more or less modal or futurate (see Chapters 1 and 3 for a discussion of my reasons for this).

The development and spread of the constructions have been treated more extensively in a number of studies, and not least the grammaticalisation of *going to* has received considerable attention; see, for example, Danchev et al. (1965), Bybee and Pagliuca (1987), Danchev and Kytö (1994), Dahl (2000), Krug (2000), Poplack and Tagliamonte (2000), and Hopper and Traugott (2003). I examine the short-term diachronic development of the use of the expressions in one of my papers (Paper IV) and discuss the issue further in Chapter 8.

2.5 What is the meaning of the expressions of future used in English?

Much, if not most, research on the expressions of future focuses on differences and similarities between the expressions, using meaning as a starting-point. The expressions are comprehensively compared and contrasted in a number of ways, within different frameworks and described with a varied set of terminology. In the current section, a brief overview will be given of some of these studies with the rationale of highlighting certain semantic aspects that have been examined in this context.

2.5.1 Tense symbolisation: Reichenbach and Close

One type of analysis that has been frequently used (and not infrequently criticised) where tense interpretation is concerned is the analysis proposed by Reichenbach (1947). In Vikner's words: "Reichenbach (1947:287-298) is widely recognised as the classical attempt at a symbolization of semantic values of verbal tenses" (Vikner 1985:81).[3]

Reichenbach uses three components "to determine the temporal location of a proposition" (Haegeman 1989:296): the moment of speech (S), the time of the event (E), and the reference point (R). Reichenbach does not deal with the difference between various ways of referring to the future in English at any length. That issue is, however, treated by Close (1962, 1977). He introduces the notion of 'the speaker's point of primary concern' (SPPC) and uses 'T' as a symbol for 'the present moment' while 'F' symbolises 'some time in the future'. Reference to the future can then be illustrated by marking the speaker's point of primary concern (SPPC) in relation to the present moment (T) or some time in the future (F). Close assumes that *will* + infinitive is typically used when the speaker's concern is in the future, as illustrated in (1) below. There "conditions for action at F ---> are imagined as fulfilled at F, but conditions for future action are not yet present at T" (1977:148).

(1) We **will** find a cure for cancer
T F
↑
SPPC
[Speaker focuses on a future point/period in time]

When conditions for future action are present at T (indications, expectations), other expressions of future are used, such as in (2) where *going to* is used to illustrate how the speaker's concern is directed to the post-present.

(2) Look at those clouds. It **is going to** pour with rain
T F
SPPC ------------------>
[Speaker concerned with the present, but with attention directed to the future. Indication at T of what will be/happen at F]

Other authors use other means to represent time reference (see, for example, Declerck 1991, Declerck and Depraetere 1995).

[3] For further illustration and discussion of Reichenbach' treatment of tenses, see also Nehls (1988), Haegeman (1989), Davidsen-Nielsen (1990), Gvozdanovic (1991), and Harder (1994).

2.5.2 Speaker's perspective

Close sees a primarily two-fold division between *will,* on the one hand, and other means to express future, on the other. *Will* differs from the other expressions in that the speaker's primary point of concern is *at* F rather than moving *towards* F. The choice between different means to express future reference is thus explained with reference to the speaker (or writer) and how s/he perceives the context or situation.[4]

The importance of the speaker's intention is something which is also noted by other writers (although the terms used may be different). Joos comments on the importance of the speaker in relation to the use of *will*. He claims that "[t]here is no sharp division between clear conditions and the remaining *will* uses, no matter what the subject. The essential point, which unites them all, is that the speaker has selected *will* because of his prudent confidence in the event's eventual occurrence"(1968:157). This can be compared to Close (1977:152), where it is noted that (modal) *will* is a "typical expression of personal attitudes to the future". Palmer (1988:38) claims that *will* "suggests willingness, not mere futurity", and that this refers to something that is envisaged or planned, which can be seen as something involving the speaker's intention. He also notes "... the future expressed by WILL being indicated as a modal judgement by the speaker, in contrast with that expressed by BE GOING TO, which makes an objective statement about current situations relevant to the future" (Palmer 1986:217). Quirk et al. (1985:213-4, 228-9) show how the volitional range of *will* extends from weak to strong volition, and they place "the more usual volitional sense of INTENTION" between the two (Quirk et al. 1985:229).

2.5.3 Distance in time/relation to present time

The distance in time from the moment of speaking to the time which the expression of future refers to has also been noted as a factor governing the choice between different expressions of future. It is generally assumed that *going to* is used when the future event is close in time while *will* is used for events seen to happen further away in time. Coates (1983:200), for example, notes that: " ... BE GOING TO refers to a future event envisaged as happening almost immediately after the moment of speaking. WILL would suggest a more distant, more leisurely future". Quirk et al. (1985:214) provide an explanation to why this may be the case when they suggest that "the association of *be going to* with the present often leads to the assumption that it indicates the proximity of the future event". Haegeman (1989:308) argues that the reason why *going to* is felt to be more immediate is because "*be going to* constrains the processing of the proposition with which it is associ-

[4] I will use the term 'speaker' throughout to denote the person producing the utterance or, in the case of a written text, the author producing the text.

ated to the present context". In her pragmatic account of *going to* and *will*, she maintains that present intention/indication/cause are not specific meanings that are inherent in *going to* but something that can be seen to "follow naturally from the present time contextualization of the construction". She claims that *going to* "... is more immediate not by inherent meaning but because one only has to access present propositions" (1989:308). This can be related to the observation made by Close (1977:148) that *going to* can express "present indications of what the future may bring", which can be illustrated as in example (2) above, repeated here as (3). There the reference (implicit or explicit) to present conditions sets the context for the expression. In Close's terminology, the conditions for future action are fulfilled at the present moment (T) and the speaker's point of primary concern (SPPC) is directed towards a point in the future (F). There are present indications for something which will happen in the future.

(3) Look at those clouds! It is *going to* pour with rain.
T F
SPPC------------------------>
[Present indication (= the clouds) present at the moment of speaking]

The notion that the present is in some way in focus when *going to* is used can also be found in works by Palmer. He notes that "[i]n most cases BE GOING TO can be interpreted in terms of current orientation ... in that it relates to the future *from the standpoint of the present*" (1988:146, my italics). Or, as said in relation to non-past forms of *going to* in a previous work by the same author: "there are features of the present time that will determine future events" (Palmer 1979:121). Nehls (1988:300), notes that "[i]f signs of the future event can be observed in the present, GOING TO+SIMPLE INFINITIVE is used; ... These signs can be present either in the surroundings of the speaker ... or they can be present in the speaker's mind". The importance of the speaker and his/her mind is here referred to once again.

Example (3) can be compared and contrasted to example (4). In (4) the conditions for the actualisation of the situation are not present at the present moment (T) but imagined as fulfilled at some time in the future (F), which is the point of concern (SPSS) for the speaker. The expression *will* is consequently chosen.

(4) It *will* rain (by the time we get there).
T F
↑
SPCC
[Conditions fulfilled at a moment in the future]

This fits in well with Palmer's claim that "[*will/shall*] occurs where only the future is involved" (1988:148) and the statement made by Nehls that "[t]he WILL-futures, on the other hand, merely predict the taking place of an event" (1988:301).

It is, however, not only the time and relation to the present that decides which expression is chosen. Nehls claims that "... the expressions of future time in English cannot be exclusively explained by referring to a chronological division of linear time; modal overtones are always implied" (1988:303). He is not the only one to point to the importance of modality in the study of expressions of future. According to Quirk et al., for example, "[f]uturity, modality, and aspect are closely interrelated" (1985:213).

2.5.4 *Will* vs. *going to*: "elliptical" uses

Binnick (1972) argues that sentences with *will* are often felt to be "elliptical" unless there is something to make them complete. In his reading, a sentence such as '*The rock will fall*' is incomplete but may be completed by another clause after which the overall sentence is not elliptical. Such a completing clause may express a condition or cause ('The rock will fall *if you push it*') or refer to a point in time ('The rock will fall *at three o'clock*'). Binnick maintains that *going to* is never felt to be elliptical, and that this is "because it [*going to*] does not depend on hypothetical conditions; if no condition is explicit, it is assumed that all conditions for the future event have been met" (1972:7). Binnick's examples, (5) and (6) below, illustrate how the choice of expression affects the meaning of the otherwise similar sentences.

(5) When we build this new one, we*'ll* have nine houses up.
(6) When we build this new one, we are *going to* have nine houses up.

Here (5) is said to express that we now have eight houses completed, while (6) denotes that the number of presently completed houses is nine. Binnick explains this with reference to causal relationship. In (5) the use of *will* indicates that "the new construction *results* in our having nine houses up, therefore we *now* have only eight" (1972:5, italics mine). In (6), however, the use of *going to* signals that all conditions are fulfilled and that having the nine houses up is not conditional. If an attempt were made to adopt Close's model to this example, it could be claimed that in (5), the choice of *will* indicates that the SPPC is at F. The conditions are fulfilled in the future and that is the time when the houses will be nine in number. In (6), the speaker's point of concern refers to 'the present moment' (T) when all conditions have been met, ('we are going to have nine houses up'), and that sets the scene for the future (F).

(5) (When we build this new one,) we*'ll* have nine houses up.

```
              T          F
                         ↑
                                      SPPC
```

(6) (When we build this new one,) we are *going to* have nine houses up.

```
              T          F
              SPPC  ---------------->
```

Haegeman (1989:306) replies that sentences like ‘The rock will fall’ are felt to be incomplete (or, in Binnick’s terminology: elliptical) not because sentence-internal material is missing but because there are no clues “as to the contextualization of the utterance”. Although not necessarily meant as such, her comment can be taken to introduce an important issue – the difficulty of interpreting an utterance found or produced in isolation.

2.6 Concluding remarks

This brief overview has shown that there exists a wide and varied literature dealing with issues related to the expressions of future in English. In addition to arguments relating to whether there is a future tense in English and how future reference can be and has been expressed in English, the meaning of the expressions has been extensively discussed by scholars and so has the choice, or suitability of one expression compared to another.

Mair (1997) points to a specific difficulty in interpreting the meaning of the expressions of future:

> In the analysis of authentic examples from corpora, moreover, there is the additional trap of circularity, as the explanandum, the use of *going to*, is itself often the only reason why we feel forced to read a present intention or cause into the example. (Mair 1997:1538)

By describing how the expressions occur in the corpus data examined I hope to contribute to a fuller understanding of the use of future reference in English without being caught in the ‘trap of circularity’ suggested by Mair.

My work is similar to many previous studies in that it compares the use of different expressions. My studies differ from most other studies, however, in that I do not examine semantic features of the expressions, nor look for explanations to patterns of variation in the meaning of the expressions as such. Instead, I have considered it of interest to see what information about the use of these expressions can be found by examining other factors than those attributed to the meaning of the expression. That such information can be useful for further work on the semantics of future reference in English is illustrated by the following quotes, which suggest that scholars, at times, see no definite difference in the meanings of these expressions.

> Inevitably, there are plenty of contexts in which either *will/shall* or BE GOING TO is equally appropriate. ... it must not be assumed that the choice of one form or the other can always be explained. ... although they may be seen to be semantically different, it does not follow that there are no situations in which either is equally appropriate, so that in a particular context there is no explanation why one is used rather than the other. (Palmer 1988:148)

> ... at the level of sentence meaning *be going to* and *shall/will* are equivalent.... (Haegeman 1989:291)

> However, it must be made clear that WILL and BE GOING TO are often interchangeable, or can be substituted for each other with only the faintest change of meaning (Coates 1983:201)

> BE GOING TO ... is often either in contrast or in free variation with WILL (Palmer 1988:146)

> In most cases there is no demonstrable difference between *will/shall* and BE GOING though many scholars have looked without success for one (Palmer 1974:163)

In what follows, the use of corpus linguistics as an analytical framework will be discussed, with particular reference to my studies of the expressions of future in English.

3. Corpus linguistics as an analytical framework

> The corpus has the benefit of rendering public the point of view used to support a theory. Corpus-based observations are intrinsically more verifiable than introspectively based judgements (McEnery and Wilson 2001:14)

3.1 Introduction

As stated in Chapter 1, this thesis aims not only to discuss expressions of future as such but also proposes to deal with issues related to the use of corpora. The present chapter is concerned with certain methodological and practical issues of relevance for the application of corpus linguistics as an analytical framework in general and in relation to my investigation of the use of the expressions of future in particular.

Even though computer-readable corpora have existed for only about 40 years, a great number of studies have been based on electronic corpus data. Biber (1996) has summarised some characteristics that he suggests are shared by many of these studies. His list can be found below (see also the comprehensive bibliographies of studies based on the so-called ICAME corpora compiled by Altenberg (1986, 1993, 1998), available through ICAME[5]).

> Corpus-based studies
> --- are empirical, analyzing the actual patterns of use in natural texts;
> --- utilize a large and principled collection of natural texts (i.e., a "corpus") as the basis for analysis;
> --- make extensive use of computers for analysis, using both automatic and interactive techniques;

[5] ICAME = International Computer Archive of Modern and Medieval English. As stated in the introduction to the *ICAME Journal* 24 (2000): "[t]he aim of the organization is to collect and distribute information on English language material available for computer processing and on linguistic research completed or in progress on the material, to compile an archive of English text corpora in machine-readable form, and to make material available to research institutions". For more information and a list of corpora distributed by ICAME, see Wang (2000) or the ICAME website (http://helmer.aksis.uib.no/icame.html).

> --- depend on both quantitative and qualitative (interpretative) analytical techniques (from Biber 1996:172).

The characteristics that, according to Biber, are shared by corpus-based studies relate to the choice of texts on which a study is based as well as to the methods employed for the analysis. Biber also suggests that corpus-based studies are essentially concerned with the analysis of language use. In what follows, my discussion will evolve around these three aspects: language use (Section 3.3), choice of texts (Section 3.4), and methods of data retrieval and analysis (3.5). Issues related to the choice and use of corpora for my studies are presented in Section 3.4.1 The discussion of the three aspects is preceded by a brief presentation of issues relating to the concept of *corpus* in general.

3.2 Corpus

> But defining a corpus is a more interesting question than one would think (Meyer 2002:xi)

Sinclair suggests that "[t]he results [of a corpus-based study] are only as good as the corpus" (1991:13). Assuming that this is the case, and I suggest my studies have illustrated that it is (for example Studies II and IV), it is disturbing to learn about Meyer's observation that "[m]ost corpus linguists conduct their analyses giving little thought as to what a corpus actually is" (2002:xi). The following sections aim to illustrate that this is not the case for the current study.

3.2.1 'Corpus' definitions

Various discussions of what constitutes a corpus can be found in the literature and there are a number of more or less detailed definitions. A wide and general one defines a corpus as 'a body of texts'.[6] Bowker and Pearson give a similar definition, stating that "[c]orpora are essentially large collections of text in electronic form" (2002:1). Sinclair has a more precise definition, suggesting a corpus is "a collection of naturally-occurring language text, chosen to characterize a state or variety of a language" (1991:171). McEnery and Wilson add the criterion that a corpus is a finite unit: "a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration" (2001:32). Biber glosses a corpus as a 'large and principled collection of natural texts' (1996:172). Other simi-

[6] Cf. *Longman Dictionary* (2000):
"**corpus** ... **1** *formal* a collection of all the writing of a particular kind or by a particular person **2** *technical* a collection of information or material to be studied".

lar definitions exist, and a summary is conveniently presented by Bernardini (2000:127).

What many of the definitions have in common is that they describe a corpus as a body of text that is *representative* of a particular language or variety of language. By a representative corpus is usually understood a sample of language which, when analysed, will yield results valid for the whole population which the sample was drawn from. As the population in this case is often a language or language variety, it is not possible to define its size and composition and, consequently, to calculate to what extent the corpus mirrors the composition of the population in every respect. Inevitably, not every possible sentence of a language can be found in a corpus. It is also possible that certain constructions may be proportionately more or less frequent in a corpus than in the language as a whole. Much criticism of the corpus-based, or empirical approach to language study, focuses on this and related issues. The criticism has been met by a number of scholars and an overview of the discussion is provided by Stubbs (1996) and McEnery and Wilson (2001). It is suggested that even if a corpus cannot contain examples of every aspect of a language, it is still possible to use a corpus as the basis for empirical explorations into language, in particular if scholars are aware of what is included in the corpus and consider their results against this background. That is the approach I have taken in my studies.

An informative discussion about representativeness in corpus design can be found in a study by Biber (1993). Related questions are dealt with by Atkins et al. (1992), Hunston (2002), and Meyer (2002) and discussed by, for example, Aston and Burnard (1998), and McEnery and Wilson (2001).

3.2.2 Annotation, markup, encoding

It has been suggested that implicit in the term *corpus* lies the idea that it is more than just a collection of texts, for example that it is computer readable and representative. It is also often the case that corpora are *annotated.* Corpus annotation has been defined as "the practice of adding **interpretative**, **linguistic** information to an electronic corpus of spoken and/or written language data ... [and] the end-product of this process" (Leech 1997:2). The term *markup* is sometimes used as an equivalent to *annotation* (for example Meyer 2002:81), while some authors choose to distinguish between the two, usually with *annotation* used in relation to linguistic information and *markup* for the notation of structure or appearance. Bowker and Pearson, for example, state that "[m]arkup can be used to determine the appearance and structure or composition of a document" (2002:83). A third term, *encoding*, can also be used in relation to the process and outcome of adding extra-textual information to corpus texts. In what follows, the three terms are used synonymously.

The kind of additional information made available in a corpus can vary between different corpora, as can the format in which this information is provided. Generally linguistic information is provided within the text in the form of codes, or *tags*, while information about the texts is given separately, either in separate documentation (for example information about the title/author of the LOB corpus texts) or in a separate part of the corpus text file (such as the headers in the BNC). The following sections give examples of the annotation of the corpora I have used for my studies.

3.2.2.1 Linguistic annotation

McEnery and Wilson suggest that annotating a corpus is making explicit (linguistic) information that is implicit in the plain text (2001:32). The most common form of linguistic annotation is part-of-speech annotation (also referred to as word-class, or POS annotation or tagging), where information is provided about the word-class of each word in the corpus, usually as codes ('tags') given in immediate proximity of each word(see below).[7] This means that it is possible to make searches in the corpus not only for a particular word or tag but also to distinguish automatically, for example, instances of *will* used as a noun from those of *will* used as a main verb or modal auxiliary.

Another form of linguistic annotation is prosodic annotation. The LLC has been prosodically annotated, and this annotation is given in the corpus. As this information has not been used in my studies, I will not deal with it in any detail. An example from the LLC is given in Section 3.2.2.2 below (see the LLC manual for further information).

Part-of-speech information can be given in various formats. Two examples are provided in (1) and (2).

(1) <w PNP>it <w VBZ>is <w AJ0>true <w CJT>that <w PNP>he
<w VBD>was ... (BNC ABU:1683)

(2) it_PP3 is_BEZ true_JJ that_CS he_PP3A was_BEDZ ... (LOB G 28:95)

The string of words is the same in both examples (*it is true that he was*), but the annotation is different. Perhaps the most immediate difference is the format of the tags in which the annotation is given. In the example from the BNC (1), the linguistic information is given as three-letter codes preceded by 'w' found within brackets (<w>). In the LOB example (2), the tags follow

[7] The notion of word-class in this context usually denotes a more fine-grained distinction than our general understanding of a word-class, identifying a higher number of different classes. Verbs are distinguished from nouns, but it is also common to identify present tense forms of a verb from past tense forms and singular nouns from those in the plural etc.

the words introduced by an underscore (_). The LOB tags consist of two or more letters.

The combination of letters in the tags also differs between the two examples. The word *is,* for example, is tagged as 'VBZ' in the BNC and as 'BEZ' in LOB. Both tags denote a third person present tense form of *BE*, so in this case the information contained in the tags is the same. This is, however, not always the case. In both examples, *it* is given a tag denoting a third person singular pronoun ('PNP' in BNC, 'PP3' in LOB). In the BNC (1), the same tag is also used for *he* and *she*, while *he/she* is tagged 'PP3A' in LOB. The difference between *it* on the one hand and *he/she* on the other is not distinguished in the POS-tags used for the BNC, so the level of delicacy is lower in the BNC in this respect. There are, however, cases where the situation is the opposite. The word *that,* for example, is tagged as 'CS' (subordinating conjunction) in the LOB and as 'CJT' in the BNC, standing for "the subordinating conjunction 'that', when introducing a relative clause" (Leech 1995).[8] In this case, BNC contains the more specific information.

These examples show how the information contained in the POS tags varies between the corpora, depending on what annotation scheme (set of tags) has been used. A natural consequence of this variation is that the information that can be retrieved through the tags may vary between different corpora. This can constitute a problem when two or more corpora are compared.

POS-tagging is not only a matter of using different tag-sets. The process of analysing the texts automatically and assigning the tags also varies between different programs and between different corpora, as does the success rate of the tagging. Automatic POS annotation is usually said to have an accuracy rate of between 96 and 98%. That means that between one word out of 25 (96%) and one word out of 50 (98%) is erroneously tagged. It is, however, not the case that the errors are evenly distributed across the words and texts so that any example/sample extracted from a corpus has the same proportion of errors. Certain words or constructions may turn out to be more difficult to recognise than others, and thereby more prone to being provided with an incorrect tag. Certain features or parts of a text may also be more difficult to analyse, resulting in a higher proportion of tagging errors in that text. Problems can be due to, for example, syntax or vocabulary, typographical mistakes, other errors or language irregularities.

The type of tagging errors that can be found in a text varies. A search in a tagged corpus for a certain word with a particular tag (for example *will* tagged as a modal auxiliary) can either over-generate (retrieve instances that are not modal auxiliaries), or under-generate (fail to retrieve all instances

[8] In the BNC users' documentation, Leech (1995) writes: "CJT The subordinating conjunction *that* [N.B. *that* is tagged CJT when it introduces not only a nominal clause, but also a relative clause, as in 'the day *that* follows Christmas'. Some theories treat *that* here as a relative pronoun, whereas others treat it as a conjunction. We have adopted the latter analysis.]"

that are modal auxiliaries) depending on the success of the tagger/tagging process. There are also cases where a search both over-generates and under-generates; in the above example, the tag-based search might retrieve examples of *will* that are not modal auxiliaries at the same time as not all instances of the word that are indeed modal auxiliaries are identified.

When the success of the tagging is to be measured, it is not enough simply to count the number of correct and incorrect tags. Leech et al. (1994) discuss how it is necessary to take a number of factors into account when evaluating the accuracy of tagging, and they point out that the error rate alone does not provide enough information. Among the factors mentioned is, for example, 'consistency', measuring how far the annotation scheme has been consistently applied. The authors conclude that "[e]rror rates are useful interim indications of success, but they have to be corroborated by checking, if only impressionistically, in terms of qualitative criteria" (1994:626). An evaluation of the tagging accuracy of *going to,* as found in the spoken part of the BNC, can be found below. A more extensive presentation of corpus annotation and issues related to the theory and practice of word-class tagging is found in Garside et al. (1997). Atwell et al. (2000) also provide an interesting description and discussion of differences in annotation schemes.

The corpora I have used in my studies have been tagged by different programs using different annotation schemes, if they are tagged at all. To avoid problems caused by this variation, I have opted to use the tagging solely for the studies that involve only one corpus (Studies II, III, V performed on the BNC). For studies where I compare two or more corpora (Studies I and IV), I have chosen not to use the annotation when retrieving the expressions but collected and analysed the data manually (see Study IV for discussion).

A case study on tagging

To investigate the usefulness of the POS annotation for my studies, I examined the tagging of *going to* in the spoken part of the BNC. A summary of the case study is provided here as an illustration of the relevance of tagging accuracy to the present study.

The BNC was automatically annotated by the CLAWS4 system, which has an error-rate of approximately 1.5% (3.3% if ambiguous tags are included) (Leech et al. 1994). The tagging of the spoken part was performed by the same system, with some minor changes to account for certain features specific to spoken language (see Garside 1995 for details).[9]

In order to evaluate the usefulness of the POS-annotation for my studies, I examined the tagging of 'to' in the expression *going to* in the spoken part of the BNC. *Going to* occurs 11,441 times in the spoken part of the BNC: 9,441

[9] The new version of the corpus *BNCWorld* was tagged with a later version of CLAWS which resulted in an even lower error rate. As my initial studies were performed on the first release of the corpus, however, it is the original data that are presented here.

times *to* is tagged as an infinitive marker (‘TO0’), 1,787 times as preposition (‘PRP’), and 7 times as adverb particle (‘AVP’). 135 instances are tagged with the portmanteau tag ‘AVP-PRP’.[10] A number of samples of the different tags were selected and manually checked. The results are displayed in Table 3.1.

Table 3.1. *Tagging accuracy evaluation. Samples of* going to *in the spoken part of the BNC*

Tag	Instances in BNCspo	Size of sample analysed	Proportion infinitival *to* in the sample	Estimated error in corpus (TO0)
TO0	9,441	5%	>99%	- 60 TO0
PRP	1,787	10%	17%	+ 304 TO0
AVP-PRP	135	100%	56%	+ 75 TO0
AVP	7	100%	43%	+ 3 TO0
Sum total in corpus	11,370 11,441	7% of 11,441	85% of 11,441	+ 322 TO0

Estimated total number of *going to*=TO0 in BNCspo: 9,441+322=9,763
Estimated error: (correct tags - given tags)/given tags= 322/9441= 3.4% (under-generation)

Within the investigated sample of *going to* where *to* is tagged as infinitival marker (‘TO0’), only three of the 472 instances examined were found to be erroneous (error rate 0.6%). The error rate was higher in the sample of *going to* tagged as preposition (‘PRP’). About 17% of the instances should have been tagged as infinitival markers (‘TO0’).[11] The erroneous examples are often instances where the word following *to* is ambiguous in the lexicon, and as such can be found both as, for example, a noun and a verb, e.g.:

(3) I'm not **going to** *second* guess her decision, we will work with her and we will work with her government. (BNC KRU 092)

(4) So one tries to build up a kind of agenda of all the things that different people involved think might be important before one tries to produce a plan as to how one's **going to** *work*, and even then there may be a chance for you actually to discuss the plan with various people as well. (BNC KRH 3619)

[10] The figures for the tagged occurrences listed here were obtained by searching the corpus with the BNCweb program. It is generally known that there are deficiencies in the index provided with the first release of the corpus, which could be one explanation as to why the total does not add up to 100% here. If any instances of *to* in the corpus remain untagged, they would also be missing from these figures.

[11] The study was repeated on two more samples – another random sample constituting 10% of all instances and one sample consisting of one example from each text, 473 examples in all. The 10% sample displayed an error rate of about 18% (plus six indefinable cases) while the one-per-text sample had an error rate as high as 26% (plus nine uncertain cases).

In addition to the cases that should have been tagged as infinitival markers, a small number of unclear cases were found, where it was impossible to say how the word should have been tagged, even after considering a large context. These unclear instances have not been included in the error rate.

Among the 135 instances tagged with the portmanteau tag ‘AVP-PRP’, 75 instances (56%) seem to be infinitival markers while a further dozen instances could not be classified, such as:

(5) It's not much use if she's not **going to** <unclear> with <laugh>
(BNC KBH 4907)

Three of the seven instances tagged as adverb particles (‘AVP’) should have had the ‘TO0’ tag, while the remaining four were impossible to classify.

A search in the corpus for instances of *going to* where *to* is tagged as an infinitival marker thus under-generates and produces too few instances (low recall). The precision is good, however; the number of instances that are not infinitival markers is low (3/472=0.6%). A search for *going to* with *to* tagged as an infinitival marker thus retrieves almost only correctly tagged instances, but some instances where *to* is an infinitival marker are not found. The overall estimated error for the set of *going to*=TO0 is 3.4%; the actual number of instances is 3.4% higher than indicated by the search result. I have considered this error-rate low enough to motivate relying on the tagging in case studies concerning *going to* in the spoken part of the BNC (Studies II, III, V), in particular when considering that if the tagging information is not used, about 15% of the retrieved instances are the unwanted variant where *to* is not used as an infinitival marker.

3.2.2.2 Textual and extra-textual information

Information about a text, such as the author/speaker, source, text category, recoding date etc. can be encoded in the corpus. The amount and format of such textual and extra-textual information (or *meta data*) varies considerably between the corpora I have used. The Brown corpus and its followers are distributed by ICAME in a format where each line of text is introduced by a code denoting the text and line number, as illustrated in an example from the Brown corpus (6).

(6) A01 0010 The Fulton County Grand Jury said Friday an investigation
A01 0020 of Atlanta's recent primary election produced "no evidence" that
A01 0030 any irregularities took place. The jury further said in term-end
A01 0040 presentments that the City Executive Committee, which had over-all
A01 0050 charge of the election, "deserves the praise and thanks of the

A01 0060 City of Atlanta" for the manner in which the election was
conducted. (Brown A01:10-60)

As seen in the example, the text and line reference is not distinguished from the text itself in any other way than by its position on the line. General purpose software, such as WordSmith Tools, will treat the references as part of the running text, and fail to identify in this extract, for example, the phrase ‘the City of Atlanta’ (lines 0050-0060), reading it as ‘the A01 0060 City of Atlanta’. The LLC is distributed in a similar format, where each line is introduced by text and line references as well as speaker identity codes. In addition to that, the LLC text contains prosodic annotation given inside the words, as illustrated in (7).

```
(7)  1 1   1  160 1 2(B   13  is . *.* ^that your . there`s ^something that your /
     1 1   1  160 1 1(B   13  :own candidate can :h\/andle# - -               /
     1 1   1  180 2 1 B   21  ((I ^won`t))                          /
     1 1   2  190 1 1 A   11  *((^y\eah#))*                       (LLC 01 160-190)
```

The prosodic annotation is valuable for certain kinds of studies, but makes the text difficult to analyse with general-purpose corpus tools.

The BNC and Sampler corpora are distributed in SGML-format where the information (linguistic and non-linguistic) that is not part of the corpus text as such is given within brackets (< >). This information is found either in direct connection to the words in the text (for example marking poems found in prose texts, headlines, unclear passages, pauses in speech) or in the header of the text (a section at the beginning of each text where metadata for the text is recorded), for example the printed source, recording date, file revision history, target audience, etc. Example (8) illustrates a section of a written BNC text which includes linguistic information (POS tags) as well as textual information (“<head type=MAIN>” indicates heading, “<s n="966">” marks the beginning of a sentence-like element, number 966 in the particular text).

(8) **<head type=MAIN><s n="966">**<w VVN>Styled <w TO0>to
<w VVI>win</head> (BNC A0V 966)

Example (9) shows a spoken passage, with markup showing POS information as well as sentence numbers (<s n=”8645”>, <s n=”8646”>), speaker identity (<u who=PS04U>), description of vocal event (<vocal desc= cough>) and non-verbal content (<event desc="tape cut out">, <pause dur=11>).

(9) **<s n="8645">**<w DT0>Such <w AT0>a <w NN1-AJ0>perishing <w
VVD>bore **<vocal desc=cough>** <w ITJ>Oh <w VVB>bless <w
PNP>you <w VVB>thank <w PNP>you <w AV0>very <w AV0>

> much<c PUN>.</u>**<u who=PS04U><s n="8646">**<w VVD>Knew
> <w PNP>I <w VHD>had <w AT0>a **<event desc="tape cut out">**
> **<pause dur=11>** <w PNP>I <w VVB>hope <w TO0>to <w VVI>get
> <w PNP>it<c PUN>.</u > (BNC KBF 8645-8646)

Certain tools, such as WordSmith Tools, recognise SGML-tags and can be set to include or disregard them when retrieving concordances, which means that it is possible to search the text automatically without the markup being confused with the text. Special SGML-aware software can be used to derive information from the markup and include it in a search thereby making it possible, for example, to search for modal instances of *will* occurring in a headline or spoken by a woman. Examples of such software are SARA and BNCweb (further presented in Section 3.5.2).

3.2.2.3 Future developments

The variation between corpora with regard to the amount and format of extra-textual information means that it can be difficult to compare different corpora (see, for example, Study IV). This has been recognized by the corpus linguistic community and efforts are made to develop standards to recommend to corpus compilers. The Expert Advisory Group on Language Engineering Standards (EAGLES) have developed a set of guidelines to 'serve as a widely accepted set of encoding standards for corpus-based work'. This Corpus Encoding Standard (CES) "… specifies a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation (marking of structural and typographic information) as well as general architecture (so as to be maximally suited for use in a text database). It also provides encoding specifications for linguistic annotation, together with a data architecture for linguistic corpora." (EAGLES). The Text Encoding Initiative (TEI) also include recommendations for corpus annotation in their Guidelines (Sperberg-McQueen and Burnard 2002). However useful such standards are for new corpus compilation projects, it is nevertheless the case that many of the corpora that are used now do not adhere to them. Some of the problems related to using corpora in different formats and with different kinds and formats of annotation have been discussed in Study IV.

3.2.3 Specific issues related to the use of spoken data

A considerable part of my research on the expressions of future has been based on spoken corpus data. Working with spoken corpora is in many ways different from using written data. One feature which is immediately evident to anyone setting out to use spoken corpora is that the amount of spoken data available in corpus format is considerably smaller than that of written language. An important reason for this is that creating computer-readable corpora of spoken data a complex, time-consuming and therefore expensive task. Discussing the process of compiling spoken corpora in any detail is beyond the scope of the present study. However, a basic understanding of how spoken corpora are compiled is necessary for the interpretation of the results of the analyses presented below. Consequently, some space will be devoted to issues that I find of particular relevance to my studies (for more detailed discussions relating to the creation of spoken corpora, see, for example, Burnard 1995, Crowdy 1993, 1994, and Leech et al. 1995).

3.2.3.1 Capturing spoken data

Capturing spoken data can be difficult. Existing recordings, such as radio and television broadcasts and 'spoken books', are probably the most easily available sources, while samples of 'spontaneous private conversation' are considerably more difficult to obtain. It is not only a matter of obtaining permission to record someone (as surreptitious recording is not only unethical but also illegal in some countries) but a question of asking to what extent a sample of language spoken by people who are aware of the fact that they are being recorded differs from every-day spontaneous discourse. Research has shown that many people tend to adapt their language (lexis, pronunciation, topic) to follow the perceived standard more closely when they are aware that they are being recorded (see, for example, Renouf 1986 for a discussion on eliciting spoken data).

The problem of obtaining naturally occurring speech has in several cases been solved by recording people surreptitiously (or semi-surreptitiously, with only one or a few of the participants aware of the recording) and then asking for permission to use the recording (and erasing the material in cases where permission was not granted). That is the case for parts of the LLC and BNC material used in the present study. It is, of course, difficult to assess to what extent the corpus material differs from un-recorded conversations. The BNC *Users Reference Guide* states:

> All conversations were recorded as unobtrusively as possible, so that the material gathered approximated closely to natural, spontaneous speech. In many cases the only person aware that the conversation was being taped was the person carrying the recorder. Although an initial unnaturalness on the part of the recruit was not uncommon this soon seemed to disappear. (Burnard 1995)

The BNC texts are, to some extent, coded for the spontaneity of the speakers, defined as 'high', 'medium', or 'low'. The code refers to the conversation as a whole and it is not possible to distinguish the level of spontaneity of individual speakers or find out when the "initial unnaturalness" referred to by Burnard disappears (Burnard 1995, see above). This, in combination with the fact that most of the classified texts are given the same code (over 85% of them are marked 'high' for degree of spontaneity), means that little information can be obtained from the BNC about how the use of the expressions of future or any other linguistic feature varies with the naturalness with which an utterance was delivered (see also Study II:43-44). Any statement about the nature of the spoken texts should be interpreted with this in mind.

3.2.3.2 Transcription

One great difference between written and spoken corpus data is that while written data exist in a readable format, spoken data have to be converted into written form.[12] The spoken corpus data used in the present study are orthographically transcribed versions of spoken language (the LLC is extensively annotated with prosodic information, but that annotation is not used in the present study and therefore not treated further here). The language studied is thus processed to some extent before it is ready for research purposes (converted from spoken to written format). This processing involves human actors (the transcribers), and the transcription is based on the transcribers' interpretation of what is said in the recording.[13] Factors outside the actual pattern of the sound-waves can thus be taken into consideration when the transcription is made. There are obvious benefits with this. Semantically unrelated homonyms, such as *for/four* and *there/their* are, for example, given different transcriptions. Moreover, in cases where the quality of the recording is poor, human transcribers can interpret the context and thereby distinguish between similar words/phrases. It is, naturally, also the case that human transcribers can make erroneous decisions when guided by their linguistic intuition. It is, however, likely that other errors are more frequent. It is obvious, for example, that the transcribers at times make mistakes that are typographical errors rather than being caused by problems in interpreting the recording. An example is given below (10), where it seems likely that the word spoken is *your* and not, as written, *you're*.

(10) And *you're* dad said, Oh no no. (BNC KSS 174)

Worth noticing is perhaps that the same kind of problem is also found in the written texts. In the following example (11) it seems as if *you're* would be a more correct variant than, as given, *your*.

(11) "*Your* not going to encourage her to marry that old goat?" (BNC B1X 2302)

[12] It is, of course, possible to study spoken language in spoken form, for example by listening to a recording. Although work on producing tools for automatic analysis of spoken data is proceeding with some success, such tools suitable for corpus analysis are not yet generally available.
[13] Development in this field is making progress, but it is not yet possible to obtain automatically transcribed data which can match the standards of human transcribers.

It could be possible, of course, that such irregularities are intentional, but considering the number of such occurrences and the contexts where they occur it is more likely that they are unintentional typographical errors. To calculate the frequency of such errors/intentional irregularities, it would be necessary to inspect all examples of either variant; for *you're* and *your* in the BNC, for example, this would entail examining over 175,000 instances in all. It is obvious that manual inspection would be extremely time-consuming and not a task that could fall within the scope studies such as the present unless the correct transcription were crucial to the result of the study.[14]

When a spoken text is transcribed (especially when the transcription is orthographic), certain features of the language are no longer possible to discern. Unless explicitly marked, intonation boundaries and variation in pronunciation patterns are among those features. The omission of such features, of course, means that a great source of information is lost. It is not only impossible to examine features of pronunciation, but it is also more difficult to study meaning (for instance, is the speaker ironic?), and certain discourse-related issues. One benefit of not marking prosodic information is that the text is easier to search for lexical items if all instances of a word are written in the same way. Some problems of using prosodically annotated corpora with software not specifically adapted for the particular corpus or annotation are illustrated in Study IV.

Even if a text is transcribed orthographically, it may be desirable to record certain typically spoken features in the transcription. If features such as humming, pauses, pause fillers, and overlapping speech are to be transcribed, it is necessary to construct ways to mark these in the transcription. Some examples of how this was done in the BNC are given in (9) above. It is furthermore necessary to make decisions whether to use punctuation and, if so, how to use it. Commas, question marks, exclamation marks, and so forth are not spoken as such, so if these are to be used in the transcription they have to be inserted according to someone's (such as the transcriber's) interpretation of what was said and how. Needless to say, it is very difficult to ensure consistency when the basis of the transcription depends on the interpretation of one, or several, person(s). This also applies to the transcription of items where pronunciation differences exist and can also be recorded in writing. Examples of such items are contractions, such as *do not/don't,* and *I am/I'm*, and deletions, such as *comin'/coming* and *'aving/having,* as well as instances of vocal activity that does not have one uniform written representation, for example 'gap-fillers' such as *um/erm,* to mention only a few.

[14] In a case where the correct rendering of a frequent item is vital, it is possible to get a estimate of the reliability and an indication of potentially problematic contexts by examining a subset of the instances. See, for example, the case-study on tagging accuracy in Section 3.2.2.1.

Even if there are only two written variants, the pronunciation is not likewise dual but probably more suitably described as a continuum. There may be cases where, for example, *going to* and *I will* are clearly pronounced in a way that corresponds to the written representations, but this is not always the case. Stubbs (1980:118), for example, suggests that when *going to* is followed by a verb it is unlikely to be pronounced as two words except in very formal speech, and he gives examples of pronunciation variants. Poplack and Tagliamonte (2000:328) state that "*[g]oing to* actually subsumes a number of phonetically distinct forms, variously realized as *goin(g)ta, gonna, gon, go*", and they study these forms separately.

In the corpora used for the present thesis, however, only two variants are defined: *going to* and *gonna.* In the BNC and Sampler corpora, both variants occur frequently. In the spoken LLC, however, there is but a handful instances of *gonna.* Whether this mirrors differences in the language recorded or differences in transcription practices is impossible to say. For that reason, no attempt has been made to compare the use of *gonna* versus *going to* between the BNC/Sampler and LLC corpora. The same applies to *will* and *'ll*, which can also be regarded as pronunciation variants (see Chapter 1).

Aston and Burnard (1998) remark in connection to the transcription of the spoken material in the BNC, that "[i]t should not be assumed, for example, that one transcriber's 'erm' or 'going to' is necessarily phonetically distinct from another's 'um' or 'gonna'" (1998:38). For the present study, it has nevertheless been considered of interest to examine the variants *gonna* and *going to* separately (in certain cases, for example, when data from different corpora are compared, the figures for the two variants have been added together to enable comparison). The same refers to the *will* and *'ll* expressions.[15]

There are several reasons for accepting the transcription in these cases. One is that even if the transcribed instances are not at all times clearly phonetically distinct, there must be something in the pronunciation that makes the transcriber choose one of the two variants. In cases where the transcribers have access to a limited number of transcriptional variants (such as *gonna* and *going to, will* and *'ll*), they have to make an active choice between the variants every time they come across instances of the constructions. The transcription can in this case be seen as a mirror of how the utterance is perceived, even if it is not beyond all doubt that the phonetic difference could at times be smaller between, for example, instances of *gonna* and *going to* than between every instance of *going to*. Krug (2000:38) argues that in cases such as that of *gonna/going to* in the BNC, the sheer amount of data available makes the use of the transcriptions more reliable, as differences will be evened out.

[15] There is a certain amount of disagreement on whether the *'ll* form is to be seen as a contracted/reduced variant of *will* or of *shall*. That issue is dealt with in Chapter 1.3 above.

Crowdy describes how the transcribers working with the BNC were trained and their transcriptions monitored for a considerable period. In addition, every fifth transcript was checked against the original recordings "to ensure consistency" (1995:228). Aston and Burnard (1998:37) note that "[e]very effort was made to achieve accuracy and consistency in the transcription, but unquestionably some errors remain". Even if it cannot be assumed that the transcriptions are always correct and in every detail consistent, that is not reason enough to doubt that the transcriptions on the whole are reasonable, or even good, representations of what was actually said during the recording and that, interpreted with an appropriate amount of caution, they can be useful for the study of spoken language. For further and more thorough discussions of issues relating to encoding, transcription, and use of spoken corpora, see, for example, Leech et al. (1995), Edwards (1995), Chafe, (1995) and Johansson (1995).

3.2.4 Summary: corpus

This section has dealt with the corpus, presenting various definitions used by different scholars. Issues relating to the creation and composition of a corpus have been discussed, as has the markup, or annotation, of corpora. Potential problems relating to the creation and, consequently, use of spoken corpora have also been presented. The aim has been to provide a background to the approach taken in my studies. In what follows, the discussion moves on to deal with the characteristics Biber (1996) lists as shared by corpus-based studies: study of language use (3.3), choice of texts (3.4) and methods of data retrieval and analysis (3.5).

3.3 Language use

The value of studies of language use is widely recognised. McMahon (1994:179), for example, states that "[u]ltimately, we suggest that successful models of language must be models of language *use*". Biber (1996:171) argues that "analyses of language use provide an important complementary perspective to traditional linguistic descriptions" and that corpora constitute a good resources for investigation of language use. Bowker and Pearson (2002:1) observe that "[c]orpora are becoming a very popular resource for people who want to learn more about language use".

The term 'empirical' is commonly used in connection with corpus-based studies, for example by Biber (1996, see also Section 3.1) where the phrase "analyzing the actual patterns of use in natural texts" further explains the term. *The Longman Dictionary of Contemporary English* (2000) defines the term 'empirical' as "based on scientific testing or practical experience, not on ideas from books". A definition more directly related to linguistics is that for ‘empiricism’ given in McEnery and Wilson (2001:198): “an approach to a subject (in our case linguistics) which is based upon the analysis of external data (such as text and corpora)”. McEnery and Wilson maintain that “[e]mpirical data enable the linguist to make statements which are objective and based on language as it really is rather than statements which are subjective and based upon the individual’s own internalised cognitive perception of the language”, and conclude that “corpus linguistics necessarily entails an empirical approach, but empirical linguistics need not entail the use of a corpus” (2001:103).

The results of my investigations show how the expressions of future are distributed across the corpora. I do not intend to suggest that there are no other ways to express similar meanings and neither do I discuss the variation in meaning between the expressions in different contexts. In my studies, I describe how the expressions of future are used, and I suggest that this is information that can be of use and interest in further studies relating to the expressions of future. To be fully useful, though, the results of a study should be given in a context where information is provided about the nature of the corpora examined. The following section provides that information.

3.4 Choice of texts

> The results are only as good as the corpus. (Sinclair 1991:13).

For my studies, I have chosen to use existing corpora rather than to go through the process of compiling corpora myself. It is a decision that is easy to understand when considering that the creation of the British National Corpus, for example, took a large team, with substantial funding and considerable professional experience of related matters, several years to complete. In addition to the time and effort saved by not compiling one’s own corpus, there are a number of other benefits in using existing corpora. One important benefit of using generally available corpora is that this “enable[s] the possibility of cumulative results and public accountability” (Biber 1996:173). Other researchers can use the same corpus either to repeat and critically examine another study, or to add further observations about the language based on the same source. McEnery and Wilson (2001:32) suggest that “[a]lthough it is not an essential part of the definition of a corpus, there is also often a tacit understanding that a corpus constitutes a standard reference for the language variety which it represents”. A standard reference corpus is widely available to other researchers and can provide “...a yard-stick by which successive studies may be measured” (ibid.). A study performed on a standard reference corpus can be repeated and the results verified (public accountability). By using the same corpus for different kinds of studies it is also possible to get a more extensive coverage of the linguistic features in the corpus and, consequently, in the language from which the corpus is drawn.

By comparing the results of one study with those of another based on the same data, conclusions can be drawn not only about the linguistic properties of the texts but also about the methods and tools used. This is to some extent done in the present thesis, where well-known corpora are used and the results are compared both to those obtained by others and to those I have obtained by using different methods and tools.

Sinclair recommends that corpus designers carefully document the corpus design criteria:

> Until we know a lot more about the effects of our design strategies, we must rely on publishing a list of exactly what is in a corpus ... Users and critics can then consider the constitution and balance of the corpus as a separate matter from the reporting of the linguistic evidence of the corpus. (Sinclair 1991:13)

I interpret this recommendation to corpus compilers as one directed to corpus users as well. Issues related to the composition of the corpus should be considered and discussed also by the user of a corpus. Only then is it possible to evaluate the result of a study based on a particular corpus.

3.4.1 Corpora used

The corpora chosen for my studies contain Present-day English texts and are available in electronic format. I have defined Present-day English as roughly equivalent to the language used since the 1960s, a date which coincides with the date of publication of the texts in the first generally available electronic corpus, the Brown corpus (1961). A further point of concern has been that the corpora should be suitable for comparison (issues related to corpus comparability are discussed further in Section 3.4.2). As I wished to examine a number of potential association patterns (associations with region, speaker properties, medium, etc.), I needed corpora where the different extra-linguistic factors could be identified. I also had to use corpora that were available to me at a low cost at the same time as I considered it important to use generally available, standard reference corpora to make my results reproducible and verifiable. For the other aim of my thesis, to discuss issues related to corpus-based studies, I was interested in using well-known and much used corpora.

On the basis of the above starting points, I chose to use the following corpora: Brown, LOB, FLOB, Kolhapur, Frown, LLC, BNC, and BNC Sampler. A schematic overview of some of the features of the corpora is given in Table 3.2. The corpora are presented in some more detail below. Issues relating to the choice of the corpora are discussed further in Section 3.4.2, together with a discussion about the comparability of the corpora.

Table 3.2. *Schematic over-view of corpus features*

	Brown	**LOB**	**FLOB**	**Kolhapur**	**Frown**	**LLC**	**BNC**	**Sampler**
Corpus size (million words)	1	1	1	1	1	0.5	100	2
Regional variety	American	British	British	Indian	American	British	British	British
Medium (written spoken)	written	written	written	written	written	spoken	written + spoken	written + spoken
Text size	2,000	2,000	2,000	2,000	2,000	5,000	varied	varied
Date of texts	1961	1961	1991	1978	1992	1953-1988	1960-1993	1975-1993
Annotation		POS				POS, prosodic	POS	POS
Metadata information about	text category	text category	text category	text category	text category	text category, speaker id.	text domain, author/ speaker	text domain, author/ speaker
Text categories	15 genres	15 genres	15 genres	15 genres	15 genres	'Adult educated English'	various	various

3.4.1.1 The Standard Corpus of Present-Day Edited American English (Brown)

Modern corpus linguistics can be said to have started with the creation of *The Standard Corpus of Present-Day Edited American English*, more commonly known as 'the Brown corpus'. The Brown corpus was the first computer-readable corpus and it was compiled from material published in 1961. An important feature of the corpus is that it is balanced; it is the result of an effort to create a corpus which represents a particular variety of language. The 500 texts in the corpus are all of the same length[16] and they are drawn

[16] Francis and Kucera describe how the sample beginning and end were set:
"Each sample begins at the beginning of a sentence but not necessarily of a paragraph or other larger division, and each ends at the first sentenceending [sic] after 2000 words. In a few cases the count erroneously extended over this limit, but the extra material has been allowed to remain. Owing to errors in the rough count, 15 samples have between 1,990 and 1,999 words, and 3 have fewer than 1,990. The average length is 2,028.6" (Francis and Kucera 1979).

from 15 different categories of texts, henceforth also referred to as 'genres', following Biber (e.g. 1988).[17] The size of the genres varies between six and 80 texts of 2,000 words each (see Table 3.3). For my studies, I have at times chosen to deal with two sub-corpora formed by the informative texts (genres A-J) on the one hand and the imaginative texts (genres K-R) on the other. I refer to these sub-corpora as 'hyper-categories' (see also Chapter 5). Further information about the corpus content and composition can be found in the corpus manual (Francis and Kucera 1979).

The Brown corpus has been annotated with part-of-speech information and exists in several tagged versions. The version used for this study, however, is the un-tagged version distributed by the International Computer Archive of Modern and Medieval English (ICAME).

The importance of the Brown corpus is emphasised by the fact that the structure and composition of the corpus has since been replicated in the compilation of new comparable corpora. Some of these comparable corpora are used in Studies I and IV as well as in the present Summary and they are discussed further below. A concise overview of the composition of the Brown corpus and its clones is offered in Table 3.3 below.

3.4.1.2 The Lancaster-Oslo/Bergen Corpus of British English (LOB)

The effort to create an American English corpus was soon followed by the creation of a British English one, built according to similar principles, intended for comparison between American and British English. The LOB corpus contains material published in Britain in 1961. The composition follows that of Brown closely with a few exceptions (see Table 3.3). The corpus is available with or without part-of-speech annotation, and both versions have been used for this study (Studies I and IV). More detailed information about the LOB corpus can be found in the manual (Johansson et al. 1978).

3.4.1.3 The Kolhapur Corpus of Indian English (Kolhapur)

The Brown corpus has also been used as a model by the compilers of a corpus of Indian English, the Kolhapur corpus. The choice of text categories follows that of Brown, copied for LOB, but the number of texts in each of the 15 genres differs somewhat from that of the Brown corpus. The General fiction genre (K) is twice as big as in the comparable American and British corpora, while the number of texts in some of the other fiction categories (M, N, P) is considerably lower. The Kolhapur corpus is distributed by ICAME in an untagged version (further information in Shastri et al. 1986, and Shastri 1988).

[17] The term 'genre' is not used in the corpus documentation for the 15 text categories. Following Biber (1988), the term has frequently been used to denote the text categories in Brown and its clones, and is also used in this thesis. For ease of reference, the genres are at times referred to by the letter of the alphabet used as a code for each text category, as illustrated in Table 3.3.

Table 3.3. *Composition of the Brown corpus and its clones. Number of texts (a 2,000 words) per genre*

Hyper-category	Genre	Brown	LOB	Kolhapur	FLOB	Frown
Informative	A Press: reportage	44	44	44	44	44
	B Press: editorial	27	27	27	27	27
	C Press: reviews	17	17	17	17	17
	D Religion	17	17	17	17	17
	E Skills, trades and hobbies	36	38	38	38	36
	F Popular lore	48	44	44	44	48
	G Belle lettres, biography, memoirs, etc.	75	77	70	77	75
	H Miscellaneous (Govt. documents, foundation reports, industry reports, college catalogue, industry house organ)	30	30	37	30	30
	J Learned	80	80	80	80	80
Sub total		*374*	*374*	*374*	*374*	*374*
Imaginative	K General fiction	29	29	58	29	29
	L Mystery and detective fiction	24	24	24	24	24
	M Science fiction	6	6	2	6	6
	N Adventure and Western fiction	29	29	15	29	29
	P Romance and love story	29	29	18	29	29
	R Humour	9	9	9	9	9
Sub total		*126*	*126*	*126*	*126*	*126*
TOTAL		500	500	500	500	500

3.4.1.4 The Freiburg - LOB Corpus of British English (FLOB)

The Freiburg - LOB corpus was created as a mirror of the LOB corpus but containing texts from 1991. The size and sampling frame is carefully constructed to make the LOB and FLOB corpora suitable for comparison to enable the study of the development of British English in the 30-year span from 1961 to 1991. The untagged version of the corpus distributed by ICAME in 1999 has been used for Study IV. For further information about the corpus, see Hundt et al. (1998).

3.4.1.5 The Freiburg - Brown Corpus of American English (Frown)

Like FLOB, the Freiburg - Brown corpus (Frown) was created to mirror an earlier corpus, Brown. The corpus contains American English texts from 1992, carefully matched on the Brown corpus. The corpus is distributed in an un-tagged version by ICAME, and that version has been used for parts of the present Summary. For further information about the corpus and its content, see Hundt et al. (1999).

3.4.1.6 The London-Lund Corpus of Spoken English (LLC)

The London-Lund Corpus is a corpus of spoken British English created in collaboration between The Survey of English Usage (London) and Lund University. The corpus contains transcribed versions of spoken language recorded in England between 1953 and 1987. The 100 texts in the LLC all contain approximately 5,000 words each.[18] They are divided into 12 categories, as illustrated in Table 3.4.

Table 3.4. *Composition of the LLC: text categories (information from Greenbaum and Svartvik 1990).*

Text category	Content
S1 (14 texts)	Conversations between equals
S2 (14 texts)	Conversations between equals
S3 (7 texts)	Conversations between disparates
S4 (7 texts)	Conversations or discussions between equals
S5 (13 texts)	Radio discussions and conversations between equals.
S6 (9 texts)	Interviews and conversations between disparates
S7 (3 texts)	Telephone conversations between equals
S8 (4 texts)	Telephone conversations between equals
S9 (5 texts)	Telephone conversations between disparates
S10 (11 texts)	Spontaneous commentary, mainly radio
S11 (6 texts)	Spontaneous oration
S12 (7 texts)	Prepared oration

The corpus is annotated with prosodic information. The annotated version distributed by ICAME has been used for Studies I and IV. A stripped version was also used for parts of Study IV. More information about the LLC is provided in the corpus manual (Greenbaum and Svartvik).

3.4.1.7 The British National Corpus (BNC)

When it was created, the Brown corpus was considered big (1 million words). It was soon obvious, however, that the corpus was too small for certain research areas (for example in the field of lexicography). The British National Corpus (BNC) was created to answer the needs of a large, balanced corpus of contemporary British English containing both written and spoken data. Considerable time and effort was spent on designing the corpus, and defining the size and composition of the various text categories. The corpus contains about 90 million words of written data, and about 10 million words of orthographically transcribed spoken data (details in Table 3.5).

The BNC texts vary in length and were sampled according to certain, predefined criteria, called selection features. The selection features for the writ-

[18] In the LLC manual (Greenbaum and Svartvik) it is mentioned that certain parts of the conversations in the corpus are not included in the word-count. The size of the corpus is, thus, over 500,000 words.

ten texts were time (within certain dates), domain (subject field) and medium (book, periodical, unpublished etc.). In addition to the selection features, some secondary criteria, or descriptive features, could be considered where a free choice of texts was available. Among these secondary criteria can be mentioned author properties, such as age, sex, and domicile, and target group characteristics (age, sex, etc.). No targets were pre-defined for the descriptive criteria, but certain efforts were made to obtain material of different kinds.

Table 3.5. *Composition of the BNC (information from Aston and Burnard 1998).*

Written 90%			**Spoken 10%**	
Sampling criteria (proportion of words in the written part)			**Context-governed (CG)**	**Demographically Sampled (DS)**
text domains	**time**	**medium**	**Categorised by topic:**	**Sampled according to:**
Imaginative 22% Arts 8% Belief and thought 3% Commerce and finance 7% Leisure 10% Natural and pure science 4% Applied science 8% Social science 14% World affairs 17% Unclassified 7%	1960-74 app. 2% 1975-93 app. 98%	Book 59% Periodical 31% Misc. published 4% Misc. unpublished 4% To-be-spoken 1.5% Unclassified 0.4%	Educational and informative 17% Business 21% Institutional/ Public 22% Leisure 22% Unclassified 18%	Speaker age Speaker sex Speaker social class Geographic region
89,740,544 words			6,154,248 words	4,202,216 words
			10,356,464 words	

It is important to be aware of the fact that the selection criteria are independent. That means that if the sampling frame states that 20% of the data are to be imaginative and 30% are to be drawn from periodicals, it is not necessarily the case that 20% of the data from the periodicals are imaginative texts, or that 30% of the imaginative texts are published in periodicals.

The spoken texts were sampled according to two principles. About half of the material (4.2 million words) was captured by demographically selected respondents who recorded their spontaneous conversations over a number of days. This data is now referred to as the Demographically Sampled component (DS). The Context-governed component (CG) contains spoken data captured in four contextually defined domains, in a number of situations (6.2

million words). The DS component is also referred to as a “component of informal encounters”, while the CG part of the corpus is described as a “component of more formal encounters” (Aston and Burnard 1998:31).

The BNC corpus is different from most other, now available corpora in that it is generously annotated not only with part-of-speech information but also with extra-linguistic information about the texts, the authors, and the speakers. The first release of the corpus has been used for Studies II and III. Further information about the BNC and the distribution of text and speaker-related markup can be found in, for example, Burnard (1995) and Study III (spoken part only).

3.4.1.8 The BNC Sampler (Sampler)

The BNC Sampler consists of a subset of two per cent of the BNC material released separately.

Table 3.6. *Composition of the BNC Sampler (information from the BNC Sampler CD)*

Written 50%			**Spoken 50%**	
Sampling criteria (proportion of words in the written part)			**Context-governed (CG)**	**Demographically Sampled (DS)**
text domains	**time**	**medium**	**Categorised by topic**:	**Sampled according to:**
Imaginative 23% Pure science 3% Applied science 11% Community and Social science 4% World affairs 27% Commerce and finance 9% Arts 5% Belief and thought 4% Leisure 13%	1975-1993	Book 61% Periodical 27% Other written 11%	Educational 16% Business 27% Leisure 27% Institutional 29%	Speaker age Speaker sex Speaker social class Geographic region
			496,852 words	493,852 words
1,002,821 words			990,704 words	

The Sampler differs from the BNC not only in size but also to some extent in composition, as can be seen by a comparison of Tables 3.5 and 3.6. In the Sampler, the written and spoken parts are of the same size (1 million words each), to be compared to a distribution of 90% written and 10% spoken material in the BNC proper. The spoken part of the Sampler consists of equal proportions of material from the DS and CG components (in the BNC, the CG component contains just over 6 million words, while the DS component contains about 4 million words). The Sampler has been tagged with a new version of the CLAWS tag-set, and the output has been manually checked and corrected for errors. The corpus has been used for Study IV. Further information about the Sampler can be found in the manual (only available on the corpus CD) and in Study IV (spoken part).

3.4.2 Corpus comparability

As suggested in Section 3.4.1, one important factor when choosing corpora for my studies has been to enable comparison. For that purpose, I looked for corpora that share as many extra-linguistic features as possible. The corpora mirrored on the Brown corpus were largely created to be comparable to Brown in sampling strategies and composition. As Table 3.3 above illustrates, the corpora are very similar with regard to a number of other extra-linguistic features and thereby suitable for comparison. The texts in the LOB corpus were published in the same year as those in Brown, so they are similar with regard to the feature time while the regional feature differs. The two corpora offer good potential for studies comparing British and American English. The FLOB and Frown texts are also British and American respectively but were published 30 years after those in LOB and Brown (or 30 and 31 years to be precise). The four corpora are all well suited for comparison, either across regional varieties of English (American Brown and Frown vs. British LOB and FLOB), across time (Brown and LOB from 1961 vs. Frown and FLOB from 1992 and 1991). The Kolhapur contains texts published in India in 1978, compiled according to a plan very similar to that for LOB. Apart from the regional aspect and the time of publication of the texts, a major difference between the Kolhapur corpus on the one hand and the Brown and LOB corpora on the other is the composition of the Imaginative hyper-category (genres K-R; see above). The Indian corpus contains a much larger number of texts in the ‘General fiction’ category, at the expense of the more specific imaginative categories such as Romance and love story (P), Adventure and Western fiction (N) and Science fiction (M). It cannot be excluded altogether that this difference in the composition may affect the distribution of the expressions of future. On the whole, however, the size and composition of the Kolhapur corpus makes it a useful starting-point for comparisons between native varieties of English (as found in Brown, Frown, LOB and FLOB) and English produced in India.

For my studies I have also used the British National Corpus and the BNC Sampler. The written parts of the corpora contain British English from about the same time as the LOB and FLOB corpora texts (from 1960 to 1993, even if the bulk of the texts are from 1974-1993). The composition of the BNC and Sampler differs considerably from that of LOB and FLOB, as do the text size and selection criteria. Due to these differences I have chosen not to make any substantial attempt to compare the distribution of the expressions of future in the BNC and Sampler with that in the Brown corpus and its clones.

With regard to the spoken data, it has been more difficult to find comparable corpora. The most well-known spoken English corpus is the London-Lund Corpus of Spoken British English. I have chosen to compare the LLC to the spoken part of the BNC Sampler as these were the only corpora containing spoken Present-day British English available to me. Admittedly, these corpora are not ideal for comparison. As described in Study IV, they differ on a number of points relating, for example, to sampling methods and material gathered. In addition to this, the transcription method varies. The Sampler (and the BNC) texts have been orthographically transcribed while the LLC has prosodic information recorded in the transcription. This difference is easily observable but not necessarily something which makes the corpora impossible to compare. What is more serious is, perhaps, that it is not possible to say anything about the extent to which the transcribers and factors relating to them (training, guidelines, personal features) may have affected what is found in the corpora. This may be particularly obvious when the distribution of *gonna* in the corpora is compared. In the LLC there are but a handful of instances while the proportion in the Sampler is very high. Because of the differences between the corpora, I have chosen not to make any far-reaching comparisons between the distribution of the expressions of future in the LLC and Sampler corpora.

The BNC and Sampler corpora both contain two components of spoken texts, the Context-governed and the Demographically Sampled components. The two components differ in composition, relating to differences in the sampling strategies employed for the components. They nevertheless provide an interesting basis for comparison of different kinds of spoken discourse recorded in roughly the same period of time and presumably transcribed by the same team of transcribers. When examined, the distribution of the expressions of future turns out to differ between the two components, as discussed in Studies IV and V, and Chapter 5. For further discussion of issues related to the comparability of the corpora, see Study IV.

3.5 Methods of data retrieval and analysis

As shown in Section 3.1 above, Biber suggests that two features that are shared by many corpus-based studies are that “they make extensive use of computers for analysis, using both automatic and interactive techniques” and that “they depend on both quantitative and qualitative (interpretative) analytical techniques” (1996:172). Biber does not describe in any detail what he means by “extensive use of computers”. For many corpus-based studies computers are used primarily as a means to retrieve concordances that are then examined and interpreted manually. Other studies make use of advanced tools and methods for automatic analysis. With the increased availability of various general purpose corpus tools (such as WordSmith Tools), it has become relatively uncomplicated to use computers not only to identify and retrieve instances of a word or phrase but also to analyse, for example, collocation patterns.

Biber (1996) also suggests that corpus-based studies use a variety of techniques for analysis; automatic and interactive, and quantitative and qualitative (interpretative). In the sections that follow, I will describe the methods I have used for data retrieval and analysis, and I will also illustrate how I have identified and examined systematic variation. The discussion also contains a presentation of the tools that I have used.

3.5.1 Data retrieval

For my studies, I have used computers and a variety of computer programs (see below) to search the corpora, identify instances of the expressions of future, and display them in the form of concordances (the expressions listed as surrounded by a certain amount of context). The expressions have then been counted and analysed. The data have, primarily, been analysed quantitatively as frequencies and proportions (see below), and as far as possible, automatic methods have been used in the form of corpus search tools or concordance programs. The tools I have used are presented further below in Section 3.5.2.

The tools/programs not only produce concordances but also have features that allow sorting of the instances according to various criteria, such as the words preceding or following the expression or according to what part of the corpus/text category the instances are retrieved from. Additional features allow the user to delete unwanted instances, change the amount of context shown, and in some cases make it possible to easily create and examine collocations in which a particular expression is found.

Where tagged corpora are used, it is possible to restrict a search to identify only those instances of an expression with a certain POS tag, such as *will* tagged as a modal auxiliary or *going to* tagged as verb+infinitival marker. As suggested above, tagged corpora may contain errors so that a POS-restricted search either does not find all relevant instances or retrieves unwanted examples (or possibly both). To estimate the extent to which this affects my figures I have evaluated the tagging accuracy for the corpora and searches where the tagging has been used. I have found that the tagging accuracy is sufficiently high to motivate the use of the statistics derived on the basis of the tagged output, and my analyses in Studies II, III, and V are thus based on searches where the POS tags have been considered.

For the studies where I have used untagged corpora, or where the corpora used in a study have not been tagged with one and the same tagger or tag-set, I have chosen to rely on manual identification of expressions of future. In these cases (Studies I and IV), I have used corpus handling programs to identify all instances of an expression and have then manually analysed the output to identify and remove the unwanted instances. In the process of identifying unwanted examples, the 'sort' function in the corpus programs has been very useful, as it enables one to locate, for example, all instances of *will* preceded by an article or a particular personal pronoun (which are highly likely to be nominal uses of the word, such as *in <u>his</u> will*) or *going to* followed by articles or certain nouns (for example, *going to <u>the market</u>*, *going to <u>school</u>*, *going to <u>bed</u>*). In cases where it has been impossible to determine whether a particular instance of an expression is to be considered an expression of future or not on the basis of the syntactic environment, the instance has been included in the overall counts. An example of such an unclear instance is given in (12):

(12) <unclear>.
Sorry yes.
I'm I'm really *going to* erm <unclear> I afraid I have experience of
Who are you sorry? (BNC D91 256-259)

As the number of manually scanned instances is very high and the number of unclear cases low, occasional individual errors will not affect the overall statistics (for Study IV, for example, over 8,000 instances of *will* and around 1,800 instances of *going to* were examined).

As described in Chapter 1, I have not made any semantic analysis of the expressions to distinguish between more or less futural instances of an expression for this study. In addition to *will* used as a modal auxiliary and *going to* where *to* is the infinitival marker, all instances of *shall, 'll,* and *gonna* have thus been included in my analyses.

3.5.2 Tools used

For my studies I have chosen to use existing tools rather than create my own or revert to manual options. The decision to use corpus handling tools rather than manual identification is to a considerable extent a matter of practicality. The process of manually going through the corpora I have used to collect the instances is far too time-consuming, not least as it is necessary to go through the data several times to ensure that the relevant instances have been identified. Moreover, any subsequent analysis, such as relatively simple cross-tabulations, becomes a laborious task.

The reason why I have chosen generally available tools is once again to ensure that my studies can be reproduced and my results verified or used for comparison. The tools I have used are all generally available. Some are general-purpose tools that can be used on any corpus or text while others are created specifically for a particular corpus. The tools are presented below.

3.5.2.1 WordSmith Tools

WordSmith Tools is a suite of programs developed by Mike Scott (see WordSmith Tools webpage Scott). The *Concord function* in Version 3 has been the principal tool used for Study IV and for any additional comparisons presented in the Summary. The concordancer finds strings (words, parts of words, or sequences of words/parts of words), and allows the user to sort, annotate and study examples listed in the concordances in a number of ways, for instance by using the *collocation*, *cluster*, and *plot* functions. For a discussion about the use of the program and other related issues, see Study IV.

3.5.2.2 Qwick

Qwick is a corpus tool developed by Oliver Mason (University of Birmingham). For the present study, the program was used primarily to compare the results obtained with other programs or through manual inspection, in particular in connection to the case-studies presented in Study IV.[19] Qwick is an easy-to-use tool which efficiently makes concordances and calculates collocations on corpora that have been indexed for use with the program. Further functions are available, but have not been used much for the present study (see also Study IV).

3.5.2.3 SARA

SGML-Aware-Retrieval-Application (SARA) is the custom-made retrieval program provided with the BNC. The program is particularly valuable as it makes it possible to combine searches for linguistic information (words or phrases with or without POS information) with searches for the extra-linguistic information contained in the corpus files (text, author or speaker related information). It is thus possible to get concordances, for example, for a particular word or phrase produced by a certain speaker or speaker category or published in a particular text category. In this study SARA (versions 0.930 and 0.931) has been used for case-studies concerning the BNC and Sampler corpora, often in combination with or as a complement to other programs. For more information, see the SARA webpage.

3.5.2.4 BNCweb

The BNCweb program is developed at the University of Zurich by Sebastian Hoffmann, Hans-Martin Lehmann, and Peter Schneider. The program is based on the SARA search and retrieval program and makes available a

[19] I am grateful to Oliver Mason for making available pre-release versions of the program and for providing valuable guidance in the installation and usage procedures.

web-based interface to the BNC.[20] The user-friendly interface not only makes it easy to search the corpus, but also offers advanced alternatives for processing the search results in a number of ways and for retrieving information about the (extra-linguistic) features of each hit. The collocation/colligation function is extremely useful for studies such as that presented in Study III. Other valuable functions that have been of major importance for my work are the database function and the distribution display. The BNCweb program has been used for all of my studies involving the BNC and for certain parts of the studies involving the Sampler corpus.

3.5.2.5 Miscellaneous

In addition to the tools listed above, occasional use has been made of other resources, such as the WordCruncher, MonoConc, and MicroConcord programs, as well as a few PERL-scripts and UNIX commands.

3.5.3 Analysis

For my studies, the retrieved instances of the expressions of future have been analysed with quantitative and qualitative analytical techniques. A certain emphasis has been placed on the quantitative analysis. The distribution of the expressions has been examined across different corpora and sub-corpora defined according to various criteria. As the main focus has not been to examine the degree of futurity in a text or corpus or the extent to which expressions of future occur at all, the overall frequencies of the expressions have only been considered briefly. The emphasis has instead been put on examining the variation between the five expressions across different samples of text.

For this purpose, it has been considered useful to base the analysis on the proportions of the different expressions. The proportion of an expression is given as a percentage, calculated on the basis of the combined frequency of all five expressions, as illustrated for *will* in LOB in Figure 3.1 (frequencies from Study I:15).

$$\frac{\text{Frequency } \textit{will}}{(\text{Frequencies } \textit{will} + \textit{ll} + \textit{shall} + \textit{going to} + \textit{gonna})} = \text{Proportion } \textit{will}$$

$$\frac{2{,}316}{(2{,}316 + 505 + 363 + 170 + 2)} = 0.69 = 69\%$$

Figure 3.1. Formula for calculating proportions. Example: *will* in LOB (frequencies from Study I:15)

[20] I am indebted to the Zurich team for allowing me to use early versions of the program.

It has been suggested (see Chapter 2) that the expressions *will/shall* and *going to* (with respective reduced forms *'ll* and *gonna*) can at times be used interchangeably. If they were truly synonymous in every sense we would expect the expressions to be found randomly distributed across all kinds of texts so that the proportion of an expression would be the same, or similar, in any sample or text. If the expressions were used interchangeably, there would be no systematic differences in the usage patterns of the expressions and no expression would occur proportionately more frequently in one kind of text than in another. As it happens, this is not the case. In my studies I have shown that the proportions of the five expressions of future vary between different samples of text (where the samples of text are defined on the basis of certain non-linguistic features). This suggests that the expressions are not distributed randomly across the texts but there are certain patterns of usage, systematic ways in which the expressions are used.

Biber (1996:173) uses the term 'association patterns' to describe "the systematic ways in which linguistic features are used in association with other linguistic and non-linguistic features".

Association patterns

A) Investigating the variability of a linguistic feature (lexical or grammatical)
i) Non-linguistic associations of the feature:
- distribution across registers
- distribution across dialects
- distribution across time

ii) Linguistic associations of the feature:
- co-occurrence with particular words
- co-occurrence with grammatical features

B) Investigating the variability among texts:
- "dimensions" = co-occurrence patterns of linguistic features

Figure 3.2. Association patterns (from Biber 1996:174)

He suggests that a corpus-based approach is useful for studying such association patterns. According to Biber, there are primarily two kinds of research questions that are investigated through the study of association patterns: "the variability of a linguistic feature, and the variability among texts" (see also Figure 3:2 above). In my studies I have dealt with the first of those, the variability of a linguistic feature: the expressions of future.

The linguistic feature under investigation in my studies is, thus, the expressions of future in the FUT paradigm presented in Chapter 1 (*will, 'll, shall, going to, gonna*). The non-linguistic associations I have examined are distribution across medium, text category, speaker properties, region, and time. I have also investigated some linguistic associations: co-occurrence with particular words and co-occurrence with items of particular grammatical classes . The results of my studies of the association patterns of the expressions of future are presented in Chapters 4-8 (non-linguistic associations) and Chapter 9 (linguistic associations).

3.5.4 Identifying systematic variation

I have used different methods to examine to what extent there is variation in the proportions of the expressions of future. The starting-point has been to look for patterns of variation, to see what indications there are of a systematic variation. One important indication has been consistency. If similar variation in the proportion of an expression is found in several samples, that is an indication of systematic variation. As an example of this kind of variation one can mention the proportion of *will* in Imaginative and Informative texts distinguished in corpora such Brown and its clones. The proportion of *will* is larger in Informative texts than in Imaginative in four comparable corpora, which suggests that this variation is systematic. The proportion of *will* is, however, not consistently larger or smaller in comparable corpora from different periods, which suggests that the use of the expression does not vary systematically with the time when the text was produced.

An important factor to be considered in this context is also the basis of variation. Variation, consistent or not, is obviously less informative if the differences between the examined samples (for example different sub-corpora) are slight, as such differences could be the result of chance or due to random variation in the samples. It is, however, not only the scale of variation that needs to be evaluated. When evaluating variation between different corpora or sub-corpora in the proportion of expressions of future, it is useful to consider not only differences between the proportions of the expressions but also to take into account what the proportions are based on. Proportions based on corpora or sub-corpora where the overall raw frequencies of the expressions are low are of limited interest as such proportions fluctuate considerably also with small differences in number. Proportions derived from texts where there is a larger number of instances are more reliable, and as such of greater interest in this context.

Using statistical tests, it is possible to determine how likely it is that a factor, linguistic or extralinguistic, influences the distributions attested in the data. The test which I have used is presented in Figure 3.3 (see, for example, Butler 1985, Mendenhall et al. 2003). The test takes into account the size of the examined sample of data is (in my case defined as the total number of expressions of future in the corpus or sub-corpus) as well as the proportion of the expression in question.

$$z = \frac{p1 - p2}{\sqrt{P * (1 - P) * (\frac{1}{N1} + \frac{1}{N2})}}$$

N1 Frequency of *will+ 'll+shall+going to+gonna* in Sample1
N2 Frequency of *will+ 'll+shall+going to+gonna* in Sample 2

p1 Proportion of examined expression in Sample 1 $\frac{n1}{N1}$, where n1 = raw frequency of examined expression of future in Sample 1

p2 Proportion of examined expression in Sample 2 $\frac{n2}{N2}$, where n2 = raw frequency of examined expression of future in Sample 2

P Overall proportion in the two samples $\frac{(n1 + n2)}{(N1 + N2)}$

Figure 3.3. Formula used for identifying statistically significant differences between proportions.

The formula in Figure 3.3 will return a value for *z* which can then be checked for statistical significance. It is customary in linguistic studies to use the significance level 5% (see, for example, Oakes 1998, Woods et al. 1986). In such a case, there is thus a 5% probability that an identified difference between two samples is merely randon and not an actual difference between the populations.

When examining real language data, it is difficult to establish to what extent factors outside one's control influence the distribution, in the present case factors such as topic of the text or stylistic considerations on the part of the speaker. Statistical tests of significance usually require that the data tested should fulfil certain criteria which are difficult to measure where natural language is concerned (such as normal distribution and independent observations). Moreover, it cannot be excluded that some of the distributions I compare (in particular *gonna* and *going* to) are not those of two discreet entities but are more to be compared to two arbitrary points on a continuous scale. To minimise the error in any claim of a significant difference, a very high threshold has been set. I have considered a 5% probability to be too large, and have instead chosen to only regard differences on the 1% level as statistically significant. Raw frequencies are presented in Appendix B to allow alternative calculations to be made, should so be desired.

Woods et al. (1986) discuss the value of statistical hypothesis testing. They suggest that “[t]he value of statistical hypothesis testing as a scientific tool has been greatly exaggerated” (1986:127) and that “[i]t makes much more sense to discuss the details of the data in a manner which throws as much light as possible on the problem which you intend to tackle” (1986:130). That is what I do in the following chapters where I sum up how the expressions are distributed across the different corpora and sub-corpora.

3.6 Summary

In this chapter, I have discussed issues relating to the use of corpus linguistics as an analytical framework against the background of the characteristics outlined by Biber (1996). Biber suggests that there are certain characteristics that corpus linguistic studies share. These characteristics relate to the choice of data and the methods of analysis as well as to the fact that the studies are concerned with the analysis of actual patterns of use. I have shown how my studies fit within this framework, and discussed issues relating to my choice and use of material in my studies.

Sinclair suggests that the quality of the results of a corpus-based study is dependent on the corpus (1991). To be able to properly evaluate a corpus-based study it is important to know about the composition of the corpus on which it is based. In this chapter I have aimed at providing information about corpora in general and the corpora I have used in particular. I have explained the motivation behind selecting the corpora I have chosen, and also described how the corpora have been used in my studies.

The following chapters will focus on the results of my case studies. The findings in my published articles (Studies I–V) are summarised and supplemented with results from further case studies to illustrate how the use of the expressions of future varies with a number of extra-linguistic and linguistic factors.

4. Medium

4.1 Introduction

In Studies I and IV, it was shown that the use of the expressions of future varies considerably between written and spoken corpora. The present chapter summarises the results of the two studies and discusses the differences that have been observed. One question that was raised in the studies was the use of the expressions in speech-like contexts. This issue is pursued here through three unpublished case studies. These studies show that some of the variation in the distribution of the expressions of future within written corpora can actually be accounted for with reference to written-spoken differences.

It is a well-known fact that written language and spoken language differ in a number of ways (for an over-view of the research, see for example Biber 1986a, 1986b). Differences are noted between written and spoken language where the use of the expressions of future are concerned, and the *going to* expression has received considerable attention in this context. In their corpus-based *Longman Grammar* for example, Biber et al. (1999) describe the distribution of the modal and semi-modal verbs across different kinds of texts. It is shown that these verbs are most common in conversation. The difference is particularly noticeable for the semi-modals (such as *be going to*) which as a group are five times more common in conversation than in what the authors refer to as 'written expository registers'. It is stated that "[t]he semi-modal *be going to* is a common way of marking future time in conversation (and fictional dialogue), but is rarely used in written exposition" (1999:490). That *going to* is more frequent in spoken language has also been noted by a number of other authors, although rarely in combination with detailed accounts of quantitative distributions. Palmer, for example, suggests in his earlier work that the pattern where *going to* is used for future reference "is very common, indeed, probably more common than sentences with WILL and SHALL in ordinary conversation" (1965:63, 1974:37). In a later publication he changes this into "[f]orms with BE GOING TO are very common in colloquial speech" (1988:38 no particular definition for 'ordinary conversation' or 'colloquial speech' is given).

In what follows, this chapter first provides an introduction to the overall distribution of the expressions of future across some spoken and written corpora. I then move on to discuss the use of expressions of future in quoted contexts, and to examine the use of *gonna* in written text.

4.2 Overall frequencies and proportions

As noted in the previous section, Palmer (1965, 1974) suggests that *going to* is probably more common than *will/shall* in conversation. With access to corpus data, it is possible to establish to what extent Palmer's claim holds. Studies I and IV showed that the distribution of not only *going to* but also the other expressions of future varies between written and spoken texts. (There is also considerable variation within the greater categories of written and spoken text, something which is discussed further in Chapter 5 below).

The expressions of future are more frequent in spoken than in written language. Study IV showed that the frequency (per million words) found in the written British English LOB and FLOB corpora varies between 3,088 (FLOB) and 3,362 (LOB), while the frequency in the spoken corpora is considerably higher, between 5,706 (LLC) and 9,392 (Sampler DS) per million words.[21] Figure 4.1 illustrates the frequency of the five expressions in these corpora: the written LOB and FLOB and the spoken LLC, Sampler CG and Sampler DS (see also Tables A.1, A.2, A.5, and A.6 in Appendix B).

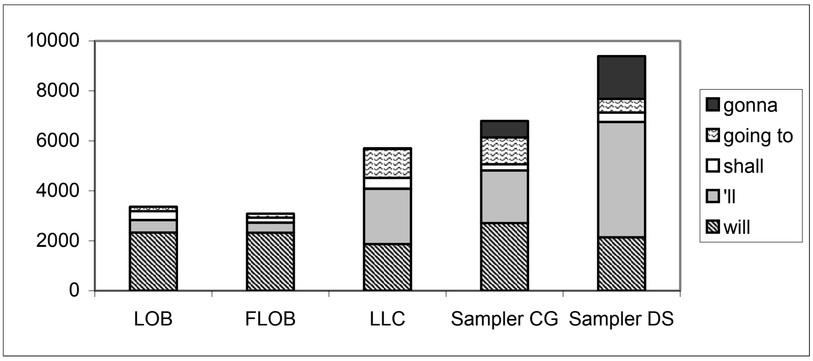


Figure 4.1. Expressions of future in the written LOB and FLOB corpora and the spoken LLC and BNC Sampler (CG and DS components). Frequency per million words (from Study IV:32)

As described in Studies I and IV, it is not only the frequency of the expressions that differs with medium but also the proportions of the different expressions. The written corpora contain large proportions of *will* and small proportions of the other expressions. *Going to* is used in only about 5% of all cases in the written corpora, less than any other expression (excluding *gonna* which is rarely found in written corpora at all). In the spoken corpora, the

[21] The frequency (per million words) of the expressions of future in LLC varies between those given in Study I and those in Study IV. The difference can be explained with differences in the method of extraction, described in Chapter 1. Study IV also discusses issues related to identifying instances of an item in LLC.

expression *'ll* (which is relatively rare in the written material) is generally the most frequent expression. The proportion of *will* is considerably smaller in the spoken than in the written corpora, while the proportion of *going to* is larger, even if the difference is not quite as noticeable as for *'ll*. That the *'ll* expression is more frequent in spoken language is, perhaps, not surprising. Kjellmer (1998:159), for example, notes that: "[i]t is a common observation that contraction is particularly frequent in speech", and he quotes a number of sources referring to this fact. My results concerning the distribution of *'ll* tie in well with those presented by Kjellmer.

At first glance there does not seem to be any consistent difference between the written and spoken corpora as far as the proportion of *shall* is concerned (see Figure 4.1). The expression is used most in the written LOB corpus (11%), followed by the spoken LLC (8%), FLOB (6%), and least in the spoken CG and DS corpora (4%). In Chapter 8 it is shown that the use of *shall* seems to have decreased from the 1960s to the 1990s. When the earlier corpora (LLC and LOB) are compared to each other, the expression *shall* turns out to be proportionally less frequent in the spoken corpus (LLC) than in the written (LOB). Similar results are obtained when the later FLOB is compared to the spoken parts of the Sampler (both CG and DS components); the proportion of *shall* is smaller in the Sampler components than in FLOB. The use of *shall* thus seems to vary with medium, being less frequent in the spoken corpora, although this difference is not as marked as that for the other expressions.[22]

The expression *gonna* is frequently referred to as a spoken, informal variant of *going to. Gonna* is almost exclusively found in the spoken corpora but to a varying degree. The proportions of *going to* also vary between different spoken corpora. In the LLC, for example, the proportion of *going to* is as high as 20% while it is only 6% in the Sampler DS, a proportion similar to that in the written corpora. The proportion of *gonna* is high in the DS component (18%) to be compared to less than 1% in the LLC. As discussed for example in Study II, it is difficult to compare the use of *gonna* and *going to* as it cannot be excluded that the distribution of the two expressions is affected by differences in transcription practices (see discussion in Section 3.2.3.2). Furthermore, it has been shown (for example in Study III) that the *gonna* and *going to* expressions are used in very similar lexical and syntactic contexts. It can thus be justified to conflate the frequencies of the two expressions to be able to get an indication of whether there are any substantial differences between the written and spoken corpora or between different spoken corpora in the use of the expressions of future. When the figures for

[22] As the LLC and Sampler corpora are not comparable in the same way as the LOB and FLOB corpora these results should be interpreted with caution. The spoken corpora differ not only as to the time of sampling but also across a number of other parameters (see Section 3.4.2).

gonna and *going to* are combined, it becomes apparent that *gonna+going to* belong predominantly to the spoken texts. In the spoken corpora, 20–26% of the expressions of future are *gonna+going to*, to be compared to a proportion of only 5% in the written material. The combined proportion of the expressions *will+'ll+shall* is around 85% in the written corpora and just over 70% in the spoken. It is therefore obvious that *going to* is more frequent in spoken than in written texts. It is, however, not the case in any of the corpora examined in my studies that the expression is more frequent than *will+'ll+shall*, as is suggested by Palmer (1965, 1974).[23]

4.3 Quoted context in writing

As shown above, the distribution of the expressions of future varies considerably between written and spoken corpora. The use of the expressions also varies between different kinds of spoken and written text, as further discussed in Chapter 5. In Study I, I compare and contrast Informative and Imaginative texts in written corpora, and suggest that the variation in the use of expressions of future may be related to the differences between media. The *'ll* and *going to* expressions are proportionately more frequent in contexts similar to spoken language, such as dialogues, quotes, reported and imagined speech (Study I:16). Such contexts are more frequent in the Imaginative text categories and that coincides with the higher proportions of *'ll* and *going to*.

I have pursued the topic of variation in the use of the expressions of future in speech-like text in three unpublished case-studies. Three subsets of written corpora were studied with regard to the distribution of the expressions in quoted contexts.[24] In the first subset, all expressions in two text categories (A: Press Reportage and K: General Fiction) in two comparable corpora (LOB and FLOB) were examined. This approach makes it possible to obtain a broad, general picture of the distribution of the expressions in quoted and non-quoted contexts. The second subset consisted of all instances of the expression *going to* in the four written comparable corpora used in my published studies (see Studies I and IV). The reason for examining all instances was that the *going to* expression is infrequent in most written text categories. With this method it was possible to obtain detailed information

[23] Higher proportions of *going to* than *will* are found by Facchinetti (1998) and Poplack and Tagliamonte (2000) in their studies on British Caribbean Creole and Afro American Vernacular English, respectively. These studies are commented on further in Section 7.3.

[24] By quoted contexts is understood that the expression is found in a clause, sentence or paragraph surrounded by single or double quotation marks. This means that other speech-related instances, such as dialogues and indirect speech are not included in the counts, while quotes that are not spoken passages are. Spot checks show, however, that the number of instances that were erroneously included or excluded was negligible.

on how the expression is used in quoted and non-quoted contexts and see to what extent the proportions vary between the corpora and text categories. This information was useful for evaluating the result of the study of the first subset, as well as when the use of the expression was examined in different types of text (Chapter 5) and compared across corpora of different regional varieties (Chapter 7). The third subset of data where the proportion of quoted context was examined was drawn from the written part of the BNC. All instances of *gonna* and a sample of occurrences of *going to* were studied. In addition to identifying instances in quoted context, the non-quoted instances were examined to determine to what extent *gonna* can be found in other than speech-related contexts. The results of the examination of quoted contexts in the three subsets are presented below. They support the initial findings made in Studies I and IV, showing that the proportions of instances in quoted contexts vary both between the expressions and between different kinds of text.

4.3.1 Subset 1: Text categories A and K in LOB and FLOB

The first subset was drawn from the two comparable corpora of written British English, LOB and FLOB. Two text categories from each corpus were examined, one Informative (A: Press: Reportage) and one Imaginative (K: General Fiction). All instances of the expressions of future in these text categories were examined, and the proportion of instances found in quoted contexts was calculated (see Table 4.1).

Table 4.1. *Quoted instances. Proportions in percentages and raw frequencies (the number of quoted instances and the total number of instances given in parentheses)*

Text category	***will***	***'ll***	***shall***	***going to***	**Total**
LOB A: **Press Reportage**	15% (46/313)	90% (9/10)	93% (13/14)	64 % (7/11)	22% (75/348)
FLOB A: **Press Reportage**	27% (98/361)	95% (20/21)	100% (3/3)	73% (16/22)	34% (137/407)
LOB K: **General Fiction**	79% (76/96)	80% (53/66)	81% (22/27)	69% (11/16)	79% (162/205)
FLOB K: **General Fiction**	47% (31/66)	88% (53/60)	50% (3/6)	59% (10/17)	65% (97/149)
Total of examined text categories	30% (251/836)	86% (135/157)	82% (41/50)	67% (44/66)	42% (471/1,109)

Table 4.1 shows that in the examined samples, the expressions of future appear in quoted context to a considerable extent. The majority of the instances of *'ll, shall* and *going to* are used in quoted context, while the proportion of quoted instances of *will* is generally smaller. This fact correlates with the findings which show that *will* is used more in written than in spoken texts. All figures, especially for the less frequent expressions, should be interpreted

with caution since the raw frequencies are low and the proportions consequently subject to great fluctuation even with small variation in the number of occurrences.

The proportion of *'ll* found in quoted contexts is large in both text categories in the two corpora. The large proportions are not surprising against the background that contractions in general are more frequent in speech and also used as speech-reflecting devices (see, for example, Axelsson 1998, and Kjellmer 1998). The very large proportion of quoted *'ll* in text category A (90% in LOB, 95% in FLOB) ties in well with the results obtained by Axelsson who, in her study of press texts, showed that contracted forms were used to a high degree in quotes. Axelsson also observes that the function of non-quoted contractions "...seems to be to make the text more accessible and reader-oriented through a conversational turn of phrase" (1998:212). Her discussion refers to British newspapers, but could also apply to the use of non-quoted contractions in other text categories, such as that studied here in (K :General Fiction). The instances of non-quoted *'ll* in this text category are often conversational, as in (1):

(1) The cabinet minister has gone inside. David's coming over. I*'ll* get a chair for him – one of the folding canvas ones. He*'ll* expect that. (FLOB K26 120-122)

Shall is found to a large extent in quoted contexts. It is similar to *'ll* in that respect. Unlike *'ll,* however, *shall* is not used in spoken texts to a great degree (see Section 4.2, Figure 4.1). This means that while the high proportion of quoted instances of *'ll* can be seen as an example of how the expression is used as a speech-reflecting device, the high proportion of quoted instances of *shall* (82% overall) has to be explained in other ways. According to Wekker (1976:47), "... *shall* often reflects greater formality than the corresponding *will*-form, and is perhaps for this reason felt to be more assertive in certain cases". That may be one explanation for the high proportion of quoted instances in text category A, which is press texts. Speakers choose *shall* when they want to be assertive, and it is those assertions that are found important enough to be rendered verbatim in the journalistic text. As the number of instances of *shall* is very low (50 instances in the four investigated text categories together), the proportions are sensitive to small differences in absolute numbers. It is also difficult to draw conclusions about the reasons for the high proportion of quoted *shall* on the basis of the occurrences examined. It is shown elsewhere (for example Section 5.2.3.2 and Study IV) that *shall* is particularly frequent in certain individual texts dealing with specific topics. I suggest that this fact contributes to making the expression rare in other contexts. It is felt to be marked, and therefore it is rarely used in the normal running text but primarily found in quotes.

As for *going to*, the pattern is that the expression is used in quotes less than *'ll* and *shall* but more than *will* (67% altogether). This corresponds to the results presented above, where it is shown that the expression is used more in spoken than in written text, but that the difference is not as marked as for *'ll*. The difference between the proportions of quoted contexts of *going to* and *'ll* can be seen in the light of the suggestion that *'ll* is a speech-reflecting device, a finding that is supported by the large proportion of *'ll* found in quotes and other contexts similar to spoken language (Study I:16). *Going to* is not as characteristic of spoken language: the difference between the written and spoken corpora is not as large as for *'ll.* A possible further indication of this can be found reflected in the difference in the proportions of *going to* and *'ll* found in quoted contexts.

4.3.2 Subset 2: *Going to* in four written corpora

As shown in Table 4.1, the number of instances of *going to* is low in all the text categories in which the proportion of quotes was examined. The proportions are, however, based on only a few instances and the result is consequently sensitive to possible variation with other factors (for example author style or the topic of single texts). To get a better understanding of the factors governing the use of *going to*, I have examined further to what extent the expression is used in quoted contexts in the four comparable written corpora that I have use in my published studies. All instances of the expression were examined and the results can be found in Table 4.2 (these results are also referred to in Chapters 5 and 7).

The proportions of quoted instances vary considerably between the text categories, which to some extent can be accounted for by the fact that the number of instances is low and the proportions therefore also sensitive to small variation in the raw frequencies. The general pattern is, however, that the proportion of quoted instances is considerably larger in the Imaginative text categories (text categories K–R) than in the Informative (text categories A–J), albeit with great differences within the hyper-categories. This pattern is found in all four corpora. The text categories that show the most deviating proportions are generally those where the number of instances is the lowest. This variation correlates with the finding that *going to* is more frequent in the fictional hyper-category (text categories K-R).

Table 4.2. *Quoted contexts of* going to *in the LOB, FLOB, Brown, and Kolhapur corpora. Proportions of all instances (%) and raw frequencies (italics = less than 15 instances of the expression in the text category; - = no instances).*

Text category*	LOB	FLOB	Brown	Kolhapur
A Press: Reportage	*64%* *11*	73% 22	*44%* *9*	*0%* *4*
B Press: Editorial	*20%* *10*	*14%* *7*	19% 16	*0%* *9*
C Press: Reviews	*0%* *3*	*15%* *13*	*50%* *2*	*0%* *3*
D Religion	*20%* *5*	- *0*	*0%* *1*	*0%* *1*
E Skills and Hobbies	*0%* *12*	*17%* *7*	*0%* *6*	*33%* *3*
F Popular Lore	*14%* *14*	*33%* *6*	*38%* *8*	*75%* *4*
G Belles Lettres, Biography, Memoirs, etc	*20%* *10*	*0%* *6*	*100%* *2*	*20%* *5*
H Miscellaneous government and official documents	*0%* *4*	*0%* *4*	*0%* *1*	*25%* *12*
J Learned	*0%* *5*	*20%* *5*	- *1*	*50%* *2*
K General Fiction	69% 16	59% 17	*100%* *15*	79% 19
L Mystery and Detective Fiction	85% 20	89% 19	92% 26	*43%* *7*
M Science Fiction	*80%* *5*	*67%* *6*	*100%* *5*	*100%* *1*
N Adventure and Western Fiction	92% 28	76% 25	94% 22	*100%* *1*
P Romance and Love Story	88% 32	81% 17	95% 24	*80%* *10*
R Humor	*50%* *2*	*83%* *7*	*50%* *4*	*100%* *5*
Total	57% 177	57% 161	71% 142	49% 86
Informative hyper-category (A-J)	19% 74	33% 70	30% 46	21% 43
Imaginative hyper-category (K-R)	84% 103	76% 91	93% 96	77% 43

* Text category labels are taken from the Brown Corpus Manual (Francis and Kucera 1979) and may vary slightly from those given in the other corpus manuals.

4.3.3 Subset 3: *Gonna* and *going to* in the BNC

In the previous section, it was shown that *going to* is relatively infrequent in many text categories, which makes it difficult to interpret the results of the calculation of quoted and non-quoted instances. The expression *gonna* is even rarer in the corpora examined above, which, of course, makes it impossible to draw any conclusions about the extent to which the expression is used in quoted and unquoted contexts on the basis of this material. Instead, I used the written part of the BNC, constituting about 90 million words, to investigate to what extent *gonna* occurs in quoted or speech-like contexts. To see to what extent the low frequency of *gonna* in the written data can be seen as a reflection of the fact that the expression is primarily a spoken variant, the distribution of the expression has been compared to that of *going to* in the same corpus.

The written part of the BNC contains 546 instances of *gonna*.[25] All occurrences were examined to find how the expression is used in written language, in order to establish to what extent it is used exclusively in spoken (or speech-like) language, as is frequently suggested in other studies. For comparison, a sample of instances of *going to* (with past and present forms of *be*) in the same corpus was also examined. There are almost 19,000 instances of the full form *going to* (where *to* is tagged as the infinitival marker *to:* TO0), a subset of which was examined for this study.[26]

The occurrences of the expressions were examined and classified according to the kind of context in which they were found. To get a fuller picture of the use of *gonna,* all instances that were not used in quoted contexts were classified further according to the context where they occur: Speech, Song, Headline, Written, and Drama, as described below. For comparison, the sampled instances of *going to* were classified in the same way.[27] The result of the analysis is found in Table 4.3. As can be seen in Table 4.3, the majority of the occurrences of *gonna* are classified as ‘quotes’ (75%). As in the studies of the other two subsets described above, ‘quote’ is used for the instances found in the text (clause, sentence, paragraph) surrounded by single or double quotation marks (“ ” or ‘ ’). The proportion of *going to* in quotes varies considerably between the different text categories, but is consistently smaller than that of *gonna* (15%–64% compared to 75%). The proportions of expressions found in quotes are largest in the Imaginative text category for both *gonna* and *going to* (84% and 64% respectively).

[25] This is the number obtained when searching for the phrase *gon na* in the first version of the corpus with the original index. Instances rendered as *gonna* have also been included (see Studies II and IV for a discussion of issues related to the index and the process of searching the corpus).

[26] The *going to* sample consists of a proportion of the instances from the ‘Imaginative’, ‘Commerce and finance’, and ‘Natural and pure science’ domains (the domains with largest, median, and lowest number of occurrences respectively). Since the populations are so varied in size, different proportions of them have been extracted for this subset: 5% of the Imaginative (which makes the sample equal in size to the *gonna* sample), 50% of the Commerce, and 100% of the small Natural science population. The random sample function in the BNCweb programme was used to make the selection.

[27] As the statistics for *going to* are derived from a limited sample, I have chosen not to estimate a ‘total’ figure. The figures in the ‘average’ column are based on the proportions found in the three examined domains. The results from each domain have been given equal weight to compensate for differences in sample size.

Table 4.3. Gonna *and* going to *in the BNC (written part).*

gonna				context classification	*going to*						
All instances of *gonna*		Imaginative			Imaginative		Natural science		Commerce		average percentage*
408	75%	272	84%	Quote	361	64%	19	15%	85	24%	34%
21	4%	7	2%	Speech	11	2%	2	2%	5	1%	2%
42	8%	7	2%	Song	0	0%	0	0%	0	0%	0%
3	<1%	0	0%	Headline	0	0%	0	0%	0	0%	0%
68	12%	33	10%	Written	187	3%	110	84%	258	74%	64%
4	1%	3	1%	Drama	1	1%	0	0%	0	0%	0%
546		**322**		**total no**	**560**		**131**		**348**		

* To compensate for differences in sample size, the average percentage was calculated by giving equal weight to each sample (for example, (64+15+24)/3=34).

A few instances of both variants have been classified as 'speech'; they are found in a speech-like context, without being rendered within quotation marks. Among these examples are some transcripts of spoken language and some occurrences from texts based on interviews. Three occurrences of *gonna* are found in headlines, and four in dramas. In all 68 instances of *gonna* have been classified as 'written', meaning instances that do not occur in quotes, speech, songs, headlines, or drama. Of these, 19 were found in e-mail messages sent to the Leeds United E-mail List. Although the texts (each consisting of several shorter e-mail messages) also display instances of *going to*, these do not co-occur with the contracted form in any one message. This may indicate that the choice of the variant is a matter of personal preference. This cannot be attested or refuted in the present material, however, as the texts have been anonymised in the corpus, making it impossible to say whether any one individual has contributed to the text with more than one message.

Over a quarter (20 instances) of the 'written' instances of *gonna* are found in one single text, H8M, an extract from the book *Underground* by Russell James.[28] That text is the one with the most instances of *gonna,* 58 occurrences altogether (found in quotes as well as in non-quoted contexts). Other texts in the corpus that include 'written' *gonna* are a number of texts from a music magazine (*New Musical Express,* six texts with ten instances altogether), from a magazine for teenagers, and from a magazine about computing. It can perhaps be speculated whether the use of *gonna* is an attempt to achieve a modern variety of language, to mark the relationship to a particular reader group (youth culture), or something else. Unfortunately the amount of available data is too limited to draw definitive conclusions.

[28] *Underground* by Russell James. London: Victor Gollancz Ltd, 1989, pp. 44-169; 46,636 words.

Among the other ‘written’ instances, many show close resemblance to spoken language, even though they do not occur within passages marked with quotation marks or passages that can be defined as ‘spoken’, as in:

(2) One she'd made herself: a black and white Mae West wearing the smile that says I'm *gonna* eatcha and you're *gonna* love it, big boy. (BNC HGF 1193)

(3) On Friday, I receive a letter. Jezebel ... Don't listen to a word they say. I'm clean, my love, I'm *gonna* make a go of it. (BNC HGL 3454)

There are only a few more instances of ‘written’ *gonna*. Even in those examples, it is easy to imagine that the use of the reduced form derives from a wish of the author to mirror spoken or informal usage, as in (4).

(4) He informed his wife that when he died he wished to be cremated, for fear that any preparation for burial might involve a pimply, eighteen-year-old undertaker poking around the corpse for a while before sadly informing Mrs Sinclair that the job would take a fortnight and that he couldn't guarantee a result seeing as how the corpse had clearly been carelessly handled in the past, and above all, it was definitely *gonna* cost her. (BNC FPS 1225)

The classification ‘song’ has been given to 42 instances of *gonna*. These are instances found in songs, or titles of songs or recordings. There were no instances of *going to* used in that context found in the sample.

Among the instances of *going to,* few examples are found in other contexts than ‘quotes’ or ‘written’. The proportion of ‘written’ instances is considerably larger in the text categories ‘Natural and pure science’ and ‘Commerce and finance’ than in the ‘Imaginative’ texts, where it is closer to that found for the *gonna* examples. Since the number of occurrences of *gonna* is small, it is not possible to compare the proportion of various contexts across the text categories. Table 4.3 includes the result of the analysis of the single largest group of *gonna*, found in the ‘Imaginative’ text category. Although the proportion of *gonna* in ‘quotes’ is larger than that of *going to* in this text category (75% vs. 64%), the distribution across the contexts is still more similar for the two expressions in the Imaginative domain than for *going to* in different text categories. It is also similar to the distribution of quoted *going to* in the Imaginative hyper-category in the four corpora examined above (see Table 4.2).

4.3.4 Conclusions regarding quoted/non-quoted contexts

The results of the analyses of quoted/non-quoted contexts can be interpreted as supporting the patterns of variation found between written and spoken data, even though the low number of instances in some of the examined text categories makes it impossible to draw anything but tentative conclusions. The expressions that are more frequent in spoken data (*'ll* and *going to*) are used more often in quotes. Medium-related factors can thus to some extent account for the choice of expression in the written texts. The expression *shall* is also frequently found in quotes but, as it is not more frequent in one medium than the other, explanations for the high proportion of quoted instances need to be looked for elsewhere. Stubbs suggests that "...the spelling *gonna* will only be found in stereotyped indications of casual speech, in, say, a novel" (1980:118). This corresponds well with the findings in my case studies where the instances of *gonna* in written text are almost exclusively found in quotes or in informal or speech-like contexts.

4.4 Summary

In this chapter and in the case studies referred to (Studies I, IV), it has been shown that there are substantial differences between the written and spoken corpora with regard to the use of expressions of future. The expressions are considerably more frequent in the spoken corpora, and the proportions of the different expressions also vary with medium. In the written corpora, *will* is the most frequent expression, and the difference between that expression and the others is considerable; no other expression is nearly as frequent. The *going to* expression is more frequent in the spoken corpora than in the written ones. It is, however, never the case that *going to* is used with a frequency close to that of *will.* The contracted form *'ll* is used more often in the spoken corpora, where its frequency is similar to that of *will.* The difference between the spoken and the written corpora is most noticeable for *'ll*. *Shall* is found to a somewhat higher extent in the written data, but that difference is evident only if earlier and later texts are compared separately. *Gonna* is almost exclusively found in the spoken material, primarily in the BNC and Sampler data. The combined proportion of *gonna+going to* is considerably larger in the spoken corpora than in the written ones.

In quoted contexts, *will* is found proportionately less frequently than the other expressions, and it is also used to a smaller extent in the spoken corpora than in the written. *Going to* is used in quotes more often than *will* but less often than *'ll.* I suggest that *'ll* is more markedly speech-related than *going to*. The expression *gonna* occurs but rarely in written texts, and then either in contexts that can be considered speech-related, (quotes and other speech-like passages), in very informal contexts, or where the distinction between spoken and written language is somewhat unclear (such as in e-mail messages).

5. Text category

5.1 Introduction

In my studies I have found that the distribution of the expressions of future varies between different categories of texts. The frequencies as well as the proportions of the expressions differ not only between corpora but also between different parts of one corpus. The variation within a corpus is often more substantial than the difference between two corpora of the same medium, written or spoken (see Studies I, II, IV, V, Chapters 4, 6, 7, 8). The present chapter describes this variation within corpora in more detail. The chapter consists of two parts: one dealing with written corpora and one with spoken. The results from my published studies (Studies I–V) are summarised and supplemented with further information from some additional case studies.

The corpora examined in my studies all consist of a number of smaller units of text, given labels such as categories, genres, domains, or hyper-categories. I use the term 'text category' as a generic term to represent all these units of text that are smaller than the corpora they are part of, as explained below and illustrated in Tables 5.1 and 5.2.

Table 5.1. *Text categories in the Brown, LOB, FLOB, and Kolhapur corpora*

Corpora	**Brown, LOB, FLOB, Kolhapur**														
Hyper-categories	Informative									Imaginative					
Genres	A Press: reportage	B Press: editorial	C Press: reviews	D Religion	E Skills, trades, hobbies	F Popular lore	G Belle lettres, biography	H Miscellaneous	J Learned	K General fiction	L Mystery and detective	M Science fiction	N Adventure and Western	P Romance and love story	R Humour

In the present context I use the term genre (following, for example, Biber 1988) to denote one of the 15 text categories in the Brown, LOB, FLOB, and Kolhapur corpora (henceforth also referred to as Brown-family). The genres are grouped into two hyper-categories: Informative and Imaginative (see Table 5.1). When the genres are mentioned in the text below, they are denoted by a letter and a short name, as listed in the table above. For a full description of the corpora, see Chapter 3.

Table 5.2. *Text categories in the BNC and Sampler corpora*

Corpora	**BNC and Sampler**														
Parts	Written										Spoken				
Components											Context-governed (CG)				Demographically Sampled (DS)
Domains	Imaginative	Natural science	Applied science	Social science	World Affairs	Commerce	Arts	Belief & thought	Leisure	Unclassified	Leisure	Business	Educational	Institutional	

Where the BNC/Sampler corpora are concerned, they contain two parts: written and spoken. The written part can be further subdivided into ten domains, listed in Table 5.2. The spoken part contains two spoken components: Context-governed (CG) and Demographically Sampled (DS). The Context-governed component consists of four domains.

5.2 Written corpora

5.2.1 Overall frequency of expressions of future

The frequencies of the expressions of future vary both between different corpora and between different text categories within the corpora (see Studies I, IV). Figure 5.1 illustrates the frequency (per 2,000 words) in the four comparable written corpora dealt with in my studies: LOB, FLOB, Brown, and Kolhapur. The figure presents the average relative frequency in each of the 15 genres (A–R), seen across the two hyper-categories in each corpus (Informative and Imaginative).[29] Raw and relative frequencies for genres and hyper-categories are given in Tables 5.3–5.6.[30]

[29] The lines have been added to facilitate the identification of the values for each hyper-category and do not suggest that the variation is continuous.

[30] Genre names in the tables are shorter forms of the text category labels given in the Brown Corpus Manual (Francis and Kucera 1979) – see Section 3.4.1 for full names.

As can be seen in the figure and tables, frequencies vary considerably between the individual genres in each hyper-category. The average frequency is higher in the Imaginative hyper-categories than in the Informative in three of the corpora, while the Brown corpus displays the reverse pattern, with more expressions of future in the Informative hyper-category.

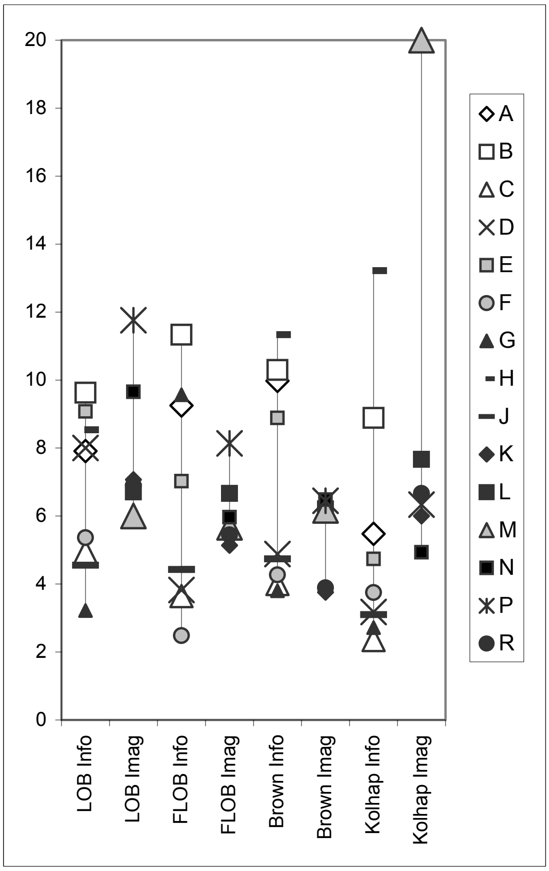


Figure 5.1. Expressions of future: frequency per 2,000 words in genres and hyper-categories in four comparable corpora (based on Tables 5.3–5.6)

Although one corpus or hyper-category as a whole may have a higher frequency of the expressions than another corpus or hyper-category, it is not necessarily the case that the average frequencies in all genres in that hyper-category or corpus are higher than in all genres in another. This indi-cates that the frequency of the expressions of future does not vary with hy-per-category: it is not always the case that Imaginative genres have more expressions of future than Informative ones, or vice versa. As will be further shown below, however, there is considerable difference between the hyper-categories regarding the proportion of the different expressions of future.

Although the frequency of the expressions of future does not appear to vary consistently with the hyper-category, it can be noted that in a number of cases the frequency order of the genres is similar in all the corpora. Genres that contain a proportionately large number of expressions of future in one corpus are often found to have a relatively high frequency of FUT in the other corpora as well. The same applies to genres with low frequencies of FUT. Genres C (Press: Reviews) and J (Learned and scientific writings), for example, contain low numbers of FUT in all corpora, while high frequencies are found in genres B (Press: Editorials) and H (Misc. Government and official documents). Therefore there does seem to be a relation between some of the genres and the frequency of the expressions of future. The Kolhapur corpus displays the most diverging pattern regarding the frequency of the expressions of future in the different genres. This may be due to differences between Indian English and the other national varieties, which have been described in Study I and which are also discussed in Chapter 7.

5.2.2 Overall proportions of expressions of future

As shown in Chapter 4 above, the proportions of the different expressions of future vary between the hyper-categories in the written corpora. I have examined this variation in Studies I and IV. It is found that not only do the overall proportions vary between the hyper-categories, but it is generally the case that all genres in one hyper-category differ distinctively from every genre in the other hyper-category with regard to the use of the expressions of future. This pattern is illustrated in Figure 5.2, based on Tables 5.3–5.6 and discussed below. The general patterns that are found in all four corpora are presented first, followed by a more detailed description of the variation within the hyper-categories in the two British English corpora. The latter part contains some hitherto unpublished results, while the first part summarises information presented in Studies I and IV.

Table 5.3. *Expressions of future in the Brown corpus. Raw and relative frequencies and proportions in the different text categories and hyper-categories*

BROWN	**Number of FUT**	***will***	***'ll***	***shall***	***going to***	***gonna***	**relative freq/2,000 words**
A Press: reportage	439	89%	8%	1%	2%	0%	10.0
B Press: editorial	278	86%	2%	7%	6%	0%	10.3
C Press: reviews	68	93%	1%	3%	3%	0%	4.0
D Religion	83	77%	1%	20%	1%	0%	4.9
E Skills and hobbies	320	86%	11%	2%	2%	0%	8.9
F Popular lore	205	82%	8%	6%	4%	0%	4.3
G Belle lettres,	285	83%	5%	11%	1%	0%	3.8
H Miscellaneous	340	71%	0%	29%	0%	0%	11.3
J Learned	379	88%	1%	11%	0%	0%	4.7
Informative	**2397**	**84%**	**4%**	**10%**	**2%**	**0%**	6.4
K General fiction	109	55%	30%	1%	13%	1%	3.8
L Mystery fiction	149	24%	56%	3%	17%	0%	6.2
M Science fiction	37	49%	30%	8%	14%	0%	6.2
N Adventure fiction	188	34%	48%	6%	9%	3%	6.5
P Romance	187	34%	\51%	2%	11%	2%	6.4
R Humour	35	57%	26%	6%	11%	0%	3.9
Imaginative	**705**	**37%**	**46%**	**4%**	**12%**	**2%**	5.6
TOTAL	3,102	73%	14%	8%	4%	0%	6.2

Table 5.4. *Expressions of future in the LOB corpus. Raw and relative frequencies and proportions in the different text categories and hyper-categories*

LOB	**Number of FUT**	***will***	***'ll***	***shall***	***going to***	***gonna***	**relative freq/2,000 words**
A Press: reportage	348	90%	3%	4%	3%	0%	7.9
B Press: editorial	260	92%	1%	3%	4%	0%	9.6
C Press: reviews	84	87%	5%	5%	4%	0%	4.9
D Religion	136	74%	4%	18%	4%	0%	8.0
E Skills and hobbies	345	91%	1%	4%	3%	0%	9.1
F Popular lore	236	84%	7%	3%	6%	0%	5.4
G Belle lettres,	248	83%	3%	10%	4%	0%	3.2
H Miscellaneous	256	61%	0%	37%	2%	0%	8.5
J Learned	364	82%	0%	16%	1%	0%	4.6
Informative	**2,277**	**83%**	**2%**	**11%**	**3%**	**0%**	**6.1**
K General fiction	205	47%	32%	13%	8%	0%	7.1
L Mystery fiction	161	32%	48%	7%	12%	0%	6.7
M Science fiction	36	56%	28%	3%	14%	0%	6.0
N Adventure fiction	280	29%	56%	5%	9%	1%	9.7
P Romance	341	38%	40%	12%	9%	0%	11.8
R Humour	62	74%	18%	5%	3%	0%	6.9
Imaginative	**1,085**	**39%**	**42%**	**9%**	**9%**	**0%**	**8.6**
TOTAL	3,362	69%	15%	11%	5%	0%	6.7

Table 5.5. *Expressions of future in the FLOB corpus. Raw and relative frequencies and proportions in the different text categories and hyper-categories*

FLOB	**Number of FUT**	***will***	***'ll***	***shall***	***going to***	***gonna***	**relative freq/2,000 words**
A Press: reportage	407	89%	5%	1%	5%	0%	9.3
B Press: editorial	306	85%	8%	5%	2%	0%	11.3
C Press: reviews	62	77%	0%	2%	21%	0%	3.6
D Religion	65	89%	2%	9%	0%	0%	3.8
E Skills and hobbies	267	89%	7%	1%	2%	0%	7.0
F Popular lore	348	93%	3%	2%	2%	0%	7.9
G Belle lettres,	191	82%	1%	14%	3%	0%	2.5
H Miscellaneous	287	84%	0%	15%	1%	0%	9.6
J Learned	354	86%	2%	11%	1%	0%	4.4
Informative	**2,287**	**87%**	**4%**	**6%**	**3%**	**0%**	**6.1**
K General fiction	149	44%	40%	4%	11%	0%	5.1
L Mystery fiction	160	36%	45%	7%	11%	1%	6.7
M Science fiction	34	41%	29%	12%	18%	0%	5.7
N Adventure fiction	173	39%	45%	2%	14%	0%	6.0
P Romance	236	43%	37%	13%	7%	0%	8.1
R Humour	49	59%	24%	2%	12%	2%	5.4
Imaginative	**801**	**42%**	**40%**	**7%**	**11%**	**0%**	**6.4**
TOTAL	3,088	75%	13%	6%	5%	0%	6.2

Table 5.6. *Expressions of future in the Kolhapur corpus. Raw and relative frequencies and proportions in the different text categories and hyper-categories*

Kolhapur	**Number of FUT**	***will***	***'ll***	***shall***	***going to***	***gonna***	**relative freq/2,000 words**
A Press: reportage	241	93%	0%	5%	2%	0%	5.5
B Press: editorial	240	95%	0%	2%	4%	0%	8.9
C Press: reviews	40	93%	0%	0%	8%	0%	2.4
D Religion	54	80%	0%	19%	2%	0%	3.2
E Skills and hobbies	180	92%	1%	5%	2%	0%	4.7
F Popular lore	165	94%	0%	4%	2%	0%	3.8
G Belle lettres,	190	83%	0%	15%	3%	0%	2.7
H Miscellaneous	489	64%	0%	33%	2%	0%	13.2
J Learned	248	82%	0%	17%	1%	0%	3.1
Informative	1,847	83%	0%	15%	2%	0%	4.9
K General fiction	348	57%	27%	10%	5%	0%	6.0
L Mystery fiction	184	53%	34%	9%	4%	0%	7.7
M Science fiction	40	48%	5%	45%	3%	0%	20.0
N Adventure fiction	74	54%	30%	15%	1%	0%	4.9
P Romance	114	58%	27%	6%	9%	0%	6.3
R Humour	60	58%	28%	5%	8%	0%	6.7
Imaginative	820	56%	28%	11%	5%	0%	6.5
TOTAL	2,667	74%	9%	14%	3%	0%	5.3

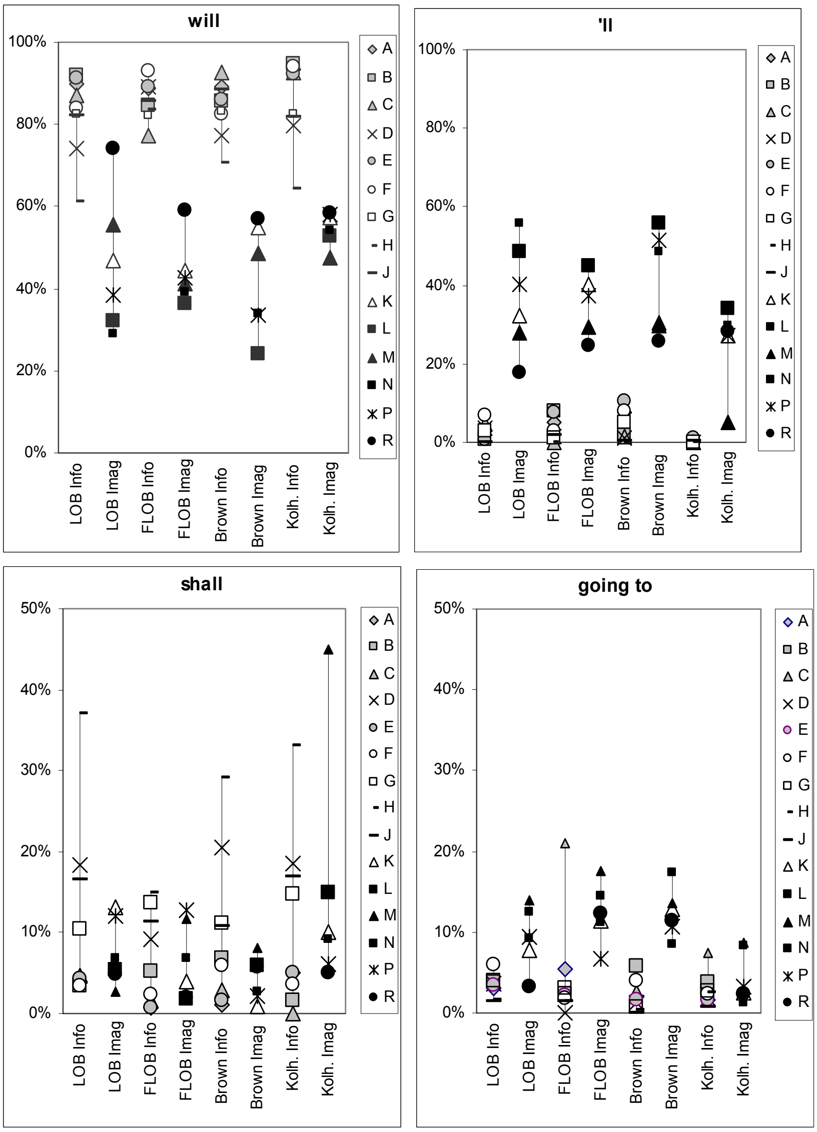


Figure 5.2. Patterns of variation across corpora and hyper-categories. Proportions of the expressions of future in 15 genres (based on Tables 5.3–5.6)

Figure 5.2 illustrates the proportions of each of the expressions in the different genres and hyper-categories in the Brown, LOB, FLOB, and Kolhapur corpora.[31] In the top left quarter of the figure, it can be seen that the proportions of *will* are considerably higher in the Informative hyper-category than in the Imaginative in all four corpora. It is even the case that all Informative genres have larger proportions of *will* than any Imaginative genre in any corpus (with the exception of genre R: Humour, in LOB). The large proportion of *will* thus appears to be a hyper-category feature, similar in all genres in a hyper-category.

The variation between the hyper-categories is equally prominent for the expression *'ll,* although the pattern is reversed. In the top right corner of Figure 5.2, it can be seen that the expression is used much more in the Imaginative genres than in the Informative. It can also be seen that the proportions of *'ll* and *will* are similar in the Imaginative hyper-categories, while *will* is much more frequent than *'ll* in the Informative hyper-categories. The proportion of *'ll* can be said to vary consistently between the hyper-categories, just as that of *will*. All imaginative genres (with the exception of genre M: Science fiction, in Kolhapur) contain a considerably larger proportion of *'ll* than any Informative genre in any corpus.

The pattern of distribution of *going to* is similar to that of *'ll,* although the difference between the two hyper-categories is not quite as clear (note that the scale in the figure is different with respect to *will* and *'ll*). The proportion of *going to* is larger in the Imaginative hyper-categories with respect to the Informative in all four corpora. The difference between Imaginative and Informative genres is, however, not as marked as for *will* and *'ll,* and in some cases there are Informative genres that contain larger proportions of *going to* than certain Imaginative ones and vice versa (for example C: Press: Reviews, in FLOB and R: Humour, in LOB). The difference between the Indian English Informative and Imaginative genres is particularly small, while there is a greater difference between the American hyper-categories (see Chapter 7 for a discussion of regional differences). Even though the pattern is not as convincing as for *will* and *'ll,* there is still good reason to claim that the proportion of the different expressions varies with hyper-category. The genres which deviate most from the general pattern all contain low frequencies of *going to.* With low frequencies of FUT, small variation in number can result in large proportional differences.

The three expressions *will, 'll* and *going to* are similar in the sense that there are marked differences in their use between the hyper-categories: *will* is used more in the Informative genres than in the Imaginative while *'ll* and *going to* are more frequent in the Imaginative genres than in the Informative in all corpora. The use of the expression *shall,* however, does not follow this pattern of variation. As illustrated in Figure 5.2, *shall* is used to a similar extent in most genres, Informative as well as Imaginative, with some notice-

[31] No diagram is provided for the expression *gonna* since its frequency is so low in these written corpora.

able exceptions. The expression is particularly frequent in certain genres, more exactly in genres H, D, J, and G.[32] It cannot be said that the use of *shall* is dependent on the text category in the same way as the other expressions of future, as its use does not vary with hyper-category or genre. The deviating pattern of *shall* has been examined in Study IV, and will also be commented on further below (Chapters 7, 8 and 9).

5.2.3 Comparison of hyper-categories

Chapter 4, as well as Studies I and IV, have shown that there is considerable variation in the use of expressions of future between the hyper-categories. To obtain a fuller picture of how the expressions are used in relation to text category, it is important to examine also the variation within the different categories of texts. Study IV briefly discusses the distribution of the expressions of future across the hyper-categories in two corpora: LOB and FLOB (Study IV:32–34). The aim of the present section is to examine in greater detail how the use of the expressions varies with text category in Present-day British English (see Study IV and Chapter 8 for a discussion of variation with time). The LOB and FLOB corpora are particularly suitable for this kind of comparison not only because they both contain samples of Present-day British English (1961–1991) but also because the genres to a certain degree can be expected to be similar with regard to the topics dealt with in the texts. Hundt and Mair (1999) declare that in the sampling of books and monographs for FLOB "...great care was taken to select works on equivalent topics [as those in LOB]". The distribution of the expressions of future in the two corpora can be found in Tables 5.4 and 5.5 above, illustrated in Figures 5.3 and 5.4.

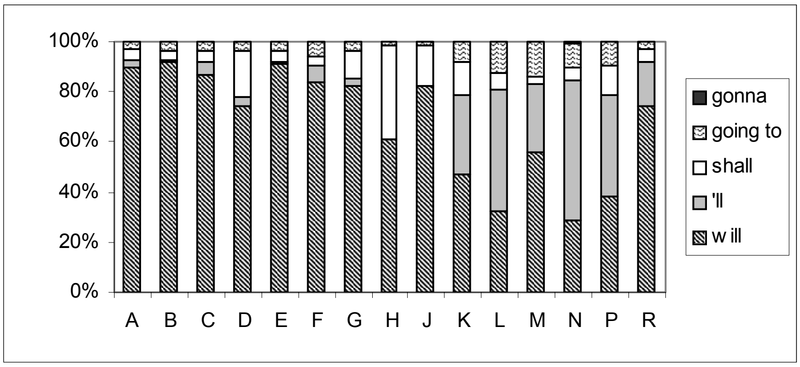


Figure 5.3. Distribution of expressions of future in the LOB corpus (proportions)

[32] H = Government and official documents, D = Religious, J= Learned, G= Belle Letters, Biography.

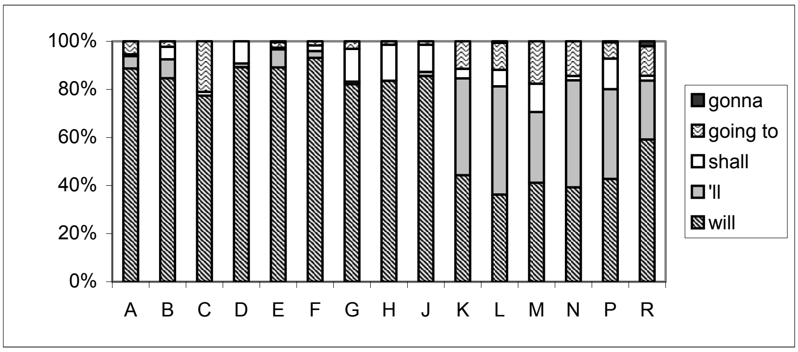


Figure 5.4. Distribution of expressions of future in the FLOB corpus (proportions)

5.2.3.1 Informative hyper-category (Genres A-J)

The Informative hyper-category is larger than the Imaginative, constituting about 75% of the text in each of the corpora (374 of the 500 texts). It contains a great variety of texts, such as press, academic writing and popular lore (see Chapter 3). The proportion of *will* is high overall: 83% in LOB and 87% in FLOB. The proportion of *going to* is very low, 3% in both corpora. The proportion of *'ll* is also low, 2% in LOB and 4% in FLOB, while s*hall* is used more, constituting 11% of the FUT in LOB and 6% in FLOB.

When the variation between the individual genres in the hyper-category is considered, it can be seen that the proportion of *will* varies considerably: from 61% (H: Miscellaneous) to 92% (B: Press: editorial) in LOB, and between 77% (C: Religion) and 93% (F: Popular lore) in FLOB. It is not possible to discern a general pattern of variation among the genres, or any consistent difference between the corpora. It was shown above that some genres with a high frequency of FUT in one corpus also have a high number of FUT in other corpora. The same does not apply where the proportions of the different expressions are concerned (with the noticeable exception of *shall,* as discussed further below). It is not the case that a genre with a high proportion of *will* in LOB also displays a high proportion of *will* in that genre in FLOB. However, the differences between the genres within the Informative hyper-category are small overall when compared to any genre in the Imaginative hyper-category, where the proportion of *will* is considerably lower in all cases.

The proportion of *'ll* is about twice as high in the Informative hyper-category in FLOB as in LOB, but in both cases very small and considerably smaller than the proportion of *will* in any of the corpora. The variation across the genres is noticeable, and the range is similar in the two corpora. In LOB the proportions vary from 0% to 7% and in FLOB from 0% to 8%. As the frequency of the expression in this hyper-category is so low (47 instances in

LOB and 84 in FLOB), the variation between the genres will not be given further consideration.

The expression *shall* is found more frequently in LOB than in FLOB in both hyper-categories. The overall difference between the corpora is considerable: the proportion of *shall* is 11% in LOB and only 6% in FLOB. The variation between the genres within the hyper-categories is also great: the proportions range from 3% to 37% in LOB and from 1% to 15% in FLOB. It is generally the case that the proportion of *shall* is larger in LOB than in the same genre in FLOB (B: Press: editorial and G: Belle lettres excepted). The variation within the hyper-category is noteworthy, but it differs from the variation for the other expressions in that it follows similar patterns in the two corpora. Four genres (D: Religion, G: Belle lettres, H: Miscellaneous, J: Learned) have high proportions of *shall* while five have considerably lower (the same pattern is also found in the Brown and Kolhapur corpora). While the use of the other expressions varies with hyper-category, *shall* is used markedly more in particular genres. When these genres are examined in more detail, it is apparent that *shall* is not evenly distributed across them, but is found in certain kinds of texts. In genre H: Miscellaneous, for example, the texts that contain large proportions of *shall* are texts related to law. This is not surprising. It is often claimed that *shall* is a formal expression primarily used in more formal texts, such as ones found in genre H: Miscellaneous. It is also claimed that the expression does not convey only future meaning but is strongly tainted with a sense of obligation (see Chapter 2). That can explain why *shall* is used frequently in legal texts (for further information on the use of shall in legal texts, see Williams 2005).

The expression *going to* is used to the same low extent in the Informative hyper-category in both British corpora: it comprises only 3% of the expressions of future. The variation between the different text categories is noteworthy, especially in FLOB where the proportion of *going to* varies between 0% (D: Religion) and 21% (C: Press: reviews). The variation in LOB only ranged between 1% (J: Learned) and 6% (F: Popular lore). The raw frequency of *going to* is very low. It exceeds 20 in only one genre (A: Press: reportage in FLOB). The absolute frequency is under 10 in 11 of the 18 genres. With such low frequencies, there is little to base any conclusions on considering the distribution across the genres. It can, however, be claimed with some confidence that *going to* is not used in British Present-day Informative prose to any great extent. The expression *gonna* only occurs once in the Informative hyper-category (in FLOB).

5.2.3.2 Imaginative hyper-category (Genres K-R)

That the proportions of the expressions of future differ between the Informative and Imaginative hyper-categories has already been shown above. This difference is very obvious where the use of *will* is concerned. In the Informative hyper-category the expression is unquestionably the most frequent. In the Imaginative hyper-category, however, the expression is used considerably less, its proportion being only about half that of *will* in the Informative hyper-category.

The expression *'ll* is used rarely in the Informative hyper-category but is found as frequently as *will* in the Imaginative genres. The average proportions of *'ll* are similar in the LOB and FLOB corpora, but the variation within the Imaginative hyper-category is substantial. In FLOB, the *'ll* expression constitutes between 24% (R: Humour) and 45% (L: Mystery, N: Adventure) of the expressions of future, and the variation in LOB is even greater, between 18% (R: Humour) and 56% (P: Romance). Despite the wide range of proportions, there is no genre in the Informative hyper-category that contains a proportion of *'ll* close to that in any of the Imaginative genres. This is a fundamental difference between the hyper-categories. It is, however, not a surprising result. It has been shown (for example Kjellmer 1998), and commented on above (Chapter 4) that contracted forms are more frequent in texts where there are greater proportions of spoken discourse, such as in fiction texts which contain dialogues or other speech-reflecting components. An interesting factor relating to the distribution of *'ll* is that a genre with a small proportion of *'ll* in LOB to a large extent also has a small proportion in FLOB, while the genres with large proportions of *'ll* in LOB also have a relatively large proportion of the contracted form in FLOB. The same pattern is not found with the other expressions (except for *shall* in the Informative hyper-category), which possibly suggests that *'ll* is more typical of particular genres than *will* and *going to*.

As described above, *shall* differs from the other expressions in that there is no clear difference between the hyper-categories in the extent to which the expression is used. In LOB, the expression is found more in the Informative hyper-category, but in FLOB the reverse is the case: *shall* constitutes 7% of the expressions of future in the Imaginative hyper-category in FLOB, while the proportion in LOB is 9%. In genre P (Romance) the proportion is equally large in both corpora, 13%. One explanation for the relatively large proportion in this category lies in the content of the texts, as shown in Study IV:53. The texts in this genre that contain high proportions of *shall* are often stories which are set in historical settings, and the instances of *shall* can be seen as attempts to mirror older language. The following example (1), also given in Study IV, illustrates the point:

(1) Good. While I am endeavouring to take some kind of bath, you can remove from my baggage those things I *shall* need here. I *shall* send the remainder back to Cairo on the next steamer. (FLOB P02 129–132)

In the Informative hyper-category, *shall* is used in a few texts in four genres in particular, and is very rare in the other genres. The same pattern is also found to some extent in the Imaginative hyper-category: the instances of *shall* occur in a limited set of texts, such as legal texts in genre H: Miscellaneous. To some degree this is an effect of the overall low frequency of the expression. In the FLOB corpus, for example, there are only 55 instances of *shall* in the 126 Imaginative texts, i.e., an average of 0.44 instances per text. It is interesting to note then that a number of texts contain a relatively high number of occurrences. One text in genre P: Romance, for example, has ten instances, and in genre M: Science fiction all four instances are found in the same text. As the frequency of the expression is very low it is difficult to evaluate the patterns of variation. It is clear, however, that the use of *shall* does not vary with hyper-category the way the other expressions do, but that the expression is used primarily in certain genres, or even in particular texts in those genres. An explanation to this variation can be found when the linguistic association patterns are examined (see Chapter 9).

As shown above, *going to* is used more in the Imaginative genres than in the Informative ones, even if the variation is not as marked as for *will* and *'ll*. The overall proportion in the Informative hyper-categories is 3% in both LOB and FLOB. In the Imaginative genres, *going to* is used in 9% and 11% of all cases (LOB and FLOB respectively). The proportions in the Imaginative genres range from 3% to 14% in LOB and between 7% and 18% in FLOB. Only one genre (C: Press: Reviews in FLOB) has a proportion of *going to* as large as any of the Imaginative genres. This is an indication that the use of *going to* varies with hyper-category, as does that of *will* and *'ll.* The frequency of the expression is low also in the Imaginative hyper-category (100 occurrences in LOB, 88 in FLOB), which makes all figures uncertain. One possible explanation for the higher frequency of the expression in the Imaginative hyper-category is provided in Chapter 4 above. The *going to* expression is more frequent in spoken language and is also used more in quotes. When the occurrences of *going to* in LOB and FLOB are examined, it is apparent that the proportion of instances found outside quotes is smaller in the Imaginative hyper-category than in the Informative. There are, however, also a number of non-quoted instances of *going to* in both hyper-categories, which indicates that the expression is not exclusively speech-related the way *gonna* and, to a considerable extent, *'ll* are.

The variant *gonna* is only found four times in the Imaginative hyper-category in FLOB and twice in LOB. It is therefore not possible to examine patterns of usage for this expression in written Present-day English on the basis of these data: it can only be concluded that the expression is very rare. To see if anything more can be discovered about the use of this expression in written Present-day English, the larger BNC has been consulted. The results are presented in the following section. As in Chapter 4, all instances of the expression were examined (including occurrences with past tense reference). Information about the distribution of *going to* has been included for comparison.

5.2.4 *Gonna* and *going to* across text categories in the BNC

The distribution of the 546 instances of *gonna* and the almost 19,000 instances of *going to* varies greatly across the different text categories (domains) in the BNC, as illustrated in Table 5.7 (for a presentation of the BNC, see Chapter 3).[33]

Table 5.7. Gonna *and* going to *in the written part of the BNC*

Domain	**No. of words**	***gonna***		***going to***	
		Raw frequency	**Frequency pmw**	**Raw frequency**	**Frequency pmw**
Imaginative	19,664,309	322	16.4	11,347	577
Arts	7,014,792	131	18.6	1,407	200.5
Leisure	8,991,792	41	4.5	1,775	197,4
Belief and thought	3,035,896	2	0.6	322	106
Commerce and finance	6,668,357	0	0	697	104.5
Social science	12,186,378	30	2.4	1,253	102.8
Applied science	7,341,375	8	1	731	99.5
World affairs	15,243,341	8	0.5	1,297	85
Natural and pure science	3,746,901	0	0	137	36.5
Total	83,893,089	542	6.4	18,966	226

Gonna is found primarily in the Arts and Imaginative domains, and it occasionally occurs in other text categories (such as Leisure and Social science) but is not found at all in Natural and pure science or Commerce and finance. *Going to* is much more frequent in the Imaginative domain than in others, and least frequent in the Natural and pure science texts, with only 36.5 occurrences per million words. The *going to* expression occurs in Commerce and finance, but *gonna* does not. It can, nevertheless, be said that, overall, the two expressions seem to display similar patterns in their distribution across the text categories, being more frequent in the 'Imaginative' and 'Arts' texts and particularly rare in the 'Natural and pure science' domain.

Figure 5.5 shows the size of the different domains compared to the proportions of *gonna* and *going to* in the corpus (how large a proportion of the corpus text each domain comprises and what proportions of all instances of *gonna* and *going to* in the corpus are found in the different domains).

[33] All instances irrespective of the tense of the auxiliary were included. The data were retrieved by using the original index provided with the first release of the corpus. There are 548 instances of *gonna* and 20,069 of *going to* (where *to* is tagged as an infinitival marker) in the written part of the BNC. The table only contains occurrences found in texts marked for text type, which is why the total number of instances differs from that obtained for the whole written corpus.

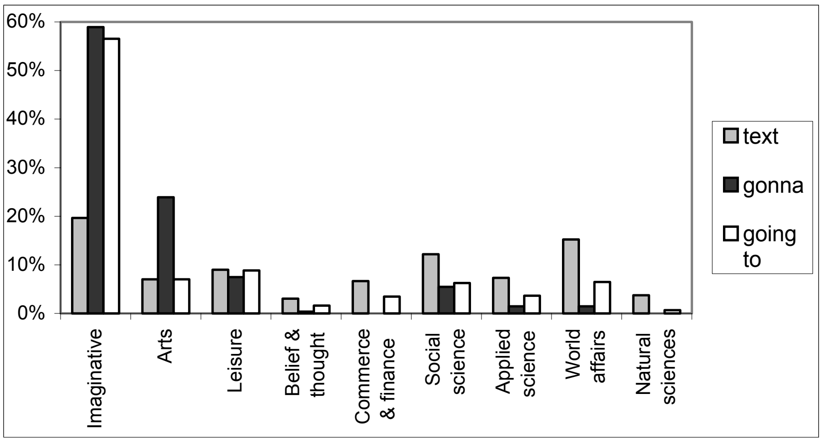


Figure 5.5. Domain size (text) compared to proportions of *gonna* and *going to* in the written part of the BNC

The figure shows that the Imaginative domain is the largest as far as the proportion of text is concerned (about 23% of the written texts defined for domain), followed by World affairs (18%) and Social science (15%). The distribution of the two expressions is very uneven across the text categories. The majority of the instances (around 60%) are found in the Imaginative domain. The proportion of Imaginative *gonna* and *going to* is thus considerably larger than the proportion of Imaginative text: the expressions can be said to be over-represented in the Imaginative domain. In the Leisure domain, the proportions of both expressions are roughly equal to the proportion of text. As far as *going to* is concerned, the same relationship is found in the Arts domain. The proportion of *gonna* is, however, unexpectedly large in this domain: almost every fourth instance of *gonna* is found in the Arts domain which constitutes just over 8 % of the text in the corpus.

The explanation can be found by inspecting the occurrences of *gonna* in the Arts texts. To a considerable extent these instances are used in texts related to (modern) music. They are found in magazines about music, in quotes from songs or in titles of songs or recordings, as exemplified in (2):

(2) One moment Harvey sounds like she's trapped in a particularly unpleas ant psychodrama, the next she's singing an exuberant snatch of I'm *Gonna* Wash that Man Right Out of My Hair. (BNC AJN:150)

The relatively large proportion of *gonna* in this domain cannot, consequently, be explained as a feature of Arts texts in general but as something more specific to certain, individual texts within that domain. In this respect the use of *gonna* is similar to that of *shall*, which was also found to be used primarily in certain texts (see Section 5.2.3.1).

About 4.5% of the text in the written part of the BNC is found in the domain Natural and pure science. There are, however, no instances of *gonna* and only a handful occurrences of *going to* in this domain. The expressions are also relatively rare in the other science domains (Applied science and Social science) as well as in the domains Belief and thought and Commerce and finance.

A number of factors make it difficult to evaluate possible similarities and differences between the BNC domains and the genres and hyper-categories in the Brown, LOB, FLOB, and Kolhapur corpora. The sampling frames are different, the text sizes differ and the text categories do not correspond. Consider, for example, the BNC Imaginative domain, which contains different kinds of Imaginative writing. In that respect it corresponds most closely to the Imaginative hyper-category in the Brown-family corpora, which comprises six genres. The Belief and thought domain in the BNC, however, contains texts which can probably be considered most closely related to one of the genres (D: Religion) in the Brown-family corpora. No exhaustive attempt will be made to compare the BNC to the Brown, LOB, FLOB, and Kolhapur corpora with regard to the distribution of the expressions of future across the text categories (domains in the BNC, genres and hyper-categories in Brown-family). It can merely be pointed out that the higher frequency of *going to* in the Imaginative domain in the BNC corresponds to the distribution in the other corpora, where the expression is more frequent in the Imaginative hyper-category. It is also worth repeating that the *going to* expression was found to be less frequent in the Informative hyper-category than in the Imaginative in the Brown-family corpora, which can be compared to the relatively small proportions of both *gonna* and *going to* in the non-imaginative domains in the BNC.

5.2.5 Summary: variation in written corpora

This part of the chapter has shown that the use of the expressions of future varies across the different text categories in the written corpora. The frequency of the expressions does not seem to vary consistently with text category, and even though some of the genres in the Brown-family corpora have more expressions of future there is no overall pattern of variation between the genres or hyper-categories. The most prominent variation is found when the proportions of the different expressions are compared between the hyper-categories. *Will* is used to a much higher extent in the Informative hyper-category, while *'ll* and *going to* are used more in the Imaginative hyper-category. The use of *shall* is quite different from that of the other expressions, as the expression is found primarily in a few genres and particularly in a limited number of texts. *Gonna* is very rare overall. When the expression is found in the BNC, it is generally in Imaginative texts or texts related to music in the Arts domain.

5.3 Spoken corpora

In my studies, I have examined the variation between different kinds of spoken text in relation to the use of two expressions in particular: *going to* and *gonna.* These studies are based on data drawn from the spoken part of the BNC (Studies II, III, V). I have been able to conclude that *going to* is the more frequent expression in more formal types of text, while *gonna* is used more in the informal settings. Considering the descriptions of the expressions that are given in standard reference works, it is not surprising that *gonna* is more frequent in informal contexts. That *gonna* is so very frequent, constituting up to 75% of the combined frequency of *gonna+going to,* is not suggested in these reference works.

One issue that cannot be overlooked when the variants *gonna* and *going to* are compared is the question of transcription. All conclusions about the distribution of the two expressions (and to some degree also of other pronunciation variants such as *will/'ll*) should be considered with an awareness that the results can be affected by transcription practices. This has been discussed in Chapter 3, and will not be dealt with further here (see also Study II and Krug 2000:38).

5.3.1 Expressions of future in the CG and DS components

The distribution of all the expressions in different kinds of spoken text has been examined in Study IV, where the data is drawn from the LLC and the spoken part of the BNC Sampler. The main findings from that comparison are summarised below. To gain a fuller understanding of the variation within the spoken corpora, my previous results are supplemented with information gathered in an additional case study. The new study concerns the distribution of the expressions across monologue/dialogue and across domains in the Context-governed component of the spoken part of the Sampler.

In what follows, the discussion will revolve around the distribution of the expressions of future across the data in the spoken Sampler corpus. The spoken part of the Sampler consists of two components of roughly the same size (just under 500,000 words): the Context-governed component (CG) and the Demographically Sampled component (DS). The DS component has been described as a "component of informal encounters" while the CG component is referred to as a "component of more formal encounters" (Aston and Burnard 1998:31).[34]

[34] The quote refers to the whole BNC from where the texts for the Sampler have been drawn.

As described in Study IV, the frequency and proportions of the expressions of future vary between the two spoken components. The frequency of the expressions is considerably higher in the less formal, DS component: 9,392 occurrences to be compared to 6,795 in the CG component of equal size. The proportions of the expressions also vary, as illustrated in Figure 5.6 (raw frequencies can be found in Table A.7, Appendix B).

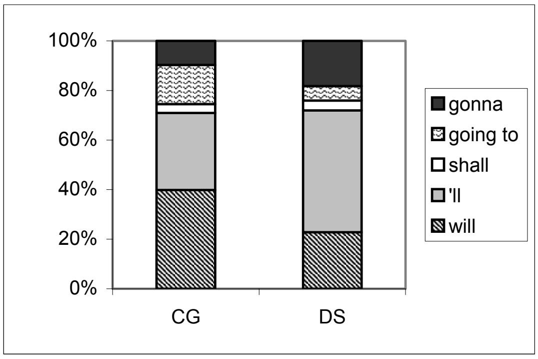


Figure 5.6. Expressions of future in the Demographically Sampled (DS) and Context-governed (CG) components of the BNC Sampler (proportions)

In the CG component, *will* has the highest frequency (40%), followed by *'ll* (31%). The reversed pattern is found in the DS component where *'ll* is more than twice as frequent as *will* (49% and 23% respectively). The combined proportion of the two expressions is, however, almost the same in both spoken components, 71% and 72%. The proportion of *shall* is the same in both components, only 4%. The proportion of *going to* is significantly larger in the more formal CG component (15%) than in the DS one (6%), while the situation is the reversed for *gonna,* which is found to a larger extent in the DS component (18%). The combined proportion of the *gonna*+*going to* expression is similar in the two components: 25% in CG and 24% in DS. Although the differences between the components are statistically significant for all expressions but *shall,* the most noticeable difference between the spoken components seems to be that the proportions of the contracted forms *'ll* and *gonna* are larger and the proportions of the full forms smaller in the less formal DS component. This is hardly surprising, considering that contractions are often avoided in more formal contexts.

5.3.1.1 Dialogue vs. monologue

As far as the use of the full vs. contracted forms is concerned, the variation between the spoken components mirrors the pattern found in the written corpora. The expression *'ll* is more frequent in the written Imaginative hyper-categories and the spoken DS component while the full form *will* is used proportionately more in the Informative hyper-categories and the CG component. In the spoken corpora the difference between the two text categories cannot be explained with differences in the proportion of speech-like text. Instead, the explanation must be looked for in other factors. One such factor is the situation where the utterance was made, whether it was used in a monologue or dialogue.

Figure 5.7 illustrates the proportion of the expressions in the monologue and dialogue texts in the CG component and in the DS (dialogue) texts[35].

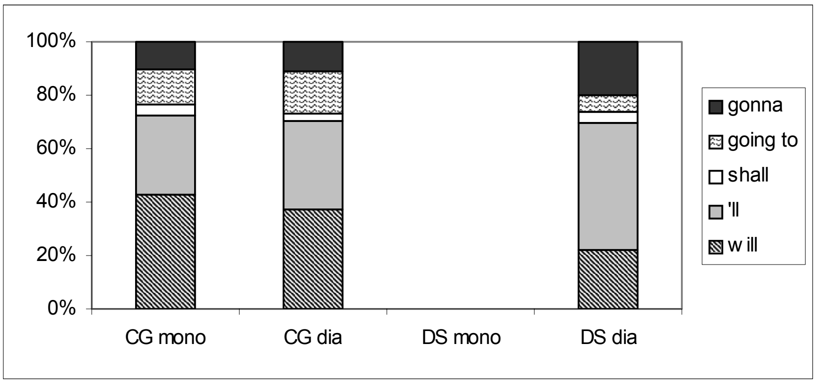


Figure 5.7. Proportions of expressions of future in dialogue (dia) and monologue (mono) texts in the two spoken components of the BNC Sampler (no monologue texts in DS component)

As can be seen in the figure, there is some variation in the distribution of the expressions of future between the monologue and dialogue texts in the CG component (as the DS component does not include monologues no comparison can be made within that component). The proportions of *will* and *going to* do not differ significantly between the monologue and dialogue texts, while those of *'ll, shall,* and *gonna* do. When the dialogue texts in the two components are compared, statistically significant differences are found for all expressions but *shall.* It can thus be concluded that the variation between the CG and DS components is more substantial than the variation between monologue and dialogue as far as the use of expressions of future is concerned.

[35] The DS component consists of conversations and only contains dialogues, while the CG texts are of a more varied origin and also include monologues. There are, nevertheless, 520 instances of expressions of future marked as monologue in the DS component, compared to 39,236 instances marked as dialogue.

5.3.1.2 Domains

The variation between the two components that can be related to speaker characteristics has been examined in Studies II and V, as described in Chapter 6. What has not been presented in my published studies is how the distribution of the expressions of future varies across the different kinds of texts found in the Context-governed component. The texts for the CG component were sampled according to the context where they were found, defined as four domains: Business, Educational/Informative, Leisure, and Institutional/Public. The frequency and proportion of the expressions of future in these four CG domains can be found in Table 5.8, illustrated in Figure 5.8 (raw frequencies given in Table A.8 in Appendix B).

Table 5.8. *Expressions of future in CG domains. Frequency (absolute/per 2,000 words) and proportions. > indicates statistically higher proportion than other domain*

DOMAIN	**Frequency (raw / per 2,000 words)**	***will***	***'ll***	***shall***	***going to***	***gonna***
Business (B)	11,543 / 17.6	37% >E,L	33% >I	2%	14%	14% >E,L,I
Educational/Informative (E)	7,253 / 14.1	29%	38% >B,I	2%	18% >B,I	12% >L,I
Leisure (L)	5,333 / 7.8	31%	38% >B,I	4% >E,B	17% >B,I	10% >I
Institutional/Public (I)	9,609 / 14.5	52% >E,B,L	24%	4% >E,B	14%	6%
Total	33,738 / 13.4	39%	32%	3%	15%	11%

One obvious difference between the domains is that the Leisure texts contain fewer expressions of future than the other domains, only 7.8 FUT per 2,000 words, compared to 14.1, 14.5 and 17.6/2,000 words in the other three domains.

The proportions of the expressions also vary between the domains. The Institutional/Public domain differs from the other in several ways. It has high proportions of *will* and *shall* and low proportions of *going to, gonna* and *'ll*. The Business domain has the same proportion of *going to* as the Institutional domain, but does not display the differences between the full forms *will* and *going to* and the contracted forms *'ll* and *gonna* that are found in the Institutional domain. The proportion of *going to*+*gonna* is similar in the Business, Education and Leisure domains, between 27% and 30%, while the proportion in the Institutional domain is only 20%. The proportion of *gonna* is particularly small in the Institutional domain, only 6%, compared to 10 –14% in the other domains.

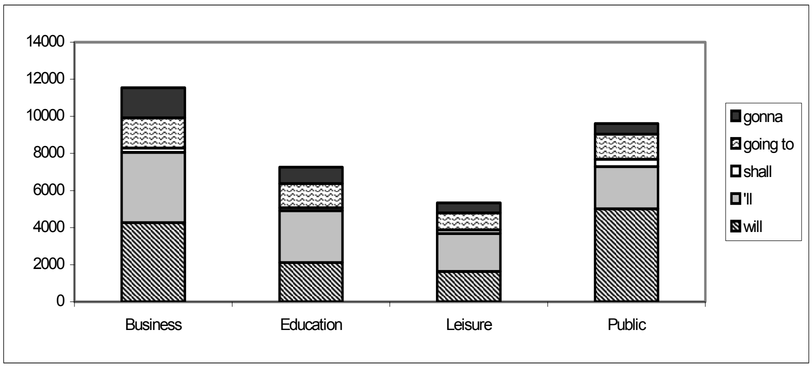


Figure 5.8. Expressions of future in CG domains (frequency per 2,000 words)

The Educational and Leisure domains display relatively similar patterns: both of them have more *'ll* than *will* and more *going to* than *gonna.* The proportion of *shall* is higher in the Leisure domain than in the Educational, 4%. This is the same proportion as in the Institutional domain. A closer look at the instances of *shall* within the Leisure domain reveals that a large proportion of these (about 30%) are from two rather special texts, recorded at Christie's auction rooms, where the language is very different from what is generally found in the corpus, as illustrated in (3):[36]

(3) Twenty five twenty five pounds, any more at twenty five and I *shall* sell at twenty five pounds if there's no further bid any more at twenty five pounds? (BNC HUR 567)

Within the Institutional domain, the expression *shall* is found in a number of different texts, but usually there are only a few instances in each text. Among the texts with higher frequency of *shall* are recordings from debates in the House of Commons and the House of Lords, county council debates and a sermon. This suggests that the pattern observed in the written material, where *shall* was particularly frequent in certain kinds of texts, can also be said to exist in the spoken data.

5.3.2 Summary: variation in spoken corpora

This study of the distribution of the expressions of future across the spoken components of the Sampler has revealed that there are differences between the different kinds of text. The most noticeable differences were found when the DS and CG components were compared. The expressions of future are

[36] See also Chapter 9.

more frequent in the DS component. The proportions of the expressions also vary between the two components. In the DS component, *'ll* and *gonna* are used proportionately more*,* while *going to* and *will* are more frequent in the CG component. There are also some differences between the domains in the CG component. It is interesting to note that the two contracted, and presumably more informal, expressions *'ll* and *gonna* do not follow the same pattern of variation. Possible explanations as to why the expressions are used differently are suggested in relation to the study of the variation with linguistic factors (see Chapter 9).

The CG component contains both monologue and dialogue texts. No significant difference could be found between the two situations regarding the proportions of the expressions. The amount of expressions of future found in monologue texts is limited and restricted to the CG component, which makes it impossible to draw any far-reaching conclusions about this particular variation.

One potential problem with comparing the two spoken components is that they are different in a number of ways. They are said to be components of informal and formal encounters respectively, but there are also a number of other features of the texts that differ between them and thus make them difficult to compare. As discussed in Studies II and V, these are differences in, for example, the kind of speakers sampled, as well as differences in the amount of text from different kinds of situations. These features make it difficult, if not impossible, to draw firm conclusions about the extent to which one particular factor affects the distribution of the expressions, and no attempt will be made to base such conclusions on these data. One central finding, however, that cannot be ignored when analysing the spoken data, is that the two components differ considerably with regard to the use of expressions of future (see also Studies II, V and Chapter 6).

5.4 Summary

This chapter has presented the results of studies concerning the use of expressions of future in various text categories in written and spoken material. It has been shown that the use of the expressions varies between different kinds of text: between the Informative and Imaginative hyper-categories in the written Brown, LOB, FLOB, and Kolhapur corpora, and between the DS and CG components in the spoken Sampler corpus. The differences are primarily related to proportions: the frequencies of the expressions do not vary to the same extent. As far as the two spoken components of the Sampler corpus are concerned, however, the difference is one of frequencies as well as proportions.

There is considerable variation between the text categories within the different corpora. The main pattern identified is that *will* is the most frequent expression overall in the written corpora. In the Imaginative hyper-category and in the spoken corpora, the expression *'ll* is as frequent, occasionally more frequent than *will.*

The variation between the spoken text categories is noteworthy, in particular with respect to the proportion of *gonna*. The combined proportion of *gonna+going to* is similar in the formal and informal text categories (CG and DS components), but *gonna* is used more in the informal, DS component. The distribution of the expressions of future varies substantially between the CG domains but there is little variation between the monologue and dialogue texts within the CG component. The most distinct variation found within the spoken part of the BNC Sampler is that between the CG and DS components.

6. Speaker properties

6.1 Introduction

Chapter 4 showed that the use of the expressions of future varies with medium: the expressions are used to different extents in written and spoken texts. It has also been shown that there is considerable variation between different text categories in the spoken and written corpora (Chapter 5). The present chapter examines the variation within the spoken corpora by looking at how the expressions are used by different kinds of speakers.

Socio-linguistic studies have shown that language use varies with a number of features related to the speaker, such as age, sex, and social class. In my studies, I have focused on how the use of *gonna* and *going to* varies with certain speaker-related factors (Studies II, V). My results will be summarised and supplemented here with further analyses of such patterns of variation across factors related to the speaker. All data are drawn from the spoken part of the BNC, which contains about 10 million words: roughly six million words of more formal conversations comprising the Context-governed component (CG) and four million words in the Demographically Sampled component of spontaneous conversation (DS). For further information about the spoken data, see Chapter 3 and the discussions in Chapters 4 and 5.

It has been shown (Chapter 5, Studies II, V) that the distribution of expressions of future varies considerably between the two components of the spoken part of the BNC. It was suggested in Chapter 5 that to some extent this variation might be the result of differences between the components in the kind of speakers that were recorded. For this reason the data for the two components are presented separately here.

6.2 Speakers' sex

Differences between male and female speakers have been examined in many sociolinguistic studies. It has been found that there are at times considerable differences between the sexes with regard to lexis and grammar as well as pronunciation. It is claimed, for example, that women are more sensitive to differences between standard and non-standard language and tend to use more standard forms than men, especially in more formal contexts (for example Chambers 2003)

Figure 6.1 (based on Table 6.1 below) illustrates the proportions in the spoken part of the BNC of the expressions of future uttered by speakers whose sex is known.

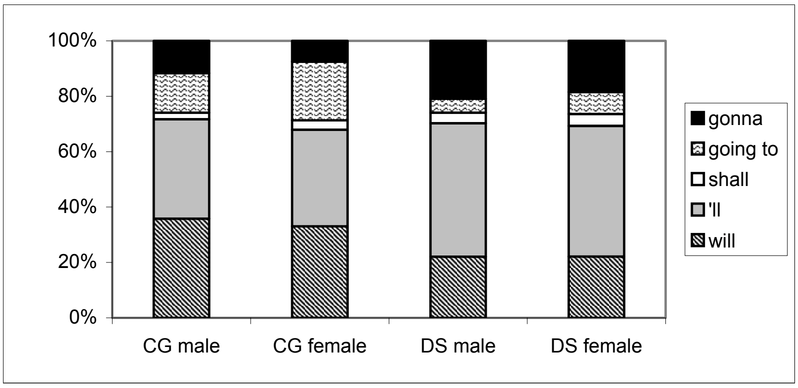


Figure 6.1. Expressions of future. Distribution (proportion) across male and female speakers in the Context-governed (CG) and Demographically Sampled (DS) components.

The figure shows that in the CG part of the BNC, women and men display patterns of usage that seem fairly similar. About one third of the FUT are *will,* and *'ll* is used to almost the same extent. Men appear to use *will* slightly more than women. The proportion of *shall* is very small but slightly larger for the female speakers than the male in this component of the corpus. The most noticeable difference between the sexes in this part of the corpus is that the women use a smaller proportion of *gonna* and a larger proportion of *going to* than the men.

In the DS part, the patterns of usage also appear to be similar for women and men. The proportion of *'ll* is more than twice as high as that of *will* for both sexes, which constitutes a difference between the CG and DS parts (see also Chapter 5). The difference between the male and female speakers is, however, very slight. The use of *shall* does not vary between the sexes either, whereas proportions of *going to* and *gonna* do appear to vary. As in the CG component, women use a larger proportion of *going to* than men, and a smaller proportion of *gonna.*

Table 6.1 presents the raw frequencies of the FUT and the proportions of each expression, and statistically significant differences between the sexes are marked (for information about the statistical significance test, see Section 3.5.4).

Table 6.1. *Expressions of future. Distribution across speakers' sex in the two spoken components of the BNC. Raw frequencies and proportions. + = significantly higher values for either sex within the component.*

	Raw frequency	*will*	*'ll*	*shall*	*going to*	*gonna*
Context-governed (CG)						
male	*18380*	*36% +*	*36%*	*2%*	*14%*	*12% +*
female	*4564*	*33%*	*35%*	*3% +*	*21% +*	*8%*
Demographically Sampled (DS)						
male	*13597*	*22%*	*48%*	*4%*	*5%*	*21% +*
female	*22070*	*22%*	*47%*	*4%*	*8% +*	*18%*

As the table shows, the differences between the sexes are greater in the CG component than in the DS. In CG, all expressions but *'ll* are used in ways that differ significantly in statistical terms. In the DS component only the proportions of *gonna* and *going to* are used to a significantly different extent by the male and female speakers. The only difference that can be found between the male and female speakers in both components is that women use a smaller proportion of *gonna*. Assuming with a number of authors that *gonna* is a non-standard form of *going to,* the current finding fits in well with theories about the sociolinguistic gender pattern (see, for example, Fasold 1990, Labov 1966). These theories state that, especially in more formal contexts, women use forms that are considered 'incorrect' less than men, which seems to be the case with *gonna* and *going to* here.

6.3 Speakers' age

In addition to sex, sociolinguistic studies have pointed to age as one influential factor accounting for variation in language use. A number of the speakers in the BNC are coded for age, given as one of six age groups (for description and discussion of this kind of information, see Study II). In Study V, it was shown that the *gonna* expression is used significantly more by younger speakers in both components. The distribution of all five expressions of future across speakers' age is given in Table 6.2.[37] In the table, proportions that are significantly higher (+) or lower (-) than those in the next higher age group have been marked (for a discussion of statistical significance, see Section 3.5.4).

[37] Note that the span of the age groups is different. The youngest and second oldest groups span over 15 years, while the oldest is open-ended. The remaining age groups cover ten years each.

Table 6.2. *Expressions of future. Distribution across speakers' age in the two spoken components of the BNC (raw frequencies and proportions). + / - = proportion significantly higher/lower than in the higher age group.*

Context-governed (CG)						
AGE	***Raw***	***%***				
	frequency	***will***	***'ll***	***shall***	***going to***	***gonna***
-14	*182*	20	54	3	5 -	18
15-	*737*	27	44 +	3	14	13
25-	*2571*	32	36	3 +	17	11 +
35-	*2171*	35 +	37	2	17	9
45-	*7033*	32 +	38 -	2 -	18 +	10 +
60-	*1354*	28	47	4	15	6
sum	*14048*	32	39	2	17	10
Demographically Sampled (DS)						
AGE	***Raw***	***%***				
	frequency	***will***	***'ll***	***shall***	***going to***	***gonna***
-14	*4003*	22	41 -	5 +	6 +	26
15-	*4221*	20	45 -	3	4 -	27 +
25-	*7441*	21	49	4	5 -	21
35-	*7206*	22	48	3	6	19 +
45-	*6559*	24	50 +	4	7 -	15 +
60-	*5472*	24	48	4	11	12
sum	*34902*	22	48	4	7	19

As illustrated in Table 6.2, no overall trend pointing to variation with speakers' age in the choice of expression of future can be found in the spoken BNC data. There are some proportions that are significantly different from those in the higher age group, but on the whole the trends are rather weak. This is illustrated in Figures 6.2 and 6.3 (CG and DS components separately).

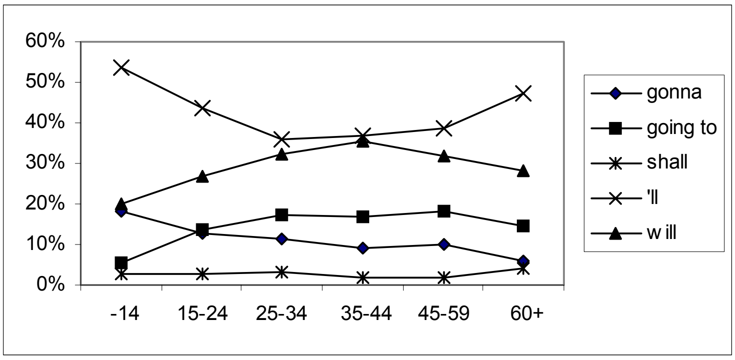


Figure 6.2. Expressions of future in the CG component of the BNC. Proportions in different age groups

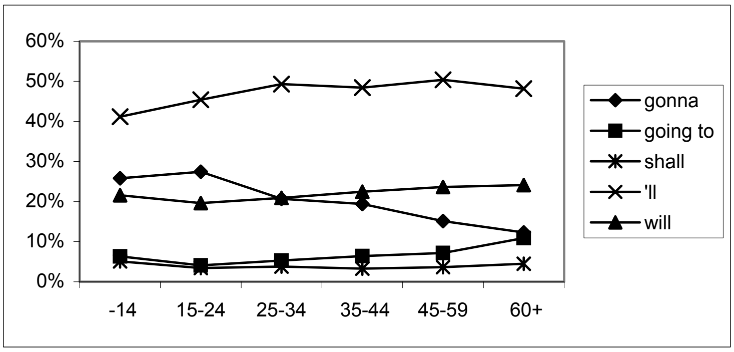


Figure 6.3. Expressions of future in the DS component of the BNC. Proportions in different age groups

Figure 6.3 illustrates how the proportions of *will, 'll,* and *shall* are more or less the same for all age groups in the DS component. The most noticeable difference is that the proportion of *gonna* decreases with age so that the older the speaker, the smaller the proportion of *gonna.* This decrease is to some extent compensated for by an increase in the use in *going to*, the proportion of which is larger for the older speakers.

The proportions of the expressions of future vary more in the CG part of the corpus. As Figures 6.2 and 6.3 show, the most diverging pattern is that found for the youngest age group. It should be noted in this context that this age group is represented by only about 1% of the expressions of future. The second youngest age group is also comparatively small: only about 5% of the expressions of future are found in that category. The proportions for the youngest speakers should thus be interpreted with some caution. As far as the other, larger groups are concerned, the most noticeable pattern is that the proportion of *will* is smaller and that of *'ll* larger for the two oldest age groups than for the younger ones. There is no apparent difference in the proportion of *going to.* The proportion of *gonna* is small overall (although larger than that of *shall*), and seems to be smaller the older the speakers are.

On the whole, the expression *shall* displays the most stable pattern, with similarly low frequencies overall (2–4% for all speaker groups but one). As *shall* has been referred to as a more formal expression, and has been shown to occur more in the earlier LOB corpus than in the later FLOB (Study IV, Chapter 8), it might have been expected that the expression would be used more by older speakers than by the speakers in the younger age groups. This, however, is not the case. The use of the expression *gonna* could also be expected to vary with speakers' age. Claims that the expression is becoming more frequent over time cannot be confirmed or refuted in this study due to lack of suitably comparable data. The current study has, however, shown that in written data, *gonna* is more frequent in texts associated with popular culture (Section 5.2.4), which could be one reason for, or a consequence of, its being used more by younger people. To judge from the data in Table 6.2, there is indeed an age-related variation in the use of *gonna.* The expression is not only used proportionately more by the younger speakers than by the older in both parts of the corpus, but it also seems to be the case that the distribution is such that the older the speaker is, the smaller the proportion of *gonna.* The reversed pattern is found for *going to* so the combined proportions of the two expressions are similar for all groups of speakers.

6.4 Speakers' social class

The BNC contains information about the social class of a number of the speakers represented in the corpus. There is no information in the corpus documentation on how the classification of the speakers has been done. In the first version of the corpus, the coding of the speakers turned out to be erroneous in part. The only speakers whose social class coding could be considered reliable were the recruited respondents in the DS part (see also Study II). The frequency distribution and proportions of the expressions of future uttered by those speakers are presented in Table 6.3 and illustrated in Figure 6.4. Statistically significant differences between social groups are marked (for information on the significance testing, see Section 3.5.4).

Table 6.3. *Expressions of future. Distribution across speakers' social class in the DS component of the BNC. Percentages and total raw frequencies. Values significantly higher (+) or lower (-) than in the next (lower) social class are marked*

Demographically Sampled (DS)						
	Raw frequency	***will***	***'ll***	***shall***	***going to***	***gonna***
AB top or middle management, administrative or professional	5824	21%	46% -	5% +	9% +	18%
C1 junior management, supervisory or clerical	4251	21%	50% +	4%	6% +	19% -
C2 skilled manual	5055	23%	46% -	4%	4%	23% +
DE semi-skilled or unskilled	2279	21%	53%	3%	3%	19%
Total	17409	3792	8382	693	1073	3469

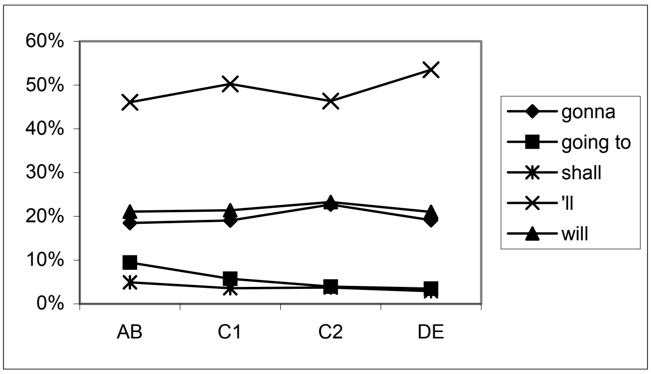


Figure 6.4. Expressions of future in the DS component of the BNC. Proportions for different social classes

The figure and table show that the difference between the social groups is slight regarding the choice of expression of future. The use of *going to* varies with social group so that the proportion is largest for the highest social class (AB: Top and middle management), smaller for the second highest group (C1: Junior management), still slightly smaller for social group C2 (Skilled manual) and smallest for the group of speakers classified as social group DE (Semi-skilled or unskilled). The use of *shall* displays a similar pattern, while the other expressions do not seem to vary consistently with the speaker's social class. There is therefore no strong indication that the speaker's social class should be a decisive factor for the choice of the expression of future. This is further emphasised by the fact that there are no consistent, statistically significant differences that correlate with social group. In Study V, however, it is shown that, as regards *gonna* and *going to*, there is some variation between the social groups. Social group AB shows a certain preference for the full form *going to*, and the two groups DE and C2 use *gonna* significantly more.

6.5 Speakers' education

Information about the education of a speaker is given for some of the BNC speakers. The coverage of the information is relatively limited: only about 20% of the material in the spoken part has this mark-up, and it is found primarily for speakers in the DS component (see also Study II). In the first version of the BNC used for my studies, there are some obvious errors in the mark-up of this feature. The two youngest speakers (15 and 16 years old) found by a search for 'left school at 14 or younger' both have the title 'student'. In the age group 25-34, one speaker who is identified as having left school at age 14 or younger has the social class code AB (Top or middle management) and the title 'teacher/house-wife'. It seems intuitively unlikely that a relatively young teacher has not received further education after the age of 14.

This means, of course, that conclusions based on this material are uncertain. Other considerations also contribute to this uncertainty. Such a consideration is that the education factor is one that can be expected to concur to a great extent with other factors. Assuming that the social class classification is based on occupation (as the labels given to the classes suggest: management, manual, etc.), it is reasonable to presume that education concurs with social class to a certain extent. The education feature can also be expected to correlate with age; today young people in Britain do not leave school at the age of 14 or under. Older people can thus be expected to be over-represented in that educational category. A further factor that makes it difficult to study the distribution of speakers across the education factor is related to the compilation of the corpus. The data collected by recruited respondents from the youngest age group come from the COLT project.[38] These respondents were recruited among the students at some London schools, and thus all belong in the 'still in education' category.

To test the above assumption about the correlation between the factors age and education, and between social class and education, the distribution has been checked for a subset of the data. Tables 6.4 and 6.5 and Figures 6.5 and 6.6 illustrate the number of words said to be produced by speakers of a particular age and social class for two of the levels of education.

Table 6.4. *Data produced by informants still in education across age groups and social class categories in the spoken part of the BNC*

Speaker still in education (310 speakers, 548,444 words)		
age group	**no. texts**	**no. words (proportion)**
0-14	129	228,893 (50%)
15- 24	61	159,619 (35%)
25-34	3	19,339 (4%)
35-44	2	47,040 (10%)
45-59	1	3 (0%)
60+	2	5,711 (1%)
Sum	198	460,605 (100%)
social class	**no. texts**	**no. words (proportion)**
AB	56	226,769 (69%)
C1	11	64,410 (19%)
C2	3	24,888 (8%)
DE	42	14,758 (4%)
Sum	112	330,825 (100%)

[38] COLT= The Bergen Corpus of London Teenager English. For further information, see the COLT webpage http://torvald.aksis.uib.no/colt/ (visited March 3 2005).

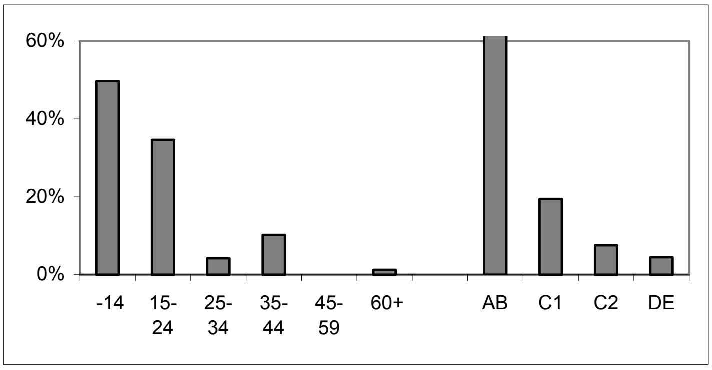


Figure 6.5. Speaker still in education. Proportion of words marked for speakers' age and social class

In the tables and figures, it is shown that the bulk of data marked 'still in education' is produced by speakers in the two lower age groups, while the data produced by speakers who left school at age 14 or younger seem to come from older speakers, primarily in the age group 60+.[39]

Table 6.5. *Data produced by informants who left school at 14 or younger across age group and social class in the spoken BNC*

Speaker left school at 14 or younger (21 speakers, 378,669 words)		
age group	**no. texts**	**no. words (proportion)**
0-14	1	5,722 (2%)
15-24	1	24,315 (6%)
25-34	2	123,735 (33%)
35-44	0	0 (0%)
45-59	2	28,294 (7%)
60+	15	196,603 (52%)
Sum	21	378,669 (100%)
social class	**no. texts**	**no. words (proportion)**
AB	5	126,134 (37%)
C1	3	20,176 (6%)
C2	7	147,853 (44%)
DE	5	45,545 (13%)
Sum	20	339,708 (100%)

[39] The data produced by speakers aged 25-35 who left school at 14 or younger are produced by only two speakers, one of whom is the teacher from social class AB whose classification already has been questioned above.

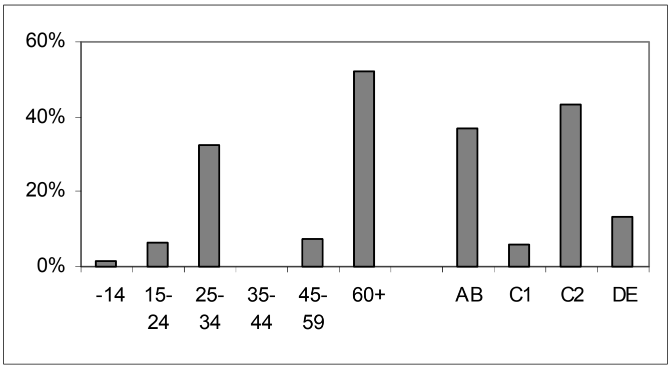


Figure 6.6. Speaker left school <14. Proportions of words marked for speakers' age and social class.

The distribution across the social classes suggests that the speakers in education come from the higher social classes to a greater extent, while the distribution is not so varied among the speakers who left school at the age of 14 or younger. (It should be noted that even if the number of words is fairly high, the number of speakers in the latter group is low: only 21 speakers are coded as 'left school at 14 or younger'). It thus seems that there is a strong correlation between age and the level of education

Against the background of the uncertain classification, uneven coverage of the data, and the possible correlation between different factors, I have found it advisable not to draw any conclusions about the distribution of the expressions of future across the factor 'education'.

6.6 Multiple variables

So far, in the sections above, the speaker-related factors have primarily been studied in isolation. The main finding, apart from the confirmation that the DS and CG components are different as to the proportions of different FUT found in them, was that the use of *gonna* and *going to* seems to vary with a number of speaker-related factors so that when the proportion of *going to* becomes larger, that of *gonna* becomes smaller. That the combined proportion of *gonna+going to* is very similar for all speaker categories supports the assumption that the two expressions are used in complementary distribution.

It is, naturally, not the case that speaker-related factors operate in isolation. To examine to what extent the factors have independent influence on the proportions of the expressions used, a variable rule analysis was performed on a subset of the data (Study V). The subset consisted of all instances of *gonna* and *going to* produced by speakers from the DS component whose age, sex, and social class were known. The analysis indicated that there is indeed an independent influence with some of the factors so that *gonna* was used to a greater extent by speakers from lower age groups and social classes C2 and DE, while older speakers and speakers from social group AB favoured the full form *going to.* No significant difference could be found between the male and female speakers in this respect. The analysis confirmed the preliminary findings made when the factors were studied in isolation (the preliminary analysis found some slight differences between the sexes, but these differences were more pronounced for the CG data). The process and result of the analysis are presented in more detail in Study V.

6.7 Summary

To conclude, it can be said that there are only slight indications that the speaker groups show consistent differences in their choice of expression of future. The main difference appears to be related to the choice between *going to* and *gonna. Shall* is the expression which seems to vary the least with the various speaker-related features. The variation between the CG and DS components is, in all cases where it is possible to study it, found to be greater than the variation between the different groups of speakers. This indicates that the variation between the two text categories cannot be explained with reference to different speaker composition, but is likely to be the result of other factors, such as the level of formality, as suggested in Chapter 5.

7. Region

> Consider, first, the question of language variation. The most obvious type is geographical variation. Everybody is aware that people from different geographical areas are likely to display differences in their speech (Aitchison 2001:39)

7.1 Introduction

It has been noted in the literature that the use of expressions of future varies between different regional varieties of English. Most interest has been devoted to the variation between British and American English, with the use of *will* and *shall* receiving particular attention. Krogvig and Johansson even go as far as to claim that "[n]o single issue has received more attention in discussions of British-American differences than the use of *shall* and *will*" (1984:70).

Among the publications dealing with this issue should be mentioned the grammars by Quirk et al. (1985) and Biber et al. (1999). Quirk et al. state that *shall* is less frequent in American English: "Among the less frequent modals, *should, shall,* and *ought to* are even less frequent in AmE than in BrE" (1985:3.39n). Biber et al. give quantitative evidence on differences between American and British English in the use of the modal verbs. The differences are particularly noticeable in modals marking volition/prediction, such as *will, would,* and *shall*, which are more frequent in British English. The opposite pattern is noted for the semi-modals: for example, *going to* is shown to be more frequent in American English (Biber et al. 1999: 6.6.2). Similar observations are made by a number of other writers. Mair (1997) notes that the *going to* construction is more frequent in certain American texts than in comparable British data (this study is discussed further in Chapter 8). Szmrecsanyi examines the use of expressions of future in spoken British and American English and finds certain regional differencies, for example in the distribution of the negated future marker (2003:305). In their study on the development of future reference, Bybee and Pagliuca say that "*shall* is rare or non-existent in most American dialects" (1987:110). Nakamura (1993) studies modals in the LOB and Brown corpora and finds that there is a difference between the corpora that can be attributed to, among other things, the use of *will.*[40]

Most studies on regional variation concern the British and American varieties. However, as mentioned in Study I, Shastri reports on a study of modals in Indian English[41], where it is found that "[b]y and large the modal usage in IndE conforms to the modal usage in the native [British and American] varieties" (Shastri 1988:17). Despite this similarity Shastri nevertheless notes that there are fewer instances of modals expressing 'futurity' and 'hypothesis' in Indian English, while modals expressing 'certainty' are rarer in the British English variety. Shastri speculates about whether this can be explained with reference to cultural differences, as a phenomenon that can be seen as "reflections of the peculiar Indian mode of thought". The author suggests that "[m]aybe the Indian mind is not given to thinking much in terms of the future and if the western culture shies away from categorical or strong views, the expression of certainty, the Indian mind does not" (Shastri 1988:18).

This chapter summarises my findings relating to regional variation in the use of expressions of future. It is based on Studies I and, to some extent, II and V where I examine the use of expressions of future in samples of language of different geographical origin. In Study I, written language from three different countries is compared, while Studies II and V deal with different British accents/dialects.

7.2 Variation between British, American, and Indian English

Study I reports on an investigation of the expressions in the Brown (American English), LOB (British English) and Kolhapur (Indian English) corpora. It is shown that the greatest regional difference is that between the Indian English corpus on the one hand and the British and American ones on the other. The difference is one of frequency as well as of proportions. The number of expressions of future is significantly lower in the Indian English Kolhapur corpus than in the comparable British one: 2,642 instances compared to 3,348. The frequency in the American corpus is higher than in the Indian but lower than in the British (3,089).[42] It is interesting to note, however, that although the Indian English corpus as a whole contains fewer in-

[40] The frequencies of the modals quoted by Nakamura differ from those I obtain when searching the corpora.

[41] The study Shastri refers to is P.B. Katikar, (1984). *The meanings of the modals in Indian English*, an unpublished PhD. dissertation from Shivaji University, Kolhapur.

[42] In Study I, only instances with overt subjects and infinitival verbs are considered, which is why the frequencies for LOB given there are slightly lower than the ones presented elsewhere (for example in Study IV and Chapters 3, 5, 8).

stances of the expressions of future, it is not the case that all genres in the Kolhapur Corpus have lower frequencies of the expressions than the corresponding British or American genres (see also Chapter 5, Figure 7.1, Study I:12).

As far as the proportions of the expressions are concerned, it has been shown above that there is considerable variation between the Informative and Imaginative hyper-categories in the LOB, Brown, and Kolhapur corpora (see Chapter 5 and Study I). *Will* is the most frequent expression overall, while *going to* is the least frequent (excluding *gonna,* which only occurs twice in LOB, 14 times in Brown and never in Kolhapur). The contracted form *'ll* is used less in the Indian corpus than in the other two, while *shall* is used more. The distribution of the expressions of future across the three full corpora (as given in Study I) and their two hyper-categories is presented in Table 7.1 and commented on further below. The table contains information about where statistically significant differences have been found (for information about the significance testing, see Section 3.5.4).

Table 7.1. *Proportions of the expressions of future and statistically significant differences between the corpora. <X reads 'less than in X', >X reads 'more than in X'. - = no significant difference. L=LOB, B=Brown, K=Kolhapur*

Whole corpora				
	will	***'ll***	***shall***	***going to* + *gonna***
LOB	69% <B <K	15% >K	11% >B <K	5% >K
Brown	73% >L	14% >K	8% <L <K	4% >K
Kolhapur	74% >L	9% <L <B	14% >L >B	3% <L <B
Informative hyper-category				
	will	***'ll***	***shall***	***going to* + *gonna***
LOB Informative	83% -	2% <B >K	11% <K	3% >B
Brown Informative	84% -	4% >L >K	10% <K	2% <L
Kolhapur Informative	83% -	0% <L <B	15% >L >B	2% -
Imaginative hyper-category				
	will	***'ll***	***shall***	***going to* + *gonna***
LOB Imaginative	39% <K	42% >K	9% >B	9% <B >K
Brown Imaginative	37% <K	46% >K	4% <L <K	12% >L >K
Kolhapur Imaginative	56% >L >B	28% <L <B	11% >B	5% <L <B

7.2.1 *Will*

As mentioned above and in Study I, *will* is the most frequent expression of future in all three corpora, constituting between 69% and 74% of all FUT on average. The proportions are much smaller in the Imaginative hyper-categories than in the Informative in all corpora, as discussed in Chapter 5.

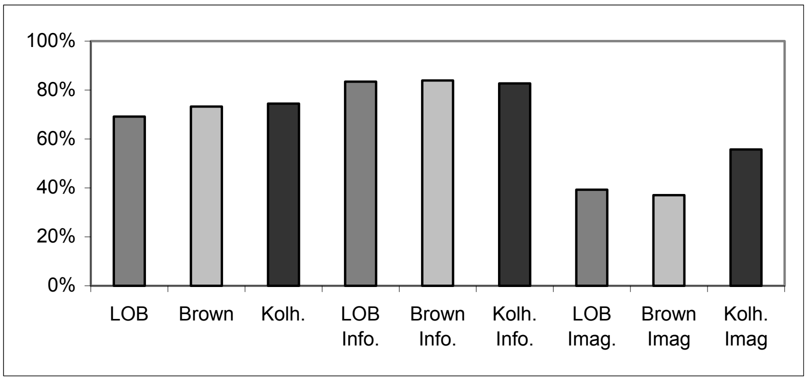


Figure 7.1. Proportions of *will* in the LOB, Brown and Kolhapur corpora (full corpus, Informative hyper-category, and Imaginative hyper-category)

As shown in Table 7.1 and Figure 7.1, the proportion of *will* is lower in LOB as a whole than in Brown and Kolhapur. However, this difference is not found if the hyper-categories are considered separately. There is no significant difference in the use of *will* between the different Informative hyper-categories while in the Imaginative hyper-category, the proportion of *will* is significantly larger in the Kolhapur corpus than in the other two. It thus appears that there is no consistent variation between the different regional varieties of English as far as the use of *will* is concerned.

7.2.2 *'ll*

The expression *'ll* is used significantly less in the Indian corpus than in the British and American corpora: the overall proportion is only 9%, compared to 15 % and 14% in LOB and Brown respectively. The difference between Kolhapur on the one hand and LOB and Brown on the other is found in the full corpora as well as in each of the two hyper-categories, as illustrated in Figure 7.2.

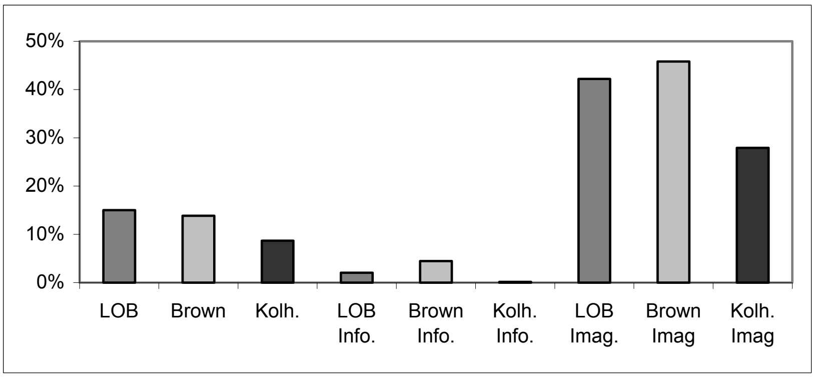


Figure 7.2. Proportions of *'ll* in the LOB, Brown and Kolhapur corpora (full corpus, Informative hyper-category, and Imaginative hyper-category)

The proportions of the expression are very low in the Informative hyper-categories: 2% in LOB, 4% in Brown (this difference is statistically significant) and 0% in Kolhapur. The proportions are much larger in the Imaginative hyper-categories. In Brown, nearly every other expression of future in the Imaginative hyper-category is *'ll* (46%)*,* and in LOB the proportion is 42% (no significant difference). In the Indian Imaginative data, however, the proportion is only 28%.

A noticeable feature of the Indian corpus is that *'ll* is used only in the Imaginative hyper-category. In the 347 Informative Indian texts, *'ll* is found only three times. The expression is also rare in the British Informative texts (47 instances) but somewhat more frequent in the corresponding American hyper-category (108 instances) (see Study I:13).

One possible explanation for the small proportion of *'ll* in the Indian data can be found in what Shastri (1988:18) refers to as "the predominance of written language over spoken in the Indian pedagogical context". The expression *'ll* is used more in spoken and speech-like contexts. If the Indian writers are primarily educated in English as a written language, they may avoid the use of the contracted form *'ll* in their writing, even in genres where British and American writers use it freely.

7.2.3 *Shall*

As shown in 7.1, it is generally understood that *shall* is not used to a great extent in American English. When the proportions of the expression in the LOB and Brown corpora are compared, it appears that there is indeed a statistically smaller proportion of *shall* in the Brown Corpus as a whole than in LOB (8% compared to 11%). This difference is, however, primarily found in the Imaginative texts. The difference is slight in the Informative hyper-category, where the proportion of *shall* is 11% in LOB and 10% in Brown.

The use of *shall* varies not only between British and American English but also between Indian English and the other two varieties (see Figure 7.3).

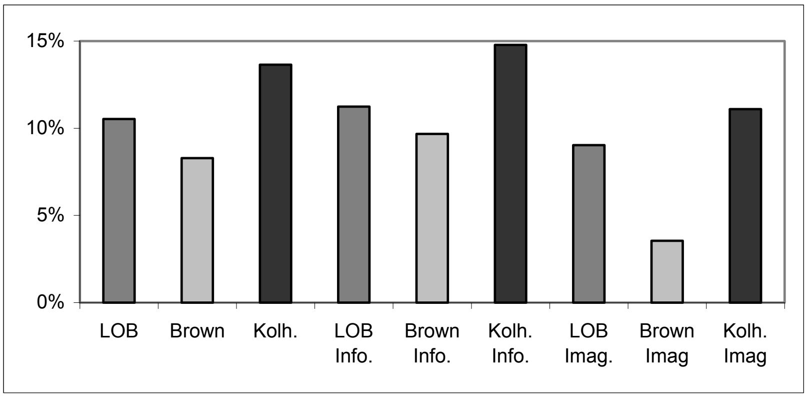


Figure 7.3. Proportions of *shall* in the LOB, Brown and Kolhapur corpora (full corpus, Informative hyper-category, and Imaginative hyper-category)

Shall is used proportionately more in Kolhapur than in the comparable LOB and Brown corpora overall, and this difference is found in both hyper-categories (the difference between the British and Indian Imaginative hyper-categories is not statistically significant). Shastri (1988) comments on the higher frequency of *shall* in the Kolhapur Corpus, and suggests that, in more than one way, it can be explained with reference to the fact that English is a second language in India. Shastri suggests that features that are no longer found in English when used as a first language are still being taught to second language learners. As shown in Chapter 8 below, *shall* is used more in the LOB and Brown corpora from 1961 than in the comparable FLOB and Frown corpora from 1991 and 1992. The proportion of *shall* in the Indian data (which is from 1978) is even higher than in the LOB corpus. This could lend support to the hypothesis that Indian usage mirrors earlier British usage (before 1961).

Krogvig and Johansson (1984) suggest that the difference between the British and American corpora in the proportion of *shall* can be explained with reference to the subject of the expression. This suggestion is discussed further in Chapter 9 where it is shown that this explanation cannot account for the diverging pattern found in the Indian data.

7.2.4 *Going to*

As noted above, it is generally assumed that *going to* is used more in American English than in British (I have not been able to find any empirical studies on this difference between other regional varieties, such as Indian English). There is, however, only a slight difference between the full LOB and Brown corpora in this respect, as shown in Figure 7.4.

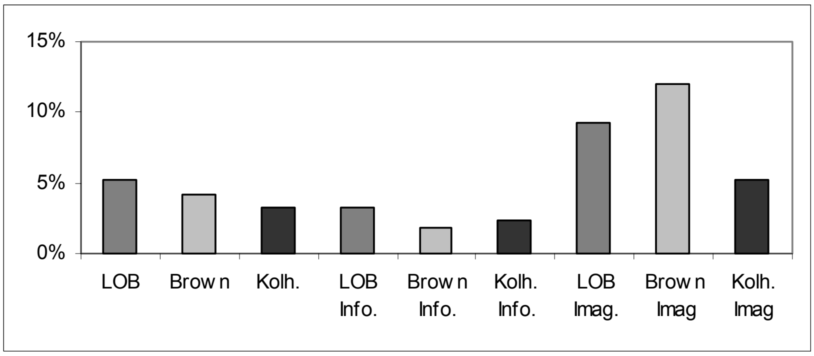


Figure 7.4. Proportions of *going to* in the LOB, Brown and Kolhapur corpora (full corpus, Informative hyper-category, and Imaginative hyper-category)

When the hyper-categories are considered separately, a rather complex pattern emerges. In the Informative hyper-category, the proportion of *going to* is significantly higher in LOB than in Brown. In the Imaginative hyper-category, however, the reverse is the case. The proportion of *going to* in Kolhapur is similar to that in the other two corpora in the Informative hyper-category but smaller in the Imaginative one. It should be pointed out in this context that in the Informative hyper-category the frequency of the *going to* expression is very low: only 74 instance in LOB, 44 in Brown and 43 in Kolhapur (corresponding to proportions of 3%, 2% and 2%). The frequencies and proportions are higher in the Imaginative hyper-categories, where the American corpus contains a higher proportion of *going to* than the others: 12% compared to 9% in LOB and only 5% in Kolhapur.

There may be several reasons for the difference between the corpora with respect to the proportion of *going to*. As the expression is more frequent in spoken language, it cannot be excluded that the low frequency in the Indian sample can be explained, at least in part, by Shastri's suggestion that Indian users are educated to write in a certain way (see above). It has also been shown that *going to* is used in quotes to a considerable extent (Chapter 4, Study I). Another possible explanation for the difference between the corpora in the use of *going to* is that the Indian sample contains fewer quotes. A comparison of the proportions of quoted instances of *going to* in the three corpora suggests that this may indeed be the case (see Chapter 4). In the

Indian sample, 49% of the 86 instances of *going to* were found in single or double quotation marks. The proportions in the American and British samples were larger, 71% and 57% respectively. As the frequencies are low it is, naturally, not possible to draw any far-reaching conclusions about the use of *going to* in relation to quoted speech on the basis of these findings. It cannot be excluded, however, that the observed differences are related to differences in the amount of speech-like text in the corpora.

7.2.5 *Gonna*

The expression *gonna* is very infrequent in the LOB, Brown and Kolhapur corpora (2, 14 and 0 occurrences respectively). It is thus not possible to draw any conclusions about differences between the regional varieties on the basis of the data available in the corpora. The expression is, admittedly, more frequent in the American corpus, but the frequencies are very low. In FLOB and Frown, the comparable British and American corpora from 1991 and 1992, *gonna* is somewhat more frequent than in the earlier LOB and Brown corpora. In Frown, *gonna* occurs 31 times (29 times in present tense contexts), which is noticeably more frequently than in the comparable British English FLOB, where the expression is found only five times (four in the present tense). Although the frequencies are low, they lend some support to claims that *gonna* is used primarily in American English, something which is suggested in dictionaries and reference grammars. The lack of comparable data, however, makes it impossible to treat the issue more extensively in the present context.

7.3 Variation within national varieties of English

So far this chapter has dealt with comparison between national varieties of English: British, American, and Indian. To my knowledge, the use of different expressions of future has not yet been examined in any greater detail with respect to regional variation within the national varieties of English. In one of the few published studies on the subject (Poplack and Tagliamonte 2000) it is noted that

> ... the few published observations on the expression of future in AAVE focus not on the opposition between *will* and *going to*, but on putative distinctions among the variant forms of *going to* (e.g. *gonna, gon*), the phonological reduction of which is said to be “highly characteristic” of AAVE (Poplack and Tagliamonte 2000:322)

In their study, Poplack and Tagliamonte examine spoken data from five Afro-American Vernacular English (AAVE) communities,[43] and find that in all but one community, the *going to* expression is more frequent than *will.* Similar results are obtained by Facchinetti (1998) in her study on British Caribbean Creole. This extensive use of *going to* is not found in any of the data examined in my studies, although the proportion of *going to* is considerably higher in the spoken BrE data than in the written (see Chapter 4). Poplack and Tagliamonte also examine four variants of *going to* (*gointa, gonna, gon, go*). They find that the variants are used to different extents in the five communities, but that the variants display such great similarities that it can safely be assumed that they are of common origin.

I deal with regional variation between *going to* and *gonna* in Study III. One section briefly treats the use of *going to* and *gonna* in different regional varieties of British English. Some variation between the regional varieties as regards the proportion of *gonna* is found but it is obvious that the pattern of variation is not always the same for the Context-governed (CG) and Demographically Sampled (DS) components of the BNC (the two components are described in Chapter 3), which suggests that regional variation is not a very influential factor.

A possible problem with the data in the BNC is related to the composition of the corpus. Information about the speakers' accent/dialect is available only for a fairly small proportion of the data. Furthermore, this classification has often been made by the respondents, i.e. the individuals who made the recordings. We cannot therefore rely fully on their being consistent or completely accurate (see Study II). For these reasons it has been considered of little value for a description of regional variation to examine the distribution of the different expressions of future across the accents as given in the BNC.

7.4 Summary

It has been shown that the use of the expressions of future varies between the three regional varieties of English represented by the LOB, Brown, and Kolhapur corpora. The frequency of the expressions is highest in the British corpus and lowest in the Indian corpus. The choice of expression varies to some extent between the corpora. As previous studies suggest, *shall* is used less in American English, and *going to* more. Indian English, as represented in the Kolhapur corpus, differs from the other two varieties in that the proportion of *shall* is larger, and the proportion of *going to* is smaller. The expression *'ll* is used less in Kolhapur than in LOB and Brown, and only in the Imaginative hyper-category. A common feature for all three corpora is that the variation between the hyper-categories within the corpora is consider-

[43] Guysborough, North Preston, Samaná, Guysborough Village, Ottawa.

able. It thus appears that even though there are regional differences, other factors (such as text category) are more important for the choice of expression. This is also the conclusion to be drawn after studying the use of *gonna* and *going to* across different dialects in the BNC. I therefore suggest that regional variation is not a prominent feature in the choice of the expressions of future.

8. Time

8.1 Introduction

Previous chapters examine the variation in the use of expressions of future relating to type of text (medium and text categories) as well as to the producer of the text (speaker properties and region). The present chapter draws together the results from my studies of variation across time. In Study IV, I examine to what extent the use of the expressions has changed from the 1960's to the 1990's by comparing the written LOB and FLOB corpora and the spoken LLC and Sampler material. These results are complemented with two new investigations, one looking at the distribution of the expressions across the press texts in four corpora, the other examining the variation in the LOB and FLOB further.

Although the use of the expressions of future in different times has received considerable attention in the literature, there are, to my knowledge, no substantial comparative studies of the use of the expressions in the near-diachronic perspective.[44] Aitchison comments on the increased use of *going to* over time and suggests that "[t]his is a construction whose progress is likely to be interesting in the next twenty or so years" (1991:100, 2001:110). She refers to data presented by Potter (1969), which are similar to those given by Danchev et al (1965).[45] Danchev et al. state that "[t]he construction *to be going to + inf.* has spread considerably during the last 50-60 years in modern English, and ... this process continues" (1965:375). They base their claim on what is referred to as "statistical data", and present raw frequencies and proportions for *going to*, found in a number of plays and novels by Brit-

[44] The term 'near-diachronic' is used here to refer to a short period of time near the present, roughly 1960-1990.

[45] Potter acknowledges that he got his figures from Dr Molhova of the University of Sofia. As a minor point it can be noted that although the raw frequencies quoted are almost identical, the percentages presented differ greatly between Potter's article and that by Danchev et al. Potter claims that in *The Catcher in the Rye*, Salinger uses *going to* in over 30 per cent of cases while Danchev et al. arrive at a proportion of "nearly 25 per cent". To judge by the frequencies presented, the proportion is just under 24 per cent (75 *going to* and 240 or 239 instances of *will/shall*).

ish and American writers from the 19th and 20th centuries.[46] These observations are of interest, even though it could be questioned whether the observed differences can be explained with the publication date alone rather than with a combination of factors such as text category or region, factors that have been shown to correlate with variation in the use of expressions of future to some extent (see, for example, Studies I and IV, Chapters 5 and 7). The choice of texts was not treated by Danchev et al. or Potter, nor was the variation between different texts dwelt upon in their publications.

8.1.1 Press texts

If the claims about the use of *going to* made by Danchev et al. and Potter were based on data which may not be well suited for comparison, the study by Mair (1997) provides a welcome approach. Based on the Press component (the three text categories A, B, C) of the four matching corpora LOB, FLOB, Brown, and Frown, Mair investigates the change in use of the *going to* future between 1961 and 1991/92 (see Chapter 3 for a presentation of the corpora). He notes that there is a significant increase in the number of *going to* in the later texts, and that this increase is particularly noticeable in the American sample. His analysis of the examples shows that *going to* is found in some new uses in the later corpora, but that the instances of these are not frequent enough to account for the increase as a whole. Mair suggests that the increased frequency of *going to* can be explained with reference to changes in the behaviour of the language user, so that *going to* now "tends to be chosen more often in contexts in which it has long been established and competing with *will* and *shall* and other expressions of futurity, for example as the stylistically informal alternative" (1997:1541).

Although Mair notes that *going to* competes with other expressions of future, he does not deal with the potential change in frequency of these other expressions. If *going to* is chosen at the expense of *will/shall*, an increase in *going to* should be mirrored in a comparable decrease in the use of other expressions of future: in other words, the proportion of *going to* should increase, not only the frequency. It is useful to consider this aspect, especially when dealing with small corpora such as these, where the absolute number of instances is comparatively small. It is not unlikely that text-related factors (e.g. the influence of stylistic, idiolectal and tense-related factors) affect the number of instances used. With few texts under investigation, occasional variation in individual samples also has a greater influence.

To examine to what extent the increase noted by Mair is also a relative increase of the *going* to construction, I derived the frequencies for the other

[46] The proportions of *going to* are calculated as proportions of the sum of all instances of *going to, will,* and *shall*.

four expressions of future in the Press categories (genres A, B, C) of the LOB, FLOB, Brown and Frown corpora and calculated the proportions of each expression in the four corpora. These are presented in Table 8.1. To enable comparison between the different expressions of future, only instances of present tense *going to* are considered (which is why the figures do not match exactly those given by Mair).

Table 8.1. *Proportions and total raw frequencies of expressions of future in the Press texts (genres A-C) of four corpora. + = significantly higher proportion. Comparisons only LOB-FLOB and Brown-Frown*

	will	***'ll***	***shall***	***going to + gonna***	**Raw frequency (100%)**
LOB	90% +	2%	4%	3%	692
FLOB	86%	6% +	3%	5%	785
Brown	88% +	5%	3% +	3%	775
Frown	83%	8%	1%	8% +	720

Table 8.1 shows that the proportions of *going to* have increased in both later corpora: from 3% in the earlier LOB and Brown corpora to 5% and 8% in the FLOB and Frown corpora respectively. When the figures for the other expressions are examined, it is found that the proportion (although not the number) of *will* seems to have decreased with time in both the British and American varieties, and so has the proportion of *shall,* while the proportion of *'ll* has increased. The decrease over time in the use of *will* is statistically significant in both the British and American material, while the other changes are statistically significant in only one of them (as illustrated in Table 8.1). It thus appears that the increase in the use of *going to* found by Mair is indeed at the expense of *will* and *shall* (although not *'ll*), and not due to an increased use of expressions of future as a whole. It is worth noting, however, that the increase in the British data is not large enough to make the results statistically significant (for information about the significance testing, see Section 3.5.4).

Regional differences have been discussed in Chapter 7. The remainder of the current chapter will focus on the variation over time in British English, as reported in Study IV.

8.1.2 LOB vs. FLOB

As mentioned above, Mair's study is concerned with a particular type of data, i.e. newspaper texts. The author notes that as his study is based on a specific type of text, it is possible that the change observed is "not a sign of grammatical change, but of stylistic development that can be documented in many other ways" (1997:1541). (For discussions on the development of newspaper language, see, for example, Westin 2002). To be able to say anything about the development in general, it is advisable to turn to balanced corpora where more than one type of text is included. The comparable corpora LOB and FLOB are well suited for near-diachronic comparisons and I have used them for this purpose.[47] In Study IV, the full corpora are compared and also studied across the Informative and Imaginative hyper-categories in an attempt to discern possible differences in the use of the expressions of future that can be attributed to changes over time. The study shows that the absolute overall frequency of the expressions of future is slightly higher in the LOB corpus than in FLOB: 3,362 occurrences (LOB) compared to 3,088 (FLOB). The frequencies in the hyper-categories vary – they are higher in LOB than in FLOB in the Imaginative hyper-category and lower in the Informative hyper-category (see also Chapter 5). As far as the number of FUT is concerned, there does not seem to be any significant overall change across the timespan covered by the LOB and FLOB corpora. The proportion of FUT, however, varies more, as indicated in Table 8.2 (raw frequencies can be found in Tables A.1 and A.2 in Appendix B).

Table 8.2. *Proportions of expressions of future in the LOB and FLOB corpora (adapted from Study IV:33). + = significantly higher proportion*

	will	***'ll***	***shall***	***going to* +*gonna***	**Raw frequency (100%)**
LOB	69%	15%	11% +	5%	3,362
FLOB	75% +	13%	6%	5%	3,088
LOB Info	83%	2%	11% +	3%	2,277
FLOB Info	87% +	4% +	6%	3%	2,287
LOB Imag	39%	42%	9%	9%	1,085
FLOB Imag	42%	40%	7%	11%	801

In Table 8.2 it is shown that the proportions of the expressions of future vary between LOB and FLOB. When the distribution across the two whole corpora is examined as well as the distribution across the hyper-categories, it is apparent that there is little consistent variation between the earlier and the later corpus. It is only for the expression *shall* that the proportions are consistently (though not always significantly) lower in the later corpus. The use of the other expressions does not seem to vary with time: there is no consistent variation between the two corpora.

[47] For a presentation of the corpora, see Section 3.4.1. Division of the corpora into text categories is described in Section 5.1.

8.1.2.1 Will

Will is the most frequent expression in both LOB and FLOB, with a significantly higher proportion in the later corpus (75% compared to 69% in LOB). However, this difference is not found in all parts of the corpus. As Figure 8.1 illustrates, the proportion of *will* fluctuates considerably between the genres.

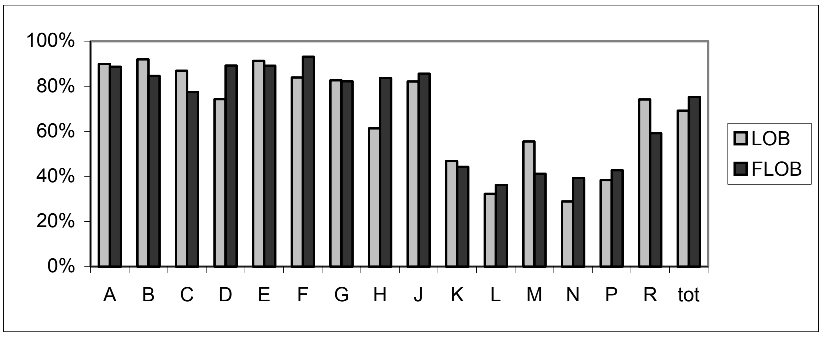


Figure 8.1. Proportions of *will* in the 15 genres in the LOB and FLOB corpora

The figure shows that the pattern of variation between the earlier LOB and later FLOB is uneven as regards the proportion of *will*. It can be seen that the expression is used more in LOB in seven of the text categories (in a statistically significant way in genre B: Press: editorial only), more in FLOB in seven of the categories (statistically significant differences in D, F, H, and N[48]), and to the same extent in one of the 15 categories (G: Belle lettres). There is no variation with hyper-category either; there are genres with larger and smaller proportions in the Informative (genres A–J) and Imaginative (genres K–R) hyper-categories in both corpora. This means that although the overall proportion of *will* is higher in FLOB than in the earlier LOB, the variation within the corpora indicates that no uniform change can be discerned in the use of *will* in these data.

An examination of spoken data presents similar results. As shown in Study IV, the proportion of *will* is somewhat larger in the earlier LLC than in the spoken part of the BNC Sampler (for a discussion of the comparability of the spoken corpora, see Section 3.4.2). However, the Sampler data are not uniform (see Chapter 5). The proportion of *will* is higher in the CG component than in the somewhat earlier LLC. In the DS component, which is from roughly the same time as the CG component, on the other hand, the proportion is lower than in the LLC (see Study IV:35). These results indicate that

[48] D= Religion, F= Popular lore, H= Miscellaneous, N= Adventure.

the difference between the LLC and Sampler corpora cannot be explained exclusively with reference to changes in language over time.

8.1.2.2 'll

In a study of variation between the press categories (genres A–C) in LOB and the Uppsala Press Corpus (UPC), a comparable corpus of press texts from 1994, Axelsson (1998) shows that the use of contracted forms in general has increased from 1961 to 1994. Similar results were obtained by Krug (1996). In Table 8.1 above, it could be seen that the proportion of the contracted form *'ll* is more than twice as high in the FLOB press texts as in the corresponding part of the LOB. This coincides with the decrease in the use of *will* in this part of the FLOB corpus. When only press texts are considered, it thus seems that the use of *'ll* increased from 1961 to 1991, at the expense of the full form *will.* When the distribution across all the genres is considered, however, the pattern of variation is less clear. As illustrated in Figure 8.2, the expression *'ll* is used more in the later corpus in seven genres and less in seven (it is not used at all in either corpus in genre H: miscelleaneous). The difference between the corpora is statistically signifi-cant for genres B, E, J (less *'ll* in LOB), and N (more *'ll* in LOB).[49]

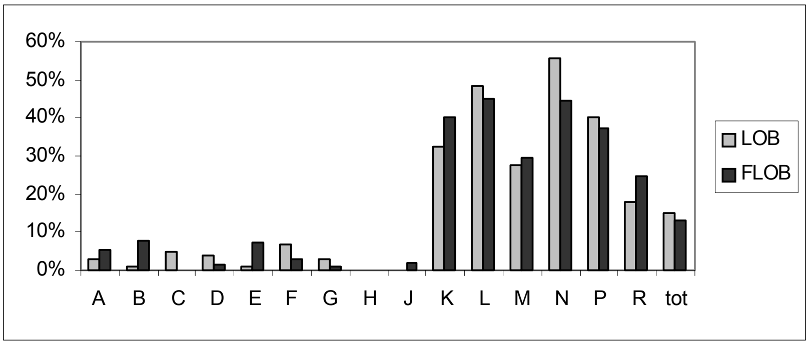


Figure 8.2. Proportions of *'ll* in the 15 genres in the LOB and FLOB corpora

There is considerable variation within the hyper-categories; it is never the case that all genres in a hyper-category show the same development. The differences between the corpora are often small: the variation between the hyper-categories within the same corpus is at all times more noticeable (this difference is examined in more detail in Chapter 5). The frequencies of *'ll* in the Informative hyper-category are very low, never over 24 occurrences (see

[49] B = Press: editorial, E = Skills, trades and hobbies, J = Learned, F = Popular lore, N = Adventure and Western fiction

Tables A.1 and A.2 in Appendix B for raw frequencies). With such low frequencies, any differences between the corpora must be interpreted with great caution, even when calculations indicate that there is a statistically significant difference. The higher frequencies in the Imaginative genres make the results for this hyper-category more reliable, pointing to a slightly smaller proportion of *'ll* in the FLOB corpus in this hyper-category. The overall distribution in the Imaginative hyper-category thus seems to indicate that there has been a slight decrease over time in the use of *'ll*. This is a development contrary to that found in studies of Press texts (Axelsson 1998, Krug 1996). It is well worth pointing out in this context that the variation between the genres is considerable, and does not indicate that there is any consistent development with regard to the use of the *'ll* expression. More detailed studies of the use of contractions have shown that the use of contracted forms varies not only with time but that the distribution is highly dependent on other factors such as the position of the contracted form in the text and whether the instance is found in quoted speech or not (see Axelsson 1998).

In the spoken corpora, the proportion of *'ll* differs between the LLC and each of the two Sampler components. The proportion in the LLC is higher than in CG but lower than in the DS component of the Sampler. There is therefore nothing in this variation that points to an overall change over time. It is worth pointing out once again that the LLC and Sampler corpora are not necessarily good comparable samples (see Section 3.4.2).

8.1.2.3 Shall

The study of the distribution of *will* and *'ll* across the genres in LOB and FLOB did not reveal an overall pattern of variation with time. A similar study of *shall*, however, turns out to be more rewarding in this respect. The proportion of *shall* is larger in LOB than in FLOB in the whole corpus as well as in the hyper-categories. Figure 8.3 illustrates the proportion of *shall* in LOB and FLOB (raw frequencies in Tables A.1 and A.2 in Appendix B).

In the figure it can be seen that the proportion of *shall* is larger in the earlier LOB Corpus in ten of the 15 genres (statistically significant for genres A, E, H, and K[50]) and smaller in three. The proportions in categories L:Mystery and P:Romance are very similar. Genre M:Science fiction displays the most diverging pattern, with a proportion of *shall* that is about four times larger in FLOB than in the earlier LOB corpus. This genre is, however, the smallest of all genres in the corpora, and contains a very low number of expressions of future overall; 36 in LOB and 34 in FLOB. There is only one instance of *shall* in genre M in LOB and four in FLOB. With such low frequencies, proportions will be subject to substantial fluctuation even with

[50] A = Press: reportage, E = Skills, trades and hobbies, H = Miscelleaneous, K = General fiction.

very modest variation in raw frequencies, and should be interpreted with appropriate caution.

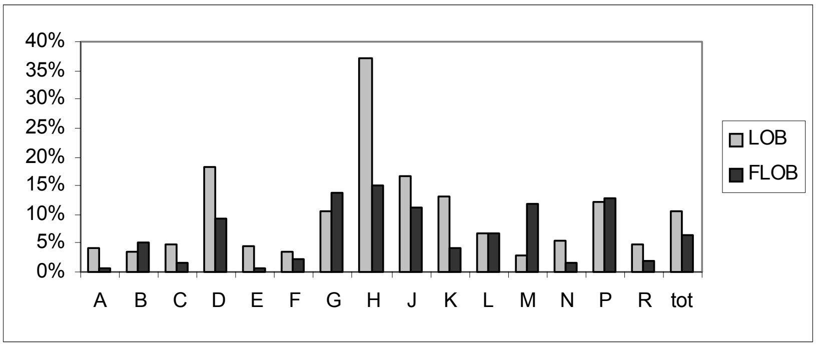


Figure 8.3. Proportions of *shall* in the 15 genres in the LOB and FLOB corpora

As already mentioned (Chapter 5, Study IV), the relatively high proportion of *shall* in the FLOB genre P: Romance can be explained with reference to the texts in that genre; they can be seen to illustrate the authors' attempts to mirror old-fashioned usage. Study IV provides further examples of how the use of *shall* seems to have changed over the relatively short timespan covered by the LOB and FLOB corpora. In the later FLOB corpus in particular, the expression is found in a limited number of texts (see Chapter 5). It is often used in quotes (see Chapter 4).

Data from the LLC and the spoken part of the BNC Sampler also suggest that the use of *shall* has decreased with time. The expression constitutes about 8% of the expressions of future in the LLC, but only 4% in the somewhat later BNC Sampler. In contrast to the situation with the *will* and *'ll* expressions, there is no difference between the DS and the CG components with respect to the proportion of *shall*.

The raw frequencies of the *shall* expression are generally low, especially when individual genres are considered. The overall result of the above analyses of the material nevertheless points in one general direction, indicating that the use of *shall* has decreased over time.

8.1.2.4 Going to

The study of the Press components (see above) indicated that the use of *going to* increased between 1961 and 1991, even if the proportionate increase in the British material was not quite as large as the increase in absolute frequencies might suggest. A comparison of the whole LOB and FLOB corpora, however, reveals a more complex pattern. When the whole corpora are compared, the use of *going to* does not seem to have increased at all. If the hyper-categories are studied separately there are even indications of a contrary development. In the Informative hyper-category (where the Press genres are included), there is a slight decrease in the proportion of *going to* from 1961 to 1991. In the Imaginative hyper-category, on the other hand, a certain increase can be discerned. This complex pattern of variation is no less opaque if the differences between individual genres are considered. As Figure 8.4 reveals, there is considerable variation between the 15 text categories, both between and within the corpora (raw frequencies in Tables A.1 and A.2 in Appendix B).

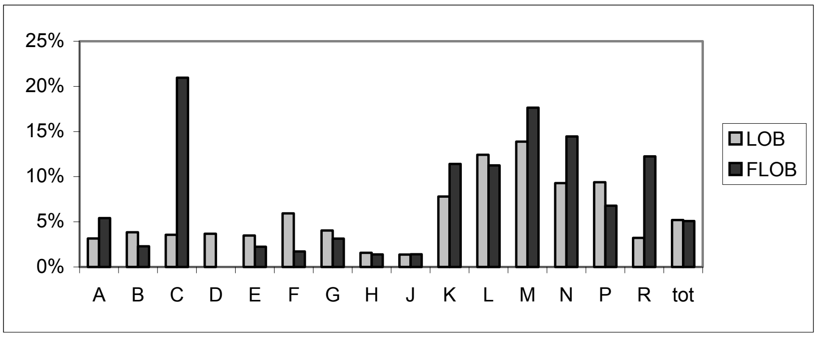


Figure 8.4. Proportions of *going to* in the 15 genres in the LOB and FLOB corpora.

As illustrated in the figure, the proportion of *going to* is larger in the later FLOB corpus in six of the 15 genres. It is larger in LOB than in FLOB in seven genres and approximately the same in two. The only statistically significant differences are found for genres C (less *going to* in LOB) and F (more *going to* in LOB). The overall proportions for the corpora, as well as for the different text categories suggest that there is no substantial change over time in the use of *going to*. This pattern does not change when the few instances of *gonna* are added to the frequencies for *going to.*

A somewhat different result is obtained when spoken material is considered. Comparisons between the LLC and the spoken part of the BNC Sampler suggest that the combined use of the variant expressions *going to* and *gonna* has increased over the last twenty or so years, at the expense of *will/'ll/shall*. Admittedly, one potential problem with these results is that they are not based on matched corpora. It cannot be excluded that the results obtained are in fact dependent, wholly or in part, on factors related to the structure or composition of the corpora, rather than on change over time.

8.1.2.5 Gonna

As regards the use of the *gonna* construction, the written LOB and FLOB corpora yield few insights as they only contain two and four instances of the expression respectively. In the spoken material, however, there is a substantial difference between the corpora. The expression is considerably more frequent in the Sampler corpora than in LLC, while the proportion of *going to* is larger in LLC (raw frequencies in Tables A.5 and A.7 in Appendix B). This could suggest that the use of *going to* has decreased over time to the benefit of an increase in the use of *gonna.* It is, however, unsuitable to base any statements about this development on the variation in the Sampler and LLC corpora, as not only differences in the composition of the corpora but also varying transcription practices may influence the distribution of the *gonna/going to* expressions. Due to the lack of suitably comparable data no statements will be made about the near-diachronic development of *gonna* (and consequently of *going to*) in spoken Present-day English.

8.2 Concluding remarks

As described in the beginning of this chapter, some previous studies suggest that the use of *going to* is increasing with time at the expense of *will/shall.* The results of the new case studies presented here indicate that the pattern is more complex than it might seem at first glance. The data available in the two matched corpora LOB and FLOB do not support the notion of an overall increase in the use of *going to*, while indications of such an increase can possibly be discerned in the spoken data. The changes found in the use of *shall* are more conclusive, and point to a factual decrease in the use of the expression in the written as well as the spoken data.

Thus the results presented in this study point to a varied but not necessarily changing pattern of use of *going to* in the near-diachronic time span. The study by Mair (1997) and Study IV, however, give reason to assume that changes may be taking place, as indicated by qualitative analyses of individual examples (Mair) and by studies of the distribution across hyper-categories and medium as well as co-occurrence patterns (Study IV). It is possible that further studies of larger quantities of comparable data from different periods (presently not available in corpus format) might provide better insights into the near-diachronic development of the choice of expressions of future. There will no doubt be reason to return to this question again in the future, to examine this potentially on-going change.

9. Linguistic association patterns

9.1 Introduction

So far, this summary has focused on the associations between the expressions of future and a number of non-linguistic factors (medium, text category, speaker properties, region, and time). The current chapter presents the results of my studies related to the linguistic association patterns, primarily Studies III and IV. In these studies I investigate the syntactic and lexical environments of the expressions of future by examining patterns where the expressions of future co-occur with particular lexical items or items from particular word-classes.[51] Study IV is based on data drawn from a number of written and spoken corpora (LOB, FLOB, LLC, and the spoken part of the BNC Sampler). The study deals with co-occurrence patterns between the FUT and personal pronouns used as subjects, as well as co-occurrences with very frequent main verbs. Study III deals exclusively with data from the spoken part of the BNC, identifying patterns for all five expressions but with a focus on *gonna* and *going to*.

I use the term ‘collocations’ to refer to the co-occurrence patterns I examine. Items co-occurring with the expressions of future are called ‘collocates’, while the expression itself is termed ‘node’ (following, for example, Stubbs 2001). The positions preceding the node are labelled -1, -2 etc, while those following the node are referred to as +1, +2, etc. (see Figure 9.1 for an illustration).[52]

[51] The term ‘word-class’ (also POS) is here used in a broad sense involving reference to items with a particular part-of-speech tag (such as VVI for infinitival form of lexical verbs, VBB for present tense forms of lexical verbs*,* VBI for infinitival *be* etc.) as well as reference to groups of words traditionally known as word-classes, such as (personal) pronouns, (infinitival) verbs, etc. (see, for example, Study III:167).

[52] In Study III, I use the term ‘colligation’ to refer to co-occurrences between items from particular word-classes and the FUT, while ‘collocation’ is used only for co-occurrences with individual lexical items. In this Summary I have opted to use the term ‘collocation’ for both types.

Collocations with:	COLLOCATES			NODE	COLLOCATES		
	Position				position		
	-3	-2	-1	node	+1	+2	+3
Lexical items			*He*	*will*	*sleep*	*longer*	*.*
	Are	*you*	*really*	*going to*	*tell*	*them*	*?*
Word classes		pronoun	auxiliary	*FUT*	infinitival verb	adjective	punctuation
	auxiliary	pronoun	adverb	*FUT*	infinitival verb	pronoun	punctuation
POS tags			PNP	*FUT*	VVI	AJC	PUN
	VBB	PNP	AV0	*FUT*	VVI	PNP	PUN

Table 9.1. *Examples of collocation patterns.(Adopted from Study III. See Burnard 1995 for a list of the POS tags)*

9.2 Collocation with word-classes

Collocation patterns where the collocates are not individual lexical items or lemmas but grouped according to their particular word-class say something about the syntactic contexts where a node, in this case an expression of future, can be found. Similarities in collocation patterns may suggest that the expressions are variants that can be used interchangeably, while dissimilarities indicate that different expressions are preferred in certain syntactic contexts. The aim of my studies has been to compare the collocation patterns of the expressions of future to see what similarities and differences there are.

Even without examining any collocation data it is obvious that there is one major difference between the expressions of future in my studies. The expressions *going to* and *gonna* are used with the auxiliary *be* (the auxiliary is occasionally omitted) while the three modal auxiliary verbs (*will, 'll, shall*) do not take this auxiliary. Thus the syntactic context is different by default; the auxiliary *be* is expected to be found preceding instances of *gonna* and *going to* to a considerable extent. This, naturally, needs to be kept in mind when the collocation patterns are compared.

Study III identifies the most frequent word-class collocates found with the expressions of future in positions -3 to +2 in the spoken part of the BNC. The study shows that there is a considerable degree of similarity between all five expressions, in particular with respect to the kind of items found in the positions closest to the nodes: they are to a very high extent preceded by pronouns and nouns and followed by verbs in the infinitive. It is interesting to examine how the collocation patterns vary between the expressions. As the sections below and my Studies III and IV illustrate, there are both similarities and differences. For example, *will* seems to differ from the other expressions in a number of ways, and *going to* and *gonna* are very similar with regard to their collocation patterns.

9.2.1 Collocations with infinitives

The most frequent collocate of all the expressions of future is the infinitival verb. The proportions and positions of these verbs differ between the expressions. As shown in Study III:182, the proportion of infinitival verbs in position +1 ranges from 43% to 97% (spoken part of the BNC).[53] If infinitival verbs in position +2 are included, the proportions are more similar: between 82% and 99%, as illustrated in Figure 9.1 (raw frequencies in Table A.17, Appendix B).[54]

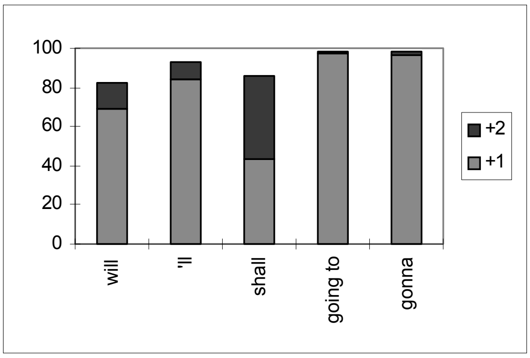


Figure 9.1. Proportions of FUT followed by verbs in the infinitive in positions +1 and +2 (BNC Spoken part).

In Figure 9.1 it can be seen that the expressions *going to* and *gonna* are very similar in the extent to which they are followed by infinitival verbs. The proportions of the two expressions that are immediately followed by infinitival verbs are larger than for the other expressions, well over 90%. It is also clear that *going to* and *gonna* are primarily found with the infinitival verb in position +1. Thus their collocation patterns are less varied than those of *will, 'll,* and *shall*. The constructions *will, 'll,* and *shall* follow less strict patterns, and also precede collocates other than infinitival verbs. This is most pro-

[53] In the BNC, infinitival verbs have four different tags: VBI for *be,* VHI for *have,* VDI for *do* and VVI for all other verbs. All these are included in the concept of 'infinitival verbs' as used here.

[54] I am aware that combing the frequencies for positions +1 and +2 may not reflect perfectly the proportions of different main verbs used with the FUT since this would include collocations where a second infinitival verb in position +2 follows the main verb in position +1. Inspection of a sample of concordance lines suggests that the number of such instances is negligible.

nounced for *will*. The expression is followed by infinitival verbs in position +1 or +2 in about 83% of all cases (69% + 14%), which is less than with the other expressions. Other frequent word-classes following *will* are pronouns and adverbs, in sentences such as (1) and (2):

(1) and you say *will* **you** come back again? (BNC D91 186)
(2) I *will* **always** help you in your future life. (BNC F72 685)

Shall is followed by infinitival verbs to a similar extent as *will* (85%). A major difference between the two expressions is, however, that infinitival verbs collocating with *shall* are equally frequent in positions +1 and +2. *Shall* differs from all the other expressions in this respect. In cases where the infinitival verb is found in position +2, it is usually preceded by a personal pronoun, which functions as the subject, as in (3) (see also Study IV:42):

(3) *Shall* **we try** that then? (BNC F7C 251)

The proportion of pronouns in position +1 is actually slightly larger than that of infinitives in position +2. This can be explained by the fact that *shall* is used to a considerable extent in sentence-final positions without an (overt) infinitive, in examples such as (4).

(4) Let's ask somebody else *shall* **we?** (BNC F7U 589)

Will is used more than *shall* with an intervening adverb between the FUT and the infinitive (8% and 4% respectively), as illustrated in (5).

(5) We *will* **never** *know* how that would develop (BNC F7T 184)

Around 9% of the instances of *'ll* are found with an adverb between the node and the infinitival verb, which makes it similar to *will* in this respect. *Going to* and *gonna* are only rarely followed by anything but infinitival verbs but they occur with an adverb between the auxiliary and the node in 7% and 6% of all instances, as in (6).

(6) In this age the price of disunity *is* **evidently** *going to* be prohibitive.
(LOB G71 17-18)

9.2.1.1 Collocations with certain frequent infinitives

All expressions of future are used with infinitives to a very high degree, which is not unexpected considering that they are auxiliary or semi-auxiliary verbs. Study III looks closer at the choice of infinitives by examining the use of very frequent infinitival collocates. The infinitives are those of the primary verbs *be, do, have* and the ten most frequent lexical verbs found with each of the expressions in positions +1 and +2 in the spoken part of the BNC. Table 9.2 lists the ranking order of the verbs (1= most frequent, 2= second most frequent, etc).

Table 9.2. *Infinitives: ranking of verbs found among the 13 most frequent infinitival collocates with at least one expression of future in the spoken part of the BNC.* ***Bold typeface*** *denotes the verbs that are found among the 13 most frequent items for all five expressions.*
* when two or more items have the same raw frequency they are all given the same ranking number

	will	***'ll***	***shall***	***going to***	***gonna***	**Overall in BNC spoken**
be	1	1	1	1	1	*1*
have	2	2	2	3	2	*2*
do	3	5	4	2	3	*3*
get	5	3	7	4	4	*4*
go	4	4	3	5	6	*5*
take	7	8	11	7	7	*11*
give	8	7	8*	9	10	*16*
say	10	12	5	6	5	*7*
come	6	9	17	10	9	*13*
know	11	19	20	54	40*	*6*
see	9	6	12	14	21	*8*
make	12	14	13*	11	11	*14*
need	13	17	21	18	26	*24*
tell	14	10	10	20	14	*17*
find	16	12	29*	24	29*	*19*
happen	17	57	-	12	13	*44*
put	22	11	6	8	8	*15*
ask	27	18	16	13	12	*23*
start	28	24	13*	19	19	*28*
sell	73	38	8*	79	65	*64*

Table 9.2 shows that there is considerable variation between the expressions with regard to the most frequent infinitival collocates. Of the verbs that are examined, eight are found among the 13 most frequent infinitival collocates with all five expressions (*be, have, do, get, go, take, give, say*). Two verbs, *make* and *come,* are ranked under 13 for one expression only, while five (*know, need, find, start, sell*) are among the most frequent infinitival collocates for only one expression.

The list of verbs collocating with the expressions of future can be compared to the frequency distribution of infinitives in the corpus overall (see column 'Overall in BNC spoken'). It can then be concluded that even though some of the collocates are also amongst the most frequent infinitives in the corpus as a whole, the expressions do not seem to collocate only with these. Items that are frequent in the corpus but not found with the expressions of future are *think* (9), *want* (10) and *like* (12), while *give* (16) and in particular *happen* (44) are proportionately much more frequent with the FUT than in the corpus as a whole.

Figure 9.2 is based on the relative frequency of the collocations given in Table 9.2 (percentages of the total number of occurrences of the FUT that collocates with the particular infinitive. Raw frequencies in Table A.19, Appendix B). The figure shows clearly how different the verb *be* is compared to other infinitival collocates as far as the frequency is concerned. The collocational patterns with be are examined further is Section 9.2.1.2.

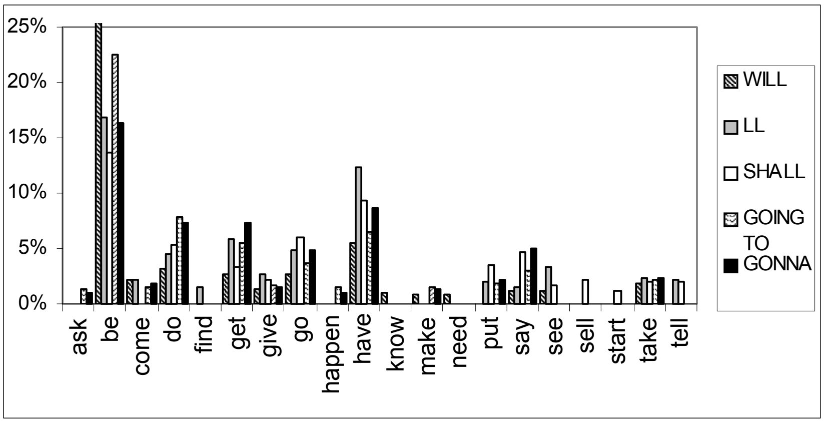


Figure 9.2. Most frequent infinitives collocating with FUT in the spoken part of the BNC

The figure shows the variation in the ways in which the different expressions collocate with different verbs. It is notable, for example, that only *shall* collocates with *sell* to any degree. The explanation for this is easily found when the texts are examined. It turns out that this collocation is specific to a particular context. All but three of the 60 instances of *sell* following *shall* (window +1 - +2) are found in two texts captured in Christie's auction rooms; texts HUR (auction of mechanical music) and HUS (auction of oriental ceramics and works of art), as exemplified in (7).

(7) I *shall sell* then at ninety pounds, any more, ninety five ninety five on my left, now any more at ninety five pounds? (BNC HUR 144)

Shall also collocates with *go* to a higher degree than the other expressions of future. As is often the case with *shall,* these collocations are, to a considerable extent, instances of the expressions of future found with a first person personal pronoun preceding the infinitive. The infinitive *go* is used with *shall* both in the literal sense to denote physical movement (as in 8) and in more figurative meaning, to start something (9). It is frequently used in combination with two infinitives, as in examples (10–11).

(8) Where *shall* we **go**. (BNC JSN 709)
(9) *Shall* I **go** first? (BNC JT5 694)
(10) So *shall* we **go** and **sing** a carol? (BNC G4R 269)
(11) *Shall* I **go** and **ask** her? (BNC G4X 1688)

Collocations with the different uses of *go* are also found with the other expressions of future, even with *going to,* as in (12–13):

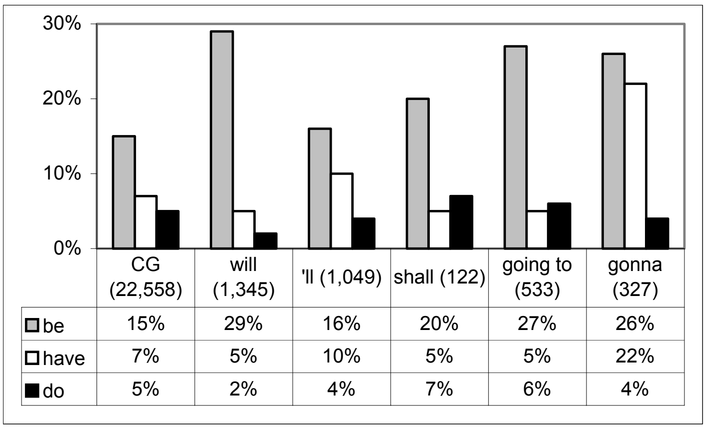
(12) Now who's *going to* **go** next door to get the gas switched on? (BNC FMB 73)
(13) We're *going to* **go** on to the effects of chilling and what damage does that do? (BNC F8L 169)

Will differs from the other FUT in that it collocates more with *be* than any of the other expressions. More than one out of four (26%) of all instances of *will* in the spoken part of the BNC are followed by the verb *be*. *Going to* is also used with *be* to a considerable degree (22%) while the proportions for the other expressions are smaller (*'ll 17%, shall* 12%, *gonna* 16%). This frequent collocation with *be* is not a feature of the BNC corpus only. Table 9.3 provides information about the proportion of *will* collocating with infinitival *be* in five other corpora (see Study IV for details).

Table 9.3. *Proportion of* will *collocating with infinitival be in five corpora of British English (based on Study IV:38)*

Corpus	**Proportion of *will* collocating with *be***
LOB (written from 1961)	34%
FLOB (written from 1991)	34%
LLC (spoken)	27%
Sampler CG (spoken context-governed)	29%
Sampler DS (spoken demographically sampled)	21%

The table shows that the three spoken corpora are similar to the spoken part of the BNC with regard to the proportion of infinitival *be* collocating with *will.* It is not surprising that Sampler CG and Sampler DS are similar to the spoken part of the BNC, considering they are a subset of the 10 million word component (0.5 million words each). The whole spoken part of the BNC consists of roughly 60% CG data and 40% DS data. Extrapolating from the figures in Table 9.3 produces a proportion very similar to that in the full component: (0.6*29%) + (0.4*21%) = 25.8%. The LLC is a spoken corpus (0.5 million words) that has proved to be different from the others in several aspects (see Chapters 3, 4 and 8). In this particular case it displays a pattern similar to that in the other spoken corpora. The written LOB and FLOB corpora also contain large proportions of *will* + *be*: 34% in both cases. The proportion of *be* collocating with *will* is in every case considerably larger than the proportion of *be* among the infinitival verbs in the corpora. The same is not found for the primary verbs *have* and *do,* as illustrated in Figure 9.3 (data from the Sampler CG corpus).



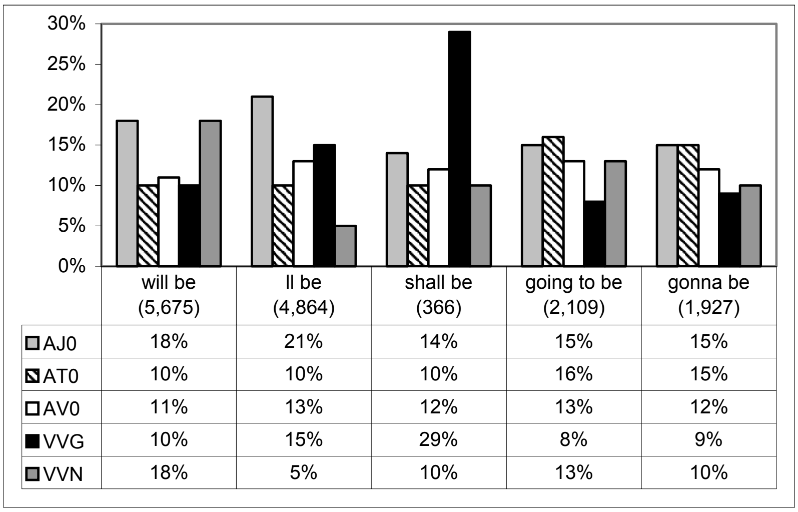
	CG (22,558)	will (1,345)	'll (1,049)	shall (122)	going to (533)	gonna (327)
▨ be	15%	29%	16%	20%	27%	26%
□ have	7%	5%	10%	5%	5%	22%
■ do	5%	2%	4%	7%	6%	4%

Figure 9.3. Expressions of future collocating with *be*, *have* and *do* in the CG component of the BNC Sampler (percentages of the raw frequencies of each expression)

Figure 9.3 shows that in the CG component of the BNC Sampler, 15% of the 22,558 infinitives are *be,* 7% are *have* and 5% *do*. The proportions of *have* and *do* among the infinitival collocates of the FUT vary and are in some cases larger than those in the corpus as a whole, in other cases smaller. The proportion of *be* collocating with the FUT, however, is in every case larger than that found in the corpus as a whole (with the possible exception of *'ll*). This suggests that there is a particularly strong collocation between the infinitive *be* and the expressions of future. In order to investigate this specific collocation in more detail, a special case study was devised.

9.2.1.2 Case study: FUT + *be*

As pointed out above, the infinitival form of *be* is a very common collocate of the expressions of future. The most frequent patterns including an expression of future, *be* and one more item are presented in Figure 9.4. The instances were found in the spoken part of the BNC.



	will be (5,675)	ll be (4,864)	shall be (366)	going to be (2,109)	gonna be (1,927)
▒ AJ0	18%	21%	14%	15%	15%
▨ AT0	10%	10%	10%	16%	15%
□ AV0	11%	13%	12%	13%	12%
■ VVG	10%	15%	29%	8%	9%
▓ VVN	18%	5%	10%	13%	10%

Figure 9.4. Word-class tags most frequently found following instances of FUT+*be* in the spoken part of the BNC (percentage of the total number of occurrences of the different FUT followed by *be*). AJ0 = Adjective (general or positive), AT0 = Article, AV0 = General adverb, VVG = The *-ing* form of lexical verbs, VVN = The past participle form of lexical verbs, (A detailed description of the tags used can be found in (Burnard 1995))

FUT + *be* + VVG

Items tagged as VVG[55] are frequent collocates of FUT+*be*. About 10% of the occurrences of *will be* in the spoken part of the BNC are followed by words such as *going, coming, looking, talking*, etc., forming what is also known as the progressive infinitive. The proportions are slightly lower in FLOB and LOB, 8% and 5% respectively. Two examples are given below (14 and 15).

(14) Yeah Erm sixteen two we *will be discussing* that draft. (BNC F7A 194)
(15) Though I think that that *will be happening* in the next cycle.
(BNC F7V 795)

The proportions of VVG following *be* are even higher for *'ll* and *shall*, 15% and 29% respectively. The *gonna* and *going to* expressions are also found followed by *be* + VVG, but not as frequently as *'ll* and *shall.*

The relatively high proportion of the progressive infinitive found with *shall* warrants further scrutiny. When the instances are examined it can be seen that although they are particularly frequent in some texts, they are as a whole distributed across a number of different texts. Thus they are not a

[55] VVG= "The *-ing* form of lexical verbs" (Burnard 1995).

feature that is specific to a particular context, as is the case with the frequent collocation *shall* + *sell* which is found primarily in the BNC auctioneering texts, as discussed above (Section 9.2.1.1). What should be noted, however, is that although the Context-governed component only contains around 35% of the instances of *shall,* the majority of the instances of *shall* + progressive infinitive (>80%) are found in the CG texts (CG texts constitute about 60% of the spoken part). As described elsewhere (Chapter 3), the CG texts are generally understood to be from "more formal encounters" (Aston and Burnard 1998:31). It could then be suggested that the *shall* + progressive infinitive is a feature of more formal texts, since that is where the bulk of the instance of *shall be* +VVG appear. What should be taken into account in this context, though, is that not all spoken texts in the first release of the corpus are marked as either CG or DS. About 23% of the instances of *shall* + progressive infinitive are found in texts not marked CG or DS. This indicates why it is difficult to draw further conclusions regarding the influence of the type of text in this case.

As noted in Chapter 2, some scholars choose to treat instances of FUT (or at least *will/shall*) with the progressive infinitive as a separate construction rather than including them as instances of the five FUT as I have chosen to do. The section above has shown that removing the instance of FUT+ *be* + VVG from the general count of FUT would not significantly have altered the overall distribution. The expression most affected would be *shall,* but even there the numbers are too low to offer scope for detailed analysis. Furthermore, there does not seem to be any correlation between the extent to which the different FUT collocate with *be* and the extent to which they are used with the progressive infinitive.

FUT + *be* + VVN/AJ0 (past participle/adjective)

In Figure 9.4 it can be seen that of the 5,675 instances of *will be*, 18% are followed by items tagged VVN (the past participle form of lexical verbs), in sentences such as (16) and (17):

(16) Mercury *will be* **mixed** with the sludge to separate the gold. (BNC HE4 59)

(17) During this year and next, more European directives *will be* **agreed**. (BNC HLY 328)

VVN is also found following the other FUT, but to a lesser extent: 13% for *going to,* 10% for *shall* and *gonna* and only 5% for *'ll*. The low proportion found with *'ll* can possibly be explained by the fact that the *be*+VVN collocation is more frequent in the CG component of the corpus. Since *'ll* is less frequent there, there are few collocations of *'ll* with *be*+VVN. This could also explain why the proportion of *gonna* found in this collocation is lower than that of *going to.*

The tag AJ0 ('general or positive adjective' (Burnard 1995)) collocates with between 14% and 21% of the occurrences of FUT+*be*, such as in (18–20):

(18) Well I think it's *going to be* **essential** isn't it? (BNC F7J 024)
(19) And that *will be* **welcome**. (BNC J43 307)
(20) She *will be* **ecstatic** (BNC J97 591)

The frequencies of VVN and AJ0 should be interpreted with some caution as they may include tagging errors of the kind illustrated in (21). The word *relieved* is in this case tagged VVN although it is an adjective in this context.

(21) You *will be* **relieved** to know I've ordered some new rubber gloves like this and some new goggles for next term. (BNC F77 106)

The reverse (past participle tagged as adjective) can also be found in a number of cases and the portmanteau tag AJ0-VVN, which is used in unclear cases, is also rather frequent. There are naturally cases when it can be difficult to decide if the word is an adjective or a past participle, but many of the instances tagged with the portmanteau tag AJ0-VVN in this version of the corpus are actually easily disambiguated by a human reader. However, it is not feasible to go through all instances of AJ0, VVN and AJ0-VVN manually to establish the success rate of the automatic tagging[56].

FUT + *be* + AT0 (articles)

Another frequent collocate of all FUT are articles (22 and 23).

(22) Is it *going to be* **the** same one, or are you just announcing a squad? (BNC KRT 5564)
(23) The result *will be* **a** low, modern, divan. (BNC D8Y 308)

Articles are more frequent with *be* following *gonna* and *going to* than when found with *will, 'll,* and *shall* + *be*. This seems to be one feature that distinguishes *gonna* and *going to* from *will, 'll,* and *shall*. A similar difference is not found with, for example, VVN or AJ0. As will be further shown below, the expressions of future differ with regard to the extent to which they collocate with nouns. *Will* collocates more with nouns in general, which would suggest that articles, forming a noun phrase with a noun, could be expected to collocate more with *will* than other expressions. It is interesting that this is

[56] This version of the BNC contains over 6.2 million items tagged AJ0, about 1.9 million tagged VVN and 126,000 instances of AJ0-VVN, nearly 8.3 million examples in all. In addition, there are over 2 million instances where a word has been tagged with another portmanteau tag which includes either VVN or AJ0.

not the case for FUT+*be*, but the complexity of potentially influential factors means decisive conclusions regarding this cannot be based on the results of this study alone.

9.2.2 Collocations with personal pronouns

All the expressions of future are used with pronouns to a considerable degree, albeit in different positions, as illustrated in Figure 9.5 (data from the spoken part of the BNC).[57]

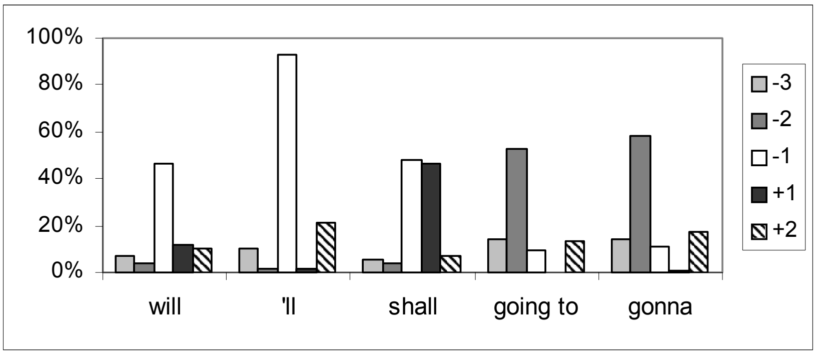


Figure 9.5. Proportions of the expressions of future collocating with pronouns in positions -3 – +2 (spoken part of the BNC)

With the exception of *shall,* the proportion of pronouns is considerably larger preceding the expressions than following them. This is most marked for *'ll*, which is found immediately preceded by a pronoun in more than nine cases out of ten (93%), as in (24):

(24) **You***'ll* see. John will get better*,* **they***'ll* send him home, **he***'ll* meet
some nice girl, (LOB G09 110-111)

This pattern has been discussed extensively in studies dealing with contractions (for example Axelsson 1998, Kjellmer 1998).

A noticeable difference between *gonna/going to* and the other expressions is that the former collocate most with pronouns in position –2. The explanation for this difference is, naturally, that the position immediately preceding the semi-auxiliaries is often occupied by the auxiliary *be,* in examples such as (25).

[57] The term 'pronoun' in this context refers to all items tagged as PNI (Indefinite pronoun), PNP (Personal pronoun), PNQ (*Wh*-pronoun), or PNX (Reflexive pronoun).

(25) Erm, so, **I**'m *gonna* start off by talking about the Local Government
Unit, where it fits in the Council organisation. (BNC D95 419)

Gonna and *going to* are used more with preceding pronouns than *will* and *shall* but less than *'ll*. Figure 9.5 shows that the proportion of pronouns collocating with *gonna* or *going to* in the position –2 is slightly larger than those of pronouns immediately preceding *will* and *shall* but considerably smaller than that found preceding *'ll*.

The expression *shall* displays a different pattern from the other expressions in that pronoun collocates are equally frequent immediately preceding and following the expression. This correlates with the fact that the expression is often used in questions, preceded by its infinitival verb collocate, as described above. That all expressions but *shall* are primarily found with preceding pronouns, suggests that they are used considerably less in tag-questions or other cases where inverted word order is required.

9.2.3 Personal pronouns as subjects

The pronouns collocating with expressions of future may have different functions, as illustrated below where *it* functions both as subject (26) and object (27) used with the expression.

(26) Mm, perhaps **it**'s *gonna* be a big one? (BNC DCH 406)
(27) Who's *gonna* do **it**? (BNC D97 795)

The fact that pronouns found close to the FUT can have different functions explains why the combined proportions of a tag in the various positions can at times add up to more than 100%; they can be both as subjects and objects in the same sentence.

In Study IV, the personal pronouns used as subjects of the FUT were identified and examined closer. Figure 9.6 shows how the proportion of personal pronoun subjects varies between five British English corpora, the written LOB and FLOB, and spoken LLC, Sampler CG and Sampler DS (see Chapter 3 for a description of the corpora). The diagram is based on figures presented in Study IV:40.

The figure shows that all the expressions of future are used with personal pronoun subjects to a considerable extent but that there is noticeable variation in the collocation patterns both between the different expressions and between the corpora, as further discussed below.

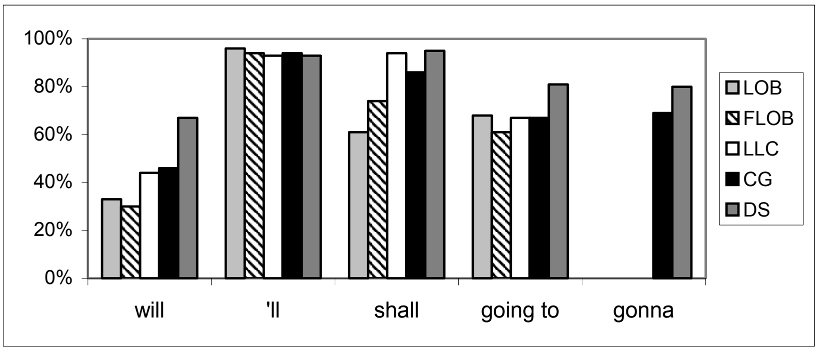


Figure 9.6. Proportions (%) of personal pronoun subjects in LOB, FLOB, LLC, Sampler CG and Sampler DS.

9.2.3.1 Variation between corpora

The proportion of FUT collocating with personal pronoun subjects is higher in the spoken corpora than in the written; 68% of all FUT in the Sampler CG corpus are found with a personal pronoun subject, and the proportion in the Sampler DS is even higher: 84%. There is less variation between the written LOB and FLOB corpora: 48% and 43% of the FUT in the two corpora respectively have personal pronouns as subjects. This difference between the written and spoken corpora is consistent for all expressions but *'ll*. Only *'ll* is found with personal pronoun subjects to a greater extent in the written than in the spoken corpora, even if the difference here is small.

One explanation for the higher frequency of personal pronoun subjects in the spoken corpora could be that pronouns are used more in spoken language: there are more pronouns overall in the spoken texts. For the corpora examined here, the number of personal pronouns varies from around 40,000 per million words (pmw) in the written corpora (41,600 instances in LOB and 39,745 in FLOB) to over 90,000 pmw in Sampler CG and more than 140,000 pmw in Sampler DS (45,565 instances in Sampler CG and 70,434 in Sampler DS). The higher proportion of personal pronoun subjects with the FUT in DS thus coincides with a large number of personal pronouns in the corpus. The relative frequency of personal pronouns is more than three times as high in Sampler DS as in the written corpora, but the proportion of personal pronoun subjects with FUT is less than twice as high so it does seem that this is not the only factor affecting the frequency. Another possible factor could be subject matter.

9.2.3.2 Variation between expressions of future

The proportions of personal pronoun subjects also vary between the five expressions of future (see Figure 9.6). The proportion is considerably lower for *will* than for the other expressions in all corpora. Even though the proportion of personal pronoun subjects of *will* is more than twice as large in Sampler DS as in LOB and FLOB, it is in every case considerably lower than for any other expression. The difference is particularly noticeable when *will* is compared to the contracted variant *'ll.* The expression *'ll* displays a large proportion of personal pronoun subjects overall, from 93% to 96% in the examined corpora, compared to between 30% and 67% for *will*. The fact that *'ll* is used with personal pronoun subjects to such a high degree is not surprising, considering the strong syntactic relationship between the expression and different pronouns (see, for example, Axelsson, 1998). That the collocation personal pronoun + *'ll* is so frequent in both the written and spoken data suggests that this is a feature related to the expression *'ll* and not dependent on the kind of text where the expression is found.

The use of *shall* with personal pronoun subjects varies between the written and spoken corpora to a similar degree as the other FUT. In the spoken material, a high proportion of the expression is used with personal pronoun subjects (86% in CG, 95% in DS). In the written corpora, the proportions are 61% in LOB and 74% in FLOB. *Shall* differs from the other expressions in that it is the only FUT which is found with personal pronoun subjects more in FLOB than in LOB. This finding should be interpreted with some caution. As mentioned above (for example Chapter 5), the frequency of *shall* in FLOB is very low and the instances are often found in specific genres or even in certain texts. The proportion of the personal pronoun subjects is thus based on a low number of instances and any conclusions regarding its use should recognise this.

More important to note is the frequent use of *shall* with a personal pronoun following the expression, with or without an infinitive in position +2, as noted in Section 9.2.1 and illustrated in (28) and (29):

(28) We'll press on, *shall we*? (BNC A0R 1194)
(29) *Shall we* try it? (BNC A73 808)

This is a pattern that is particularly frequent in questions and tag questions, and it seems to be a use of *shall* which has not declined at the same rate as other uses (see discussion of the decrease over time in the use of *shall* in Study IV and Chapter 8).

Going to is used with personal pronoun subjects in about two out of three cases in the LOB and somewhat less in the FLOB. This is more than *will* but less than *'ll.* In the spoken corpora, the use of personal pronoun subjects follow similar patterns for *going to* as for the other expressions and are found more in DS than in CG. In both cases this is less than *'ll* and more than *will. Gonna* is similar to *going to* in the extent to which it is used with personal pronoun subjects in the spoken corpora (no comparison is made with the written corpora where *gonna* is practically non-existent). One explanation for the varying degree to which the expressions take personal pronoun subjects is found in the study of noun collocates. Expressions that have a large proportion of pronominal subjects (such as *'ll*) collocate considerably less with nouns, while for example *will,* which has the lowest proportion of pronominal subjects, collocates more frequently with nouns (see Section 9.2.4).

9.2.3.3 Choice of personal pronoun subjects

It was shown above that all expressions are used with personal pronoun subjects to a high degree, even though the proportions of such subjects can be seen to vary both between different corpora and between the expressions. A closer study of the personal pronoun subjects reveals that there are differences between the expressions also where the choice of personal pronoun is concerned. Figures 9.7 and 9.8 illustrate the proportions of the different personal pronoun used as subject with the FUT in one written (FLOB) and one spoken (Sampler DS) corpus (raw data are given in Table A.13, Appendix B).

It is immediately obvious from Figures 9.7 and 9.8 that *shall* shows a different collocation pattern from the other expressions. *Shall* is used almost exclusively with only two pronouns: *we* and *I.* In the very rare case of a different pronoun being found with *shall* it is in a text that is mimicking old usage (see discussion of the use of *shall* in FLOB, Chapter 8).

(30) Dear Clementina, *you shall* tell us nothing at all if you don't wish to. (FLOB P01 87-88)

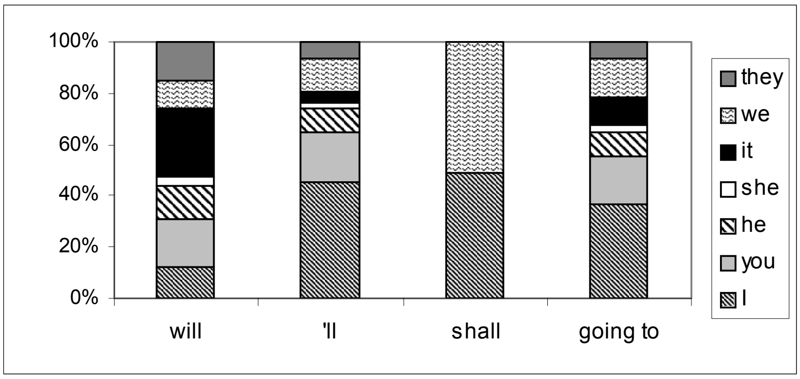


Figure 9.7. Proportions of different personal pronouns used with the FUT in the FLOB corpus (100% = all personal pronoun subjects used with each expression)

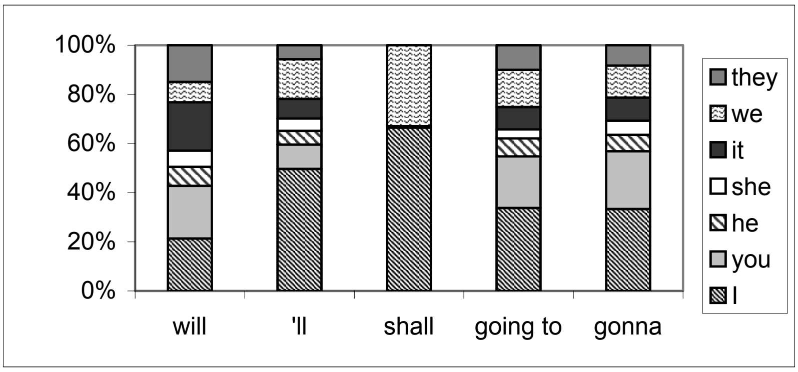


Figure 9.8. Proportions of different personal pronouns used with the FUT in the Sampler DS (100% = all personal pronoun subjects used with each expression)

The other expressions display a more varied choice of personal pronoun subjects. The collocation pattern found with *'ll* is similar to that of *going to* in the written corpus and of *going to* and *gonna* in the spoken. The proportions of first person pronouns are high, even if not as high as for *shall.* There is some slight variation between the *'ll* and *going to* (+*gonna*) expressions in the extent to which they are used with third person pronouns. *It* and in the spoken corpus also *they* are used slightly more with *going to* than with *'ll. Going to* and *gonna* show very similar collocation patterns in the Sampler DS where the proportions of the various pronouns are almost identical (no comparison between *going to* and *gonna* is made with the written corpus, in which there are only a handful of instances of *gonna*). The same applies to the full spoken part of the BNC, as is shown in Study III, where the occurrences of *going to* and *gonna* are compared in more detail.

It was described above how the expression *will* is found with personal pronoun subjects to a much lesser extent than the other expressions. As illustrated in Figures 9.7 and 9.8, it is not only the frequency that differs. The choice of pronoun is another feature which shows dissimilarities. Most apparent is that the proportion of *I* used as subject is much lower for *will* than for the other expressions, while the proportion of third person singular *it* is considerably higher. The proportion of *they* is also higher with *will,* while *we* seems to be used more with the other expressions. The difference between *will* and the other expressions is found in both the written and the spoken corpora. This variation is also discussed in Study IV, where two written and three spoken corpora are included in the analysis. *Shall* differs from the other expression in that it occurs almost exclusively with first person pronouns.

The difference between *shall* and the other expressions is striking where the choice of personal pronoun subject is concerned. Krogvig and Johansson (1984) present some interesting findings with regard to the use of *shall* in the

LOB and Brown corpora. They examined all instances of *will, 'll* and *shall* and discovered that *shall* was found to a similar low extent in both corpora when used with second and third person subjects. Among the instances of *will/shall/'ll* used with a first person subject, however, the proportion of *shall* was larger in the British data; 32.9% compared to 22.5% in Brown (see Krogvig and Johansson 1984 for details). The authors thus concluded that the higher proportion of *shall* in LOB could be explained as a British preference for the expression when used with first person reference.

As mentioned in Chapter 7, the proportion of *shall* is higher in the Indian Kolhapur corpus than in LOB and Brown. In analogy with the results of the Krogvig and Johansson study, it could then be expected that the proportion of *shall* used with first person subjects would be larger as well. However, this does not however turn out to be the case. Among the first person instances of *will/shall/'ll* in the Indian English corpus (334 instances in all), *shall* constitutes about 31%, which is more than in the American data but slightly less than in the British[58]. It must be concluded, therefore, that the high frequency of *shall* in the Indian corpus cannot be explained with reference to the subject of the expression alone. The distribution of the expression across the corpora suggests that genre and certain genre-related features are important for the distribution of the expression *shall*. This variation was examined in more detail in Chapter 5.

9.2.4 Collocations with nouns

Figure 9.9 shows how the expressions of future collocate with nouns in the spoken part of the BNC, both with regard to the relative frequency and position of the collocations.

Going to is preceded by nouns in positions -3 and -2 to a similar extent, about 9% in both cases, as in (31).

(31) The problem is *going to* be the sheer weight of administration of schemes like this … (BNC KRT 6482)

The proportions for *gonna* are slightly lower, 7% in position -3 and 5% in position -2. This is one of the few cases where there is a noticeable difference between the two expressions. *Shall* collocates with nouns to a similar extent in the three positions preceding the expression but is rarely followed by items from that word class.

[58] The first person use with *will/shall/'ll* was calculated by retrieving the instances of the expressions that are found with *I* or *we* at a distance of maximally 10 words on either side of the head word. The concordance lines were then manually scanned and irrelevant instances discarded. The potential error (instances where a first person subject is not expressed as a first person pronoun found within this span) is considered too slight to influence the overall picture.

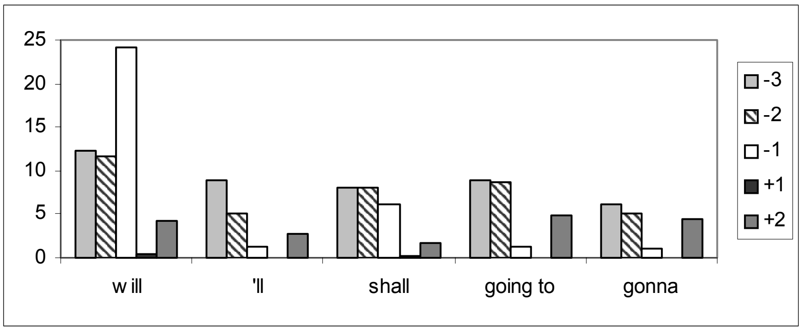


Figure 9.9. Collocations noun+FUT in the spoken part of the BNC (collocations per 100 instances of the FUT)

As far as *will* is concerned, the proportion of nouns used with the expression is larger than for any of the other expressions, in particular in position -1. Almost one out of four instances of *will* are found immediately preceded by a noun, in examples such as (32):

(32) I shouldn't think his **car** *will* start. (BNC KBK 492)

Will is also found followed by nouns, but to a much lower extent, similar to the other expressions.

The proportion of *'ll* immediately preceded by a noun is small, only around one per cent, in examples such as 33:

(33) **Daddy**'*ll* climb right up again, he always does. (BNC G0S 1673)

The proportions of nouns collocating with *'ll* in positions -3 and -2 are larger, 9% and 5% respectively. The expression is similar to the other expressions in this respect.

Like the pronouns described above, nouns can have different functions in the sentence: they can be subjects and objects. A closer look at the instances of nouns in positions -2 and -3 reveals that these nouns are only rarely the subject of the FUT (e.g. example 34).

(34) … the first part of the **talk** we*'ll* keep the lights on (BNC F71 10)

As illustrated in Figure 9.9, nouns are less frequent following the expressions of future than preceding them. Nouns are particularly rare in the position immediately following the FUT. This correlates with the findings above which showed that the expressions are rarely found without an infinitive in this position (see Figure 9.1). Even though the expression *will* has the largest proportion of collocating nouns in position +1, there are only about one hundred examples in all in the whole spoken part of the BNC, for example:

(35) So it should be quite an intimate performance, *will* **people** come along (BNC KRT:3749)

When nouns are found in position +2, they are often the object of the sentence, as illustrated in (36–38).

(36) Are you *going to* do **photographs**? (BNC F7R 426)
(37) And are there *going to* be **flags** in the street? (BNC HUV 975)
(38) …a squirrel *will* eat **berries**… (BNC J3Y 159)

Some instances of nouns following the expression of future are also the result of incomplete sentences and other features of spoken discourse, or possibly an example of accidental omissions in the transcriptions (see for example 39).

(39) I think I'm *gonna* **prawn** cocktails first for Christmas day (BNC KBE 6053)

As the collocates are identified on the basis of the part-of-speech tagging, it is inevitable that errors in the tagging result in errors in the identification of the collocates, in examples such as (40) where *fork* is tagged as a noun.

(40) Well he, are you *gonna* **fork** out three hundred quid? (BNC KP1 879)

However, spot checks suggest that the number of such errors is too low to affect the overall results.

9.3 *Going to* and *gonna*

Throughout the sections above it has been apparent that the expressions *gonna* and *going to* display exceptionally similar collocation patterns but that they generally differ considerably from the other expressions. The same similarity is not found between *will* and *'ll*. The patterns with *gonna* and *going to* have been investigated further in Study III, where the emphasis is on the comparison between the two expressions in the spoken part of the BNC. It is shown that *gonna* and *going to* are used in very similar lexical and syntactic environments. A strong syntactic relationship between the auxiliary *be* and *gonna* and *going to* is attested. Forms of *be* are found in the two positions preceding both expressions (-1, -2) to a far greater extent than in the corpus as a whole. This is not surprising. The relationship is so strong that the auxiliary is at times considered part of the expression, as can be seen when the expressions, especially the full variant *going to,* are discussed in grammars or listed in dictionaries, as in *Collins Cobuild English Grammar* (41):

(41) "You can use 'be going to' instead of 'will' to say that ... "
(*Collins COBUILD English grammar* 1990: 4.237)

Even though the auxiliary is missing in constructions with *gonna* to a somewhat larger extent than in constructions with *going to*, forms of *be* are used in the vast majority of all instances of both expressions. In the spoken part of the BNC, there are 11,914 instances of *gonna* and 9,441 examples of *going to* (where *to* is tagged as infinitival marker).[59] Some 92% and 96% of the expressions respectively are preceded by forms of *be* in one of the positions -3 to -1. Most of the cases with missing auxiliaries are examples of false starts, incomplete sentences and repetitions which are frequently found in normal spoken discourse, as exemplified in (42) and (43).

(42) ...somebody or other's gonna take it over, *gonna* try and get
(BNC KCS 2341)

(43) I wonder when er *going to* have the erm (BNC KNS 394)

A further similarity between *going to* and *gonna* is the proportions of different forms of *be* preceding the variants. The tag VBB (used for finite or base forms of *be,* such as *are*), for example, is found immediately preceding 35.8% of all instances of *gonna* and 35.5% of *going to* (44–45).

(44) We **are** *going to* do some miming this afternoon. (BNC F8M 605)

(45) But what we **are** *gonna* do is, is give out your study guides and…
(BNC KBU 440)

Past tense forms of *be* are found immediately preceding 11.5% and 10.9% of all instances of *gonna* and *going to* respectively (46), while marks of punctuation (tagged PUN) occur in position +2 with 8.4% and 8.7% of the occurrences of the expressions and in position +3 with 9.7% and 9.0% (47).

(46) I **was** *going to* say there was a flourishing black market
(BNC D8Y 142)

(47) Tell them what you're *going to* do**.** (BNC KS6 168)

[59] These are the frequencies obtained from the first version of the BNC with the original index, which has been shown to be erroneous. As discussed in Study III, the figures differ from those obtained from the corpus with a new version of the index.

The same kind of similarity cannot be found between the other expressions of future (see Study III for details).

A large proportion (187/546=34%) of *gonna* in the written part of the BNC co-occurs with words that could be classified as 'slang', colloquial, or non-standard, or in sentences/constructions that do not adhere to generally accepted rules of what is considered standard language (as defined by reference grammars). Examples include passages written with non-standard orthography to signal spoken or dialectal usage, such as in (48):

(48) Now yer know what a cowson that Frank is. 'E told 'er ter piss orf out of it in no uncertain terms an' Maudie told 'im she was *gonna* send 'er ole man round ter sort 'im out. (BNC EA5 694)

A frequent slang word co-occurring with *gonna* is *fuck*, and derivations of that, either spelled out or just indicated, as in (49).

(49) The fact is, something really f--;ed up is *gonna* happen in one year, two years into your relationship; someone's *gonna* f-- you over or you're *gonna* f-- them over. (BNC CK6 515)

Double negations (50) and use of non-standard verb forms (51–52) also occur, as well as instances where the auxiliary has been omitted (53).

(50) "There ain't *gonna* be no war." (BNC FPS 1429)
(51) But they *gonna* get took home soon." (BNC AC5 1827)
(52) "I thought we was all *gonna* get nicked!" (BNC CR6 3636)
(53) But everything *gonna* be OK. (BNC ATE 1461)

The existence of such non-standard features co-occurring with *gonna* further strengthens the impression that the variant is perceived as a colloquial feature and is used to illustrate spoken or informal language in written text. There are hardly any instances of *going to* co-occurring with non-standard features in any of the categories. This seems to be a main difference between the variants.

9.4 Summary

This section has shown that collocations where expressions of future co-occur with infinitives are very frequent. As pointed out, this is not particularly remarkable considering that the expressions are auxiliary and semi-auxiliary verbs. Focus has therefore been placed on comparing the different expressions to see which infinitives are found in the collocations, where these are placed and to what extent these patterns vary between the five expressions.

The study has shown that there are some differences between the expressions of future with regard to how they collocate with verbs. The expressions *gonna* and *going to* are very similar overall. They are found with the same kinds of collocates to similar degrees, and that is the case for word class collocates as well as collocating individual lexical items. The *will/'ll/shall* expressions are at times found to be similar, in particular when compared to *gonna* and *going to,* but there is a greater degree of variation between them. *Will* in particular displays a different collocational pattern. This difference is more noticeable in the extent to which the expression collocates with items of particular word classes than in the choice of lexical items in that class. *Will* is, for example, followed by an infinitive in about 80% of all cases, while the corresponding proportion for *gonna* and *going to* is almost 100%. Of these instances of *gonna* and *going to,* the infinitives are almost exclusively found in position +1. Infinitives collocating with *shall* are found in positions +1 and +2 to similar degrees.

The choice of infinitive verb used with the expressions does vary, but to a large extent the same verbs are found among the most frequent collocates for all expressions. Mostly these verbs are also very frequent in the corpus as a whole, but there are some exceptions. Private verbs such as *think* and *feel* are less frequent with expressions of future than in the corpus generally, while *happen* is considerably more frequent when collocating with expressions of future than otherwise (see Table 9.2).

The results of the case study of FUT+*be* fit well with my other results concerning the co-occurrence patterns of the FUT. *Gonna* and *going to* are generally more similar, on the one hand, than *will, 'll,* and *shall* on the other. Within the group of *will, 'll,* and *shall,* all three expressions seem to display collocational patterns that at times differ considerably from the other FUT in that group.

Overall, the expressions display similar patterns where the collocations with personal pronoun subjects in different corpora are concerned. Personal pronoun subjects are less frequently found with the FUT in the written corpora than in the spoken (with the exception of *'ll*). *Will* is always used least with personal pronoun subjects, and *'ll* most. *Going to* is used with personal pronoun subjects more than *will* but less than *'ll* and *shall. Gonna* displays a collocational pattern almost identical to that of *going to* where the proportion and distribution across the spoken corpora are concerned.

As regards the choice of the subject pronoun, *shall* is almost exclusively found with first person pronouns (*I* and *we*). *Will* is used less than the other expressions with *I* but considerably more with *it* and *they. Gonna* and *going to* are, once again, very similar in the proportions to which they are used with different personal pronoun subjects in the DS corpus. As stated in Chapter 1, I have included all instances of *shall* in this study, irrespective of the degree of futurity they express. One argument raised against this practice could be that *shall* would have future reference only when used with first person subjects. This chapter has shown that excluding any example that has other than a first person subject would make no difference to the overall pattern since the frequency of second or third person uses with *shall* is so very low.

Nouns do not collocate much with the expressions of future. They are most frequently found with *will,* and then often immediately preceding the FUT. If found in positions further from the expression itself (-3, -2 and +2, +2) the nouns do not normally function as the subject.

Throughout this chapter, it has been shown that the collocational patterns of *gonna* and *going to* are very similar. The other expressions are found to display similar patterns at times but there is no overall similarity between them. This suggests that as far as the linguistic association patterns are concerned, *gonna* and *going to* can be seen as interchangeable. The only major difference between the expressions is the extent to which they collocate with non-standard features in the written language samples. *Gonna* is very rare in the written texts, and primarily found in quotes or speech-like contexts.

A further important observation in this chapter is that even the most frequent of certain collocational patterns can at times be too infrequent to give reliable indications of the nature of the linguistic association patterns of the expressions of future.

10. Summary and conclusions

10.1 Introduction

Biber suggests that corpora are useful for examining what he calls association patterns or “the systematic ways in which linguistic features are used in association with other linguistic and non-linguistic features” (Biber 1996:173). As stated in the introduction to this Summary (Chapter 1), this thesis examines the association patterns of five expressions of future (FUT): the auxiliary and semi-auxiliary verbs *will, ’ll, shall, going to* and *gonna.* In Chapter 2 some previous research is presented as are some other means of expressing future reference in English. Chapter 3 deals with issues relating to the use of corpus linguistics as an analytical framework. The results of my studies are presented in Chapters 4-9, reporting on non-linguistic (Chapters 4-8) as well as the linguistic association patterns (Chapter 9). The current chapter summarises the main results of my examination of the association patterns (Section 10.2) as well as comments on the analytical framework (10.3) in addition to presenting the concluding remarks (10.4).

10.2 Association patterns

10.2.1 *Will*

Although the proportions of the expressions of future vary considerably between different categories of text, *will* is overall the most frequent of the FUT in my studies. The relative frequency of the expression is similar across all text categories, but the proportion is considerably higher in the written corpora (on average 70% or more, compared to around 30% in the spoken corpora). Within the written corpora, *will* is found to be more frequent in the Informative hyper-category than in the Imaginative. When different kinds of spoken data are compared, it turns out that the expression is more frequent in the more formal Context-governed component (CG) of the BNC and BNC Sampler than in the conversational, less formal Demographically Sampled

(DS) component. The DS corpus is the only text category which does not see *will* as the most frequent of the five examined expressions.

The collocational patterns of *will* differ from those found with the other expressions in a number of ways. Generally, *will* seems to display a somewhat more varied pattern. It is, for example, the only expression which is used with a noun subject to any degree. When used with personal pronouns as subject, the choice of pronoun is more varied than for the other expressions. In collocations with infinitives, *will* displays a pattern where the infinitive is the most frequent class of word following the expression, just like the other expressions. The proportion of instances not following this general pattern is, however, larger for *will* than for the other expressions.

10.2.2 *'ll*

It is generally understood that *'ll* is the contracted form of *will* (see Chapter 1). My studies have shown that the expression is often found in patterns directly reversed to those of *will.* The frequency and proportion of the *'ll* expression varies considerably between different kinds of text. The expression is relatively rare in the Informative genres but much more frequent in the Imaginative. It is also found in the spoken corpora more than in the written. In the informal DS component, *'ll* is the most frequent of the five expressions I examine, even more frequent than *will*. That *'ll* appears to be a feature of spoken or speech-like language is further emphasised by the high proportion of the instances of the expression found in quotes.

The expression *'ll* is less frequent in the Indian English texts than in the comparable British and American corpora examined here. The expression is particularly rare in the Informative genres in the Indian corpus. There are no noticeable differences between the other two regional varieties as far as where the use of *'ll* is concerned.

Although it has been suggested that the use of contracted forms in print is increasing over time, there is no overall difference in the proportion of *'ll* between the earlier LOB and later FLOB corpora which could provide compelling evidence for an increased use of the expression in this context.

As far as the linguistic association patterns with *'ll* are concerned, my studies have shown that the expression is similar to the other FUT in that it is found followed by infinitives to a high degree. It is almost exclusively used with personal pronouns as subjects, even more so than the other expressions of future. Where the choice of pronoun is concerned, *'ll* displays a pattern more similar to that of *going to/gonna* than *will*. This may suggest that the expression is not exclusively a spoken or speech-like variant of *will* but that there are some areas where the two differ and the two expressions can be seen as independent variants.

10.2.3 *Shall*

Overall, the expression *shall* is used less than both *will* and *'ll* in all three regional varieties of English examined here. The proportion of the expression varies considerably within and between the three corpora, being most frequent in the Indian corpus and least frequent in the American data. Similarly to the other expressions, the use of *shall* varies between the hyper-categories. The overall frequencies seem to suggest that the expression is used more in the Informative hyper-categories. A closer inspection of the instances, however, reveals that the expression is not evenly distributed across the hyper-categories but occurs primarily in certain genres or even individual texts. A similar pattern is found in the spoken data.

Shall is the only expression which shows a consistent decrease over time, being found less in the later British and American corpora alike. The proportion of the expression is also smaller in the later BNC Sampler corpus than in the LLC.

As far as the collocational patterns with *shall* are concerned, some interesting differences are found when the expression is compared to the others included in my studies. Like the other expressions, *shall* is used with infinitives to a very high degree. It differs from the other expressions, however, in that the infinitives are not primarily found immediately following the expression (position +1) but occur equally often in position +2. Moreover, the expression is almost exclusively used with first person pronouns as subjects.

My results thus conform to the idea that *shall* is becoming rare in Present-day English. Moreover, when the expression is used, it is in certain kinds of texts (such as legal texts, auctioneering scenes, contexts mimicking old-fashioned usage) as well as in particular collocational patterns: with a first-person subject and frequently in questions or tag-questions with inverted word-order. This makes the expression different from the others in my studies. The overall low frequencies of the expression mean that is it at times difficult to examine patterns of usage in any detail.

10.2.4 *Going to*

Considering the amount of research that discusses the *going to* expression and compares it to *will*, it is surprising to find that the proportion of *going to* in written language is only around 5% of the combined frequencies for *will, 'll, shall* and *going to/gonna.* The expression is much more frequent in spoken language but not even there is it nearly as frequent as *will/'ll.*

The use of *going to* differs somewhat between the regional varieties of English examined here, but not to a great extent. The Indian corpus displays a slightly lower proportion of the expression overall, while the difference between the British and American data is more difficult to evaluate since there are inconsistencies between the different hyper-categories. It is overall the case that the difference between the two hyper-categories in one corpus is larger than the difference between regional varieties.

Like the other expressions, *going to* is used to varying degrees in different kinds of text. It is more frequent in the Imaginative categories, thus more similar to *'ll* than *will*. In the spoken corpora, the expression is less frequent in the less formal DS component than in the CS component, which at first may seem surprising against the background of the patterns found for *'ll*. The simple explanation is that the variant form *gonna* is preferred to *going to* in the more informal context. That *going to* and *gonna* are indeed variant forms of one expression is obvious when the collocational patterns are examined. The expressions are remarkably similar with regard to their word-class collocates as well as lexical items with which they co-occur. It is frequently the case that not only are the collocates the same for both expressions but the proportions of these collocates are also almost identical.

However, *going to* and *gonna* are not similar in every respect. When factors related to the speakers are taken into account, there appears to be some variation between the two expressions in that the contracted variant is preferred by younger people and people from certain social groups. There is also an indication that male speakers use *gonna* more than women.

10.2.5 *Gonna*

Not surprisingly, the most marked difference between *gonna* and the other expressions in my studies is that *gonna* is almost only found in spoken data. When found in the written corpora, such as in the BNC, the expression is almost exclusively used in speech-like contexts, e.g. quotes, and is often found together with features of non-standard language or non-standard orthography.

In the spoken corpora, *gonna* is used more in the less formal DS component of the BNC and BNC Sampler, where it is more frequent than *going to.* In the CS component of more formal discourse, *going to* is more frequent. In the LLC, the expression is almost non-existent. It is not possible to say whether this difference is due to differences in transcription practices or related to other factors, such as differences in the time when the texts in the corpora were produced, in the speakers recorded or in the contexts in which the texts were captured.

It has been suggested that *gonna* is an expression used more in American English. Although my studies do include two American English corpora, it has not been possible to evaluate this claim since the expression is extremely rare in both Brown and Frown. The later corpus does display a higher frequency of the expression, but the raw frequencies are far too low to allow conclusions regarding the patterns of usage in this language variety, to compare it to other regional varieties or even to evaluate whether this difference is a general increase over time.

10.3 Analytical framework

As pointed out in the introduction (Chapter 1) this thesis had two aims. One was to examine the patterns of usage of the expressions of future. Despite occasional problems in doing this, a number of patterns were indeed found. These have been presented in my studies and in this Summary, as recapitulated above. The second aim of this thesis was to examine how the corpus-based approach could be used for an investigation such as the present one. The results of my linguistic studies have shown that the chosen approach was well suited for this task.

In Chapter 3, three central characteristics of corpus-based studies were presented (based on Biber 1996). These related to the study of language use, choice of texts/corpora, and methods employed for the analysis. My studies followed the tradition of previous corpus-based studies in the way these three characteristics were taken into consideration.

10.3.1 Language use

Biber (1996:172) suggests that corpus-based studies are empirical, "analyzing the actual patterns of use in natural texts". Throughout my studies, the focus has been on examining how the expressions of future are used, to find patterns of use. This has resulted in presentations of how the distribution of the expressions varies between different kinds of text, how the different expressions occur with certain collocates, and how the expressions are similar or dissimilar with regard to how they are used (see sections 10.2.1–10.2.5 for a summary of the main findings).

10.3.2 Choice of text

According to Sinclair, the results of a corpus-based study are only as good as the corpus upon which they are based (1991:13). The composition of a corpus is important and any result of a corpus-based study should be interpreted with this in mind, as I have done in my studies.

I have chosen to base my studies on generally available, carefully documented and well-known corpora of Present-day English, as described in Chapter 3. As far as possible I have tried to use corpora that are well suited for comparison, such as the Brown corpus and its clones. Even so, no conclusions regarding patterns of use, their similarities and differences can be drawn without critically examining the corpora used for the study. It cannot be assumed that corpora that seem similar in one respect are necessarily suitable for comparison in all cases. Not only the composition of the corpus but also the way the corpus was created can affect the outcome of a study. As an example the spoken British English corpora LLC and BNC (spoken part) can be mentioned. At first glance it may seem that the major difference between the corpora is the time when the data were collected. However, it cannot be assumed that the spoken data were treated in the same way when transcribed. Differences between the corpora regarding the use of *gonna* may therefore turn out to be due to differences in the transcription practices rather than the result of change in the language use over time.

As stated above, to allow for any results of a corpus-based study to be critically examined, it is important to know what corpora it is based on. It is against this background that I have chosen to describe the corpora in my studies and interpret the results of my studies with the composition of the corpora in mind. A further motivation for describing what corpora, tools and methods have been used for my studies is that by doing so I make the results open for scrutiny and comparison. The existence of carefully documented and generally available corpora makes it possible for us to produce linguistic studies that can be reproduced and thus verified.

10.3.3 Methods of data retrieval and analysis

As mentioned above, Biber (1996:172) suggests that one feature that corpus-based studies share is that they use computers for analysis, use both automatic and interactive methods, and that the analyses make use of quantitative and qualitative techniques. My studies are similar to other corpus-based studies in this respect.

The tools that I have used for my studies are presented in Chapter 3. It would have been, if not impossible at least very time-consuming and impractical to perform my studies without these tools. Not only do the tools facilitate and considerably speed up the process of identifying and counting instances of an expression, they also make this process less prone to errors and variation that is difficult to avoid when manually going through a large body of text and counting individual occurrences of a particular item. Automatic processes are more consistent, and it is usually a very simple process to recalculate and compare various frequencies, for example to see the relative frequency of an item in a particular body of text. Not all tools are ideally suited to all kinds of corpora or every kind of analysis, as discussed further in Study IV.

In my studies, my method of analysis has been primarily quantitative. I have examined the association patterns of the expressions of future starting from a quantitative angle. I have aimed at accounting for the absolute raw frequency of the expressions[60] even though my discussions generally focus on relative frequencies (to facilitate comparison between corpora/text categories of different size) or focus on the proportions of one expression in comparison to the other. A central motivation for focusing on the quantita-

[60] In addition to including some raw frequencies in tables and discussions I also provide complete frequency tables in Appendix B.

tive aspect is that quantitative results can be verified and reproduced; presenting the raw frequencies as well facilitates this further. Patterns based on low-frequency items are prone to considerable variation with only small deviations in frequency. By providing the raw frequencies, I allow for critical scrutiny of my claims regarding the patterns of use of these items. Moreover, with access to the raw frequencies, readers of my studies can make their own statistical calculations to evaluate the statistical significance of the patterns identified. Lastly, by providing this information I facilitate for others to either repeat my studies or use my data for alternative, comparative studies.

As shown in my studies, it is not always sufficient to look at large categories of text, such as whole corpora, to identify the patterns that govern the use of the expressions of future. A pattern that is identified when a whole corpus is examined can turn out to be not a feature of the language as a whole but the result of the distribution in one part of the corpus. By examining the data from different angles and also considering the distribution across various subsets of the corpora, it is possible to identify patterns of use that would not otherwise emerge. Thus, by considering the CG and DS components of the spoken part of the BNC separately, I could show that women do not use *gonna* more than men, even though it may seem so if the whole spoken part of the corpus is examined.

One problem with looking at smaller components of a corpus is that frequencies of certain items can be very low. Throughout my studies, I have found that the amount of available data is at times insufficient to be used as the basis for certain conclusions. This is particularly noticeable when relatively small corpora are divided into even smaller units, such as the text categories examined in my studies. Repeated comments about the lack of data illustrate one of the consistent problems of using corpora for linguistic study: there are not enough data. Low frequencies not only mean that there are few instances to look at but also that patterns found with these low frequencies are uncertain. Although finding a low frequency is a result in itself, and well worth noting, it is not always enough. If the aim of a study is to investigate certain linguistic and non-linguistic association patterns of an expression, this cannot be done if the expression does not feature in the corpus chosen for the study. With increasing access to new and bigger corpora, some of these problems may disappear with time. In my studies I have dealt with this problem by clearly stating where the frequencies are low, avoided basing far-reaching conclusions on the data in question and, wherever possible, tried to repeat the study on a larger set of data or support it with results from other analyses.

10.4 Concluding remarks

In this thesis I have examined the use of five expressions of future by looking at distribution and collocation patterns in a number of Present-day English corpora. I have been able to show that the use of these expressions varies with a number of factors and that the expressions display both similarities and differences with regard to their usage.

A result that stands out is that the use of the expressions of future varies considerably with the kind of texts where they are found. A major overall difference is found between the Informative and Imaginative hyper-categories in the Brown corpus and its clones and between the Context-governed and Demographically Sampled components of the spoken BNC and BNC Sampler corpora. There are also substantial differences between the spoken and written material with regard to the use of expressions of future. Not only are the expressions more frequent in the spoken data, the proportions of the different expressions are also different, with *going to/gonna* and *'ll* being used more in the spoken corpora, at the expense of *will.*

It has been suggested that the use of the expressions of future is changing with time. Apart from a decrease over time in the use of *shall,* no apparent changes can be found by examining the corpora in my studies. When regional varieties of English are compared, the Indian English corpus displays a pattern of use of the FUT that differs from that in the British and American English corpora.

My investigations of the collocational patterns of the expressions of future suggest that there are certain features that all expressions share, such as that they are used extensively with infinitives and personal pronouns. The expressions of future are hardly different from other auxiliary and semi-auxiliary verbs in this respect. It has been more interesting to find that there are certain differences between the expressions, for example that *will* is used with a more varied set of collocates than the other expressions while the collocations found with *going to* are almost identical to those with *gonna*.

It has been stated repeatedly that the choice of a corpus is important for the end result of a study. Throughout my studies, I have examined the variation in the use of the expressions of future in relation to the composition of the corpus, looking for patterns of systematic variation across whole corpora as well as across smaller components of text. I have strived to include in the discussions of my results considerations of the corpus and methods that have been used.

In his study of verb phrases with future reference, Close suggests that "…whereas the past is a chronicle of fact, the future is a tale untold, a mirage that each interprets in his own fashion" (1977:146). Through this thesis I hope to have made the English future a somewhat more tangible construction by presenting results that can be incorporated into further studies, thereby making future interpretations less of a mirage and more of a description of the actual patterns used.

References

Aitchison, Jean. 1991. *Language change: progress or decay?* Cambridge: Cambridge University Press. Second edition.

Aitchison, Jean. 2001. *Language change: progress or decay?* Cambridge: Cambridge University Press. Third edition.

Altenberg, Bengt. 1986. ICAME bibliography. *ICAME News* 10:67-79.

Altenberg, Bengt. 1993. *ICAME bibliography 2* http://nora.hd.uib.no/icame/icame-bib2.txt (visited March 5 2005).

Altenberg, Bengt. 1998. *ICAME bibliography 3 (1990-1998)* http://helmer.aksis.uib.no/icame/icame-bib3.htm (visited March 5 2005).

Aston, Guy, and Burnard, Lou. 1998. *The BNC handbook: exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Atkins, Sue, Clear, Jeremy, and Ostler, Nicholas. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7:1-16.

Atwell, Eric, Demetriou, George, Hughes, John, Schiffrin, Amanda, Souter, Clive, and Wilcock, Sean. 2000. A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal* 24:7-23.

Axelsson, Margareta Westergren. 1998. *Contraction in British newspapers in the late 20th century*. Uppsala: Acta Universitatis Upsaliensis.

Bernardini, Silvia. 2000. *Competence, capacity, corpora: a study in corpus-aided language learning*. Forli: Biblioteca della Scuola Superiore di Lingue Moderne per Interpreti e Traduttori.

Biber, Douglas. 1986a. Spoken and written textual dimensions in English: resolving the contradictory findings. *Language* 62:384-414.

Biber, Douglas. 1986b. On the investigation of spoken/written differences. *Studia Linguistica* 40:1-21.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

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

Biber, Douglas. 1996. Investigating language use through corpus-based analyses of association patterns. *International Journal of Corpus Linguistics* 1:171-197.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan, and Finegan, Edward. 1999. *The Longman grammar*. Harlow: Pearson Education Limited.

Binnick, Robert I. 1972. *Will* and *be going to* II. In *Papers from the eighth regional meeting Chicago linguistic society*, eds. Paul M. Peranteau, Judith N. Levi and Gloria C. Phares, 3-9. Chicago: Chicago Linguistic Society.

Bowker, Lynne, and Pearson, Jennifer. 2002. *Working with specialized language: a practical guide to using corpora*. London: Routledge.

Burnard, Lou. 1995. *Users Reference Guide for the British National Corpus Version 1.0*. Oxford: British National Corpus Consortium, Oxford University Computing Services.

Butler, Christopher. 1985. *Statistics in linguistics*. Oxford: Blackwell.

Bybee, Joan L., and Pagliuca, William. 1987. The evolution of future meaning. In *Papers from the 7th International Conference on Historical Linguistics*, eds. Anna Giacalone Ramat, Onofrio Carruba and Giuliano Bernini, 109-122. Amsterdam and Philadelphia: Benjamins.

Chafe, Wallace. 1995. The realis-irrealis distinction in Caddo, the Northern Iroquoian Languages, and English. In *Modality in grammar and discourse*, eds. Joan Bybee and Suzanne Fleischman, 349-365. Amsterdam: Benjamins.

Chambers, J. K. 2003. *Sociolinguistic theory: linguistic variation and its social significance*. Oxford: Blackwell. Second edition.

Close, R. A. 1962. *English as a foreign language: grammar and syntax for teachers and advanced students*. London: Allen & Unwin.

Close, R.A. 1977. Some observations on the meaning and function of verb phrases having future references. In *Studies in English usage: the resources of a present-day English corpus for linguistic analysis*, eds. Wolf-Dietrich Bald and Robert Ilson, 125-156. Frankfurt am Main: Peter Lang.

Coates, Jennifer. 1983. *The semantics of the modal auxiliaries*. London and Canberra: Croom Helm.

Collins COBUILD English grammar. 1990. London: Collins.

Crowdy, Steve. 1993. Spoken corpus design. *Literary and Linguistic Computing* 8:259-265.

Crowdy, Steve. 1994. Spoken corpus transcription. *Literary and Linguistic Computing* 9:25-28.

Crowdy, Steve. 1995. The BNC spoken corpus. In *Spoken English on computer: transcription, mark-up and application*, eds. Geoffrey Leech, Greg Myers and Jenny Thomas, 224-234. London: Longman.

Crystal, David. 2003. *The Cambridge encyclopedia of the English language*. Cambridge: Cambridge University Press.

Dahl, Östen. 2000. The grammar of future time reference in European languages. In *Tense and Aspect in the Languages of Europe*, ed. Östen Dahl, 309-328. Berlin: Mouton de Gruyter.

Danchev, A., Pavlova, A., Nalchadjan, M., and Zlatareva, O. 1965. The construction *going to* + inf. in Modern English. *Zeitschrift für Anglistik und Amerikanistik* 13:375-386.

Danchev, Andrei, and Kytö, Merja. 1994. The construction *be going to* + infinitive in Early Modern English. In *Studies in Early Modern English*, ed. Dieter Kastovsky, 59-77. Berlin and New York: Mouton de Gruyter.

Davidsen-Nielsen, Niels. 1990. *Tense and mood in English: a comparison with Danish*. Berlin & New York: Mouton de Gruyter.

Declerck, Renaat. 1991. *Tense in English: its structure and use in discourse*. London: Routledge.

Declerck, Renaat, and Depraetere, Ilse. 1995. The double system of tense forms referring to future time in English. *Journal of Semantics* 12:269-310.

EAGLES. *Corpus Encoding Standard*: EAGLES: Expert Advisory Group on Language Engineering Standards, http://www.cs.vassar.edu/CES/ (visited 27 February 2005).

Edwards, Jane A. 1995. Principles and alternative systems in the transcription, coding and mark-up of spoken discourse. In *Spoken English on computer: transcription, mark-up and application*, eds. Geoffrey Leech, Greg Myers and Jenny Thomas, 19-34. London: Longman.

Facchinetti, Roberta. 1998. Expressions of futurity in British Caribbean Creole. *ICAME Journal* 22:7-22.

Fasold, Ralph. 1990. *Sociolinguistics of language*. Cambridge USA/Oxford, UK: Blackwell.

Francis, W. Nelson, and Kucera, Henry. 1979. *Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with digital computers* http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM (visited 11 March 2005).

Garside, Roger. 1995. Grammatical tagging of the spoken part of the British National Corpus: a progress report. In *Spoken English on computer: transcription, mark-up and application*, eds. Geoffrey Leech, Greg Myers and Jenny Thomas, 161-167. Harlow: Longman.

Garside, Roger, Leech, Geoffrey, and McEnery, Anthony. 1997. *Corpus annotation: linguistic information from computer text corpora*. London: Addison Wesley Longman.

Greenbaum, Sidney, and Svartvik, Jan. *The London-Lund Corpus of Spoken English* http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM (visited March 11 2005).

Gvozdanovic, Jadranka. 1991. Meaning and interpretation of tense. In *The function of tense in text*, eds. Jadranka Gvozdanovic, Theo A. J. M. Janssen and Östen Dahl, 125-141. Amsterdam: North Holland.

Haegeman, Liliane. 1989. *Be going to* and *will*: a pragmatic account. *Journal of Linguistics* 25:291-317.

Harder, Peter. 1994. Verbal time reference in English: structure and functions. In *Tense, aspect and action: empirical and typological contributions to language typology*, eds. Carl Bache, Hans Basbøll and Carl-Erik Lindberg, 59-77. Berlin and New York: Mouton de Gruyter.

Hopper, Paul J., and Traugott, Elizabeth Closs. 2003. *Grammaticalization*. Cambridge: Cambridge University Press. 2nd.

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

Hundt, Marianne, Sand, Andrea, and Siemund, Rainer. 1998. *Manual of information to accompany The Freiburg - LOB Corpus of British English ('FLOB')*: Englisches Seminar: Albert-Ludwigs-Universität, http://khnt.hit.uib.no/icame/manuals/flob/INDEX.HTM (visited March 11 2005).

Hundt, Marianne, and Mair, Christian. 1999. "Agile" and "uptight" genres: the corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4:221-242.

Hundt, Marianne, Sand, Andrea, and Skandera, Paul. 1999. *Manual of information to accompany The Freiburg - Brown Corpus of American English ('Frown')* http://khnt.hit.uib.no/icame/manuals/frown/INDEX.HTM (visited March 11 2005).

Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Jespersen, Otto. 1933. Essentials of English grammar. London: George Allen & Unwin Ltd.

Johansson, Stig, Leech, Geoffrey N., and Goodluck, Helen. 1978. *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers* http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM (visited 11 March 2005).

Johansson, Stig. 1995. The approach of the Text Encoding Initiative to the encoding of spoken discourse. In *Spoken English on computer: transcription, mark-up and application*, eds. Geoffrey Leech, Greg Myers and Jenny Thomas, 82-98. London: Longman.

Joos, Martin. 1968. *The English verb: form and meanings*. Madison/London: The University of Wisconsin Press, Ltd. 2nd.

Kjellmer, Göran. 1998. On contraction in Modern English. *Studia Neophilologica* 69:155-186.

Krogvig, Inger, and Johansson, Stig. 1984. *Shall* and *will* in British and American English: a frequency study. *Studia Linguistica* 38:70-87.

Krug, Manfred. 1996. Language change in progress: Contractions in journalese in 1961 and 1991/92. In *Proceedings of the 1995 Graduate Research Conference on Language and Linguistics.*, ed. S McGill, 17-28. Exeter.

Krug, Manfred G. 2000. *Emerging English modals: a corpus-based study of grammaticalization*. Berlin and New York: Mouton de Gruyter.

Labov, William. 1966. The social stratification of English in New York City. Washington, DC: Centre for Applied Linguistics.

Leech, Geoffrey. 1971. *Meaning and the English verb*. London: Longman.

Leech, Geoffrey, Garside, Roger, and Bryant, Mary. 1994. CLAWS4: The tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94). Kyoto, Japan*, 622-628 http://www.comp.lancs.ac.uk/computing/research/ucrel/papers/coling1994paper.pdf.

Leech, Geoffrey. 1995. *A brief users' guide to the grammatical tagging of the British National Corpus* http://www.natcorp.ox.ac.uk/what/gramtag.html (visited January 13 2005).

Leech, Geoffrey, Myers, Greg, and Thomas, Jenny. 1995. *Spoken English on computer: transcription, mark-up and application*. London: Longman.

Leech, Geoffrey. 1997. Introducing corpus annotation. In *Corpus annotation: linguistic information from computer text corpora*, eds. Roger Garside, Geoffrey Leech and Anthony McEnery, 1-18. London: Longman.

Leech, Geoffrey. 2004. *Meaning and the English verb*. London: Pearson Education.

Longman dictionary of contemporary English. 2000. Harlow: Pearson Education Limited. 2nd.

Mair, Christian. 1997. The spread of the *going-to*-future in written English: a corpus-based investigation into language change in progress. In *Language history and linguistic modelling. A festschrift for Jacek Fisiak on his 60th birthday*, eds. Raymond Hickey and Stanislaw Puppel, 1537-1543. Berlin and New York: Mouton de Gruyter.

McEnery, Tony, and Wilson, Andrew. 2001. *Corpus linguistics*. Edinburgh: Edinburgh University Press. 2nd.

McMahon, April M. S. 1994. *Understanding language change*. Cambridge: Cambridge University Press.

Mendenhall, William, Beaver, Robert J., and Beaver, Barbara M. 2003. *Introduction to probability and statistics*. Pacific Grove: Brooks/Cole-Thomson Learning. 11.

Meyer, Charles F. 2002. *English corpus linguistics: an introduction*. Cambridge: Cambridge University Press.

Nakamura, Junsaku. 1993. Quantitative comparison of modals in the Brown and LOB corpora. *ICAME Journal* 17:29-48.

Nehls, Dietrich. 1988. Modality and the expression of future time in English. *International Review of Applied Linguistics in Language Teaching* 26:295-307.

Oakes, Michael P. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Oxford English Dictionary. 1989. Oxford: Clarendon Press. 2nd ed.

Palmer, F. R. 1965. *A linguistic study of the English verb*. London: Longman.

Palmer, F. R. 1974. *The English verb*. London: Longman.

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

Palmer, F. R. 1986. *Mood and modality*. Cambridge: Cambridge University Press.

Palmer, F. R. 1988. *The English verb*. London: Longman. Second edition.

Poplack, Shana, and Tagliamonte, Sali. 2000. The grammaticization of *going to* in (African American) English. *Language Variation and Change* 11:315-342.

Potter, Simeon. 1969. *Changing English*. London: Andre Deutsch.

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

Reichenbach, H. 1947. *Elements of symbolic logic*. New York: Macmillan.

Renouf, Antoinette. 1986. The elicitation of spoken English. In *English in speech and writing: a symposium*, eds. Gunnel Tottie and Ingegerd Bäcklund, 177-197. Stockholm: Almqvist & Wiksell.

Scott, Mike. *WordSmith Tools webpage* http://www.lexically.net/wordsmith/ (visited March 9 2005).

Shastri, S. V., Patilkulkarni, C. T., and Shastri, Geeta S. 1986. *Manual of information to accompany the Kolhapur corpus of Indian English, for use with digital computers* http://khnt.hit.uib.no/icame/manuals/kolhapur/INDEX.HTM (visited March 11 2005).

Shastri, S. V. 1988. The Kolhapur Corpus of Indian English and work done on its basis so far. *ICAME Journal* 12:15-26.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sperberg-McQueen, C. M., and Burnard, Lou. 2002. *Guidelines for text encoding and interchange.* Oxford: Humanities Computing Unit, University of Oxford.

Stubbs, Michael. 1980. *Language and literacy: the sociolinguistics of reading and writing*. London: Routledge.

Stubbs, Michael. 1996. *Text and corpus analysis: computer-assisted studies of language and culture*. Oxford: Blackwell.

Stubbs, Michael. 2001. *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.

Szmrecsanyi, Benedikt. 2003. *Be going to* versus *will/shall*. Does syntax matter? *Journal of English Linguistics* 31:295-323.

Vikner, Sten. 1985. Reichenbach revisited: one, two or three temporal relations? *Acta Linguistica Hafniensia* 192:81-98.

Visser, Fredericus Theodorus. 1973. *An historical syntax of the English language*. Leiden: E. J. Brill.

Wang, Jianxin. 2000. HIT and ICAME - a visiting researcher's observation. *ICAME Journal* 24:151-154 http://helmer.aksis.uib.no/icame/ij24/hit-icame.pdf.

Wekker, Hans Christian. 1976. *The expression of future time in contemporary British English*. Amsterdam: North-Holland Publishing Company.

Westin, Ingrid. 2002. *Language change in English newspaper editorials*. Amsterdam: Rodopi.

Williams, Christopher. 2002. *Non-progressive and progressive aspect in English*. Fasano: Schena editore.

Williams, Christopher. 2005. *Tradition and change in legal English: verbal constructions in prescriptive texts*. Bern: Peter Lang.

Woods, Anthony, Fletcher, Paul, and Hughes, Arthur. 1986. *Statistics in language studies*. Cambridge: Cambridge University Press.

Appendix A: Abstracts of Studies I–V

Study I

Future in present-day English: corpus-based evidence on the rivalry of expressions. ***ICAME Journal*** **21:7-20 (1997)**

Expressions: *will, 'll, shall, going to, gonna*
Corpora: LOB, Brown, Kolhapur, LLC
Factors: variation between regional varieties (British, American, Indian);
variation between written and spoken British English;
variation between hyper-categories (Informative vs Imaginative)
Tools and methods: WordCruncher was used for retrieving the concordances and manual inspection was used to identify relevant instances. Instance of *going to/gonna* used with past tense forms of the auxilliary *be* were omitted from the study. Instances not used with an explicit infinitive were excluded.

The study examines the use of five expressions of future in three written corpora (LOB, Brown, Kolhapur) and one spoken corpus (LLC). As far as the frequency of the expressions is concerned, the main difference between the written corpora is that the Indian corpus (Kolhapur) contains fewer instances of the expressions than the American (Brown) and British (LOB). However, this difference is not found across all text categories in the corpora: there are categories with high frequencies in the Kolhapur and low in the LOB, for example. The differences between the Informative and Imaginative hyper-categories are greater than the variation between the corpora – there are more expressions of future in the Imaginative hyper-category than in the Informative in all three corpora.

A similar pattern is found when the proportions of the expressions are examined: the variation is greater between the Informative and Imaginative hyper-categories than between the corpora of different regional varieties. *Will* and *shall* are used proportionately more in the Informative hyper-categories, while the proportions of *'ll* and *going to* are larger in the Imaginative hyper-categories.

The comparison between the two British English corpora (LOB and LLC) reveals that the expressions of future are more frequent in the spoken corpus.

The proportions of the different expressions also vary between the corpora, so that *will* is more frequent in the written corpus while *'ll* and *going to* are used more in the spoken than in the written corpus. The proportions of the expressions in the spoken corpus are more similar to those in the Imaginative hyper-category than to those in the Informative texts. One explanation for this is that the Imaginative texts contain larger proportions of speech-like text, such as imagined speech and dialogues.

Study II

Exploiting a large spoken corpus: an end-user's way to the BNC. *International Journal of Corpus Linguistics* 4(1): 29-52 (1999)

Expressions: *going to, gonna*
Corpora: BNC spoken part
Factors: variation between parts of the corpus;
variation with speaker properties
Other features: methodology of using large spoken corpora;
issues related to creating and using spoken corpora;
mark-up, transcription
Tools: BNCweb and SARA used for corpus searches

This study explores the spoken part of the British National Corpus (BNC), and discusses factors to consider when the corpus is used for linguistic study. The paper first presents an overview of the BNC (spoken part) and describes the compilation, transcription and mark-up of the corpus. The features of the corpus are then explored further in the context of a study of the distribution of the two expressions *gonna* and *going to.* The focus is on how the features of the corpus affect the results of such a study.

Among the issues dealt with are the importance of transcription. It is suggested that users of the corpus need to be aware of how transcription practice may vary and consider what may be affected by this. It is particularly difficult to judge the transcription in cases such as *gonna* and *going to*, where the pronunciation can be said to vary across a continuum but is transcribed as two distinct variants.

The mark-up of the corpus is also examined, in particular with regard to the coverage of the information. It is shown how not all texts are provided with the same kind of information about the speakers. A consequence of this is that a study that examines how a particular feature is used by different kinds of speakers cannot be based on the whole 10-million-word corpus. For studies examining language produced by speakers whose age, sex and social group are known, the amount of data that is available is only about 1.8 million words. Information about the speakers is primarily available for the texts in the Demographically Sampled component of conversational data.

The study also shows that there are considerable differences between the two spoken components in the corpus. The proportion of male speakers, for example, is larger in the more formal Context-governed (CG) component while the Demographically Sampled (DS) component contains more data produced by women. Generally the texts in the DS component have more information about the speakers producing the text. The study illustrates how important it is to consider such factors when evaluating the result of searches in the corpus.

Furthermore, the study points to some errors in the corpus and offers suggestions of how these can be identified and how the effect of them can be minimized when the corpus is used.

It is concluded that the spoken BNC is a valuable resource. The large amount of text (10 million words) and the rich metadata provide great potential for different kinds of studies. However, it is shown that it is vital to evaluate the information provided about the texts carefully to establish to what extent the examined data is documented and how reliable the information is and. Failure to do so may result in a study based on skewed data or fewer data than expected, which could lead the user to draw erroneous conclusions.

Study III

"You're gonna, you're not going to": a corpus-based study of colligation and collocation patterns of the (BE) going to construction in Present-day spoken British English. In Lewandowska-Tomaszczyk, B. and P. J. Melia (eds.) ***PALC'99: Practical Applications in Language Corpora. Papers from the International Conference at the University of Lódz, 15-18 April 1999.*** **Frankfurt am Main: Peter Lang. 161-192 (2000)**

Expressions: primarily *going to, gonna.* Also *will, 'll, shall*
Corpora: BNC spoken part
Factors: variation with linguistics factors (collocations and colligations)
Other features: comparing frequencies and frequency order as well as Observed/Expected values
Tools: BNCweb

The study looks at the expression *going to* and examines the collocational patterns of the two variants *going to* and *gonna.* It considers collocations with lexical items and collocations with items of the same part-of-speech (here colligations) in the positions immediately preceding and following the expression (positions -1, -2, -3 and +1, +2). The patterns found are also compared with those of the other expressions of future.

The first section gives a brief overview of previous research on the going to expression. Examples are given of how the expression has been described, often with a focus on semantic features, and how the expression has been compared and contrasted with other expressions, in particular will. The distribution of the expressions of future in the BNC is then presented, with particular reference to the variation between the written and spoken parts with regard to the frequency and proportion of *going to/gonna*.

The main part of the study deals with the collocation and colligation patterns of *going to* and *gonna*. The patterns that are examined are those with lexical items and those with items from a particular word class. Comparisons are made with raw frequencies and frequency order. Observed/Expected (O/E) values are also calculated to get a clearer picture of the strength of the collocations. The result of the study suggests that both *going to* and *gonna* are preceded by forms of be and followed by verbs in the infinitive, which is hardly surprising. What is more noteworthy is that the two variants display very similar proportions of different collocates, both when word classes and lexical items are examined. When these patterns are compared to those found with the other expressions of future, it is clear that the variation between *will, 'll* and *shall* is much greater than the variation between *going to* and *gonna*. The O/E value is a crude measurement, in particular as far as low frequency elements are concerned, but here it provides a further illu-stration of the similarity between *going to* and *gonna* as the values are very similar. Where differences are found, it is where the strong collocates are very infrequent items. When comparisons are made with the other expressions of future, these all differ more between each other than *going to* and *gonna*.

Study IV

Utilising Present-day English corpora: a case study concerning expressions of future. *ICAME Journal* 24:25–63 (2000)

Expressions: *will, 'll, shall, going to, gonna*
Corpora: LOB, FLOB, LLC, BNC Sampler spoken part (CG and DS components)

Factors: variation with time;
variation with text category;
variation with linguistic factors (collocations, clusters)
Other features: comparability of the corpora, how they can be used for a linguistic study of this kind, and evaluation of how different tools can be used to exploit the resources
Tools: WordSmith Tools, QWICK, SARA

The study exploits a selection of spoken and written corpora in order to examine how the use of the expressions of future vary with a number of factors. Experiences of using the corpora and tools are also discussed.

The different corpora are first presented and compared. It is concluded that the written LOB and FLOB corpora are very well suited for comparison, while the spoken LLC and the two components of the BNC Sampler corpus differ in a number of ways that make comparisons more difficult. The five expressions of future are then introduced, and the frequency distributions of the expressions across the different corpora are illustrated. It is shown that the frequency of the expressions is higher in the spoken corpora than in the written, and that there are differences between the spoken corpora as far as frequency of the expressions is concerned.

When the proportions of the expressions are compared, it is found that the written LOB and FLOB corpora are very similar, both when the whole corpora are compared and when the Informative and Imaginative hyper-categories are considered. The difference between the earlier LOB and later FLOB is smaller than the difference between two hyper-categories in the same corpus. The only consistent difference is that the proportion of *shall* is larger in the earlier LOB corpus.

The most noticeable difference between the spoken corpora (LLC, Sampler CG and Sampler DS) is that the proportion of *gonna* is much larger in the two Sampler corpora (CG and DS). Since differences in transcription practices are a potentially influential factor, it is more interesting to note that the combined proportion of *going to* + *gonna* is higher in the Sampler corpora than in the LLC. The proportion of *will* + *'ll* is similar in all corpora, and lower than in the written corpora. *Shall* is used less in the slightly later Sampler corpora than in the LLC.

A section of the article examines the linguistic context of the expressions. The study shows how the expressions of future are used with some very frequent main verbs. The use of personal pronoun subjects is explored in some detail. It turns out that *will* is used less with personal pronoun subjects than the other expressions in all corpora, while *'ll* has the highest proportion of personal pronoun subjects. The choice of pronoun varies between the expressions, and the pattern of variation is basically the same in the written and spoken corpora. The frequent use of personal pronouns with the expressions of future is reflected in the so-called clusters found. However, it is not only the most frequent items that co-occur with the expressions in these co-occurrence patterns, but the most frequent clusters also contain other, less frequent items.

The study is concluded with a discussion of the results, and an addendum which reviews experiences of using the tools and corpora for an investigation of this kind.

Study V

***Gonna* and *going to* in the spoken component of the British National Corpus. In Mair, C. and M. Hundt (eds.) *Corpus linguistics and linguistic theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20) Freiburg im Breisgau 1999.* Amsterdam: Rodopi. 35-49 (2000)**

Expressions: *going to, gonna*
Corpora: BNC, primarily spoken part
Factors: variation with text category (formal/informal) and speaker-related factors (sex, age, social class, dialect, education)
Other features: variable rule analysis
Tools: BNCweb

The study examines how the use of *going to* and *gonna* varies between different kinds of text and when used by different kinds of speakers in the spoken part of the BNC.

The frequency of the expression in the written and spoken part of the BNC is first examined briefly, and it is shown that the *gonna/going to* expression is more frequent in the spoken part of the corpus. The proportion of the *gonna* variant (calculated as the proportion of *gonna* of the combined frequency of *gonna* + *going to*) is considerably larger in the spoken part. A comparison between the Context-governed component of more formal encounters (CG) and the Demographically Sampled component comprising more informal discourse (DS) shows that the *gonna* + *going to* expression is more frequent in the DS component, and also that the so-called informal variant *gonna* is proportionately more frequent there.

When the use of the variant forms is compared between different kinds of speaker groups, the patterns found follow well-known patterns of variation attested in other socio-linguistic studies. Men seem to use the informal variant *gonna* more than women, younger speakers use *gonna* more than older speakers, and speakers from lower social classes show a greater preference for the contracted variant than speakers from higher social classes. This variation between different speaker groups follows similar patterns in both components in the spoken corpus so that a speaker group that uses a large proportion of *gonna* in the CG component is also found to use a large proportion in the DS component. When the variation between groups of speakers from different educational backgrounds is examined, no conclusive pattern can be found. Similarly, the comparison of speakers grouped according to accent/dialect does not reveal any clear patterns. Reasons for this may be that the amount of data is too small, that the coding of the data in the corpus is unreliable, that other factors are stronger and affect the patterns of usage more, or that several factors interact.

To establish to what extent the different speaker-related factors work in combination, a variable rule analysis was performed on a part of the data (all instances of *gonna* + *going to* in the DS component produced by speakers whose age, sex and social class were known). The analysis suggests that the main factors affecting the choice of variant are age (younger speakers preferring *gonna*, older speakers using more *going to*) and social class (the lowest social classes using more *gonna*, the highest social class preferring *going to*). The variable rule analysis did not confirm the previous finding that men use *gonna* more than women. This can possibly be explained by the fact that the analysis was performed on the DS component only, where the difference in the use of the exprssion sof future between the sexes is smaller.

Appendix B. Tables with raw frequencies

Table A.1. *Expressions of future in the LOB corpus (genres, hyper-categories and total)*[*]

	will	***'ll***	***shall***	***going to***	***gonna***	**sum**
A	313	10	14	11	0	348
B	239	2	9	10	0	260
C	73	4	4	3	0	84
D	101	5	25	5	0	136
E	315	3	15	12	0	345
F	198	16	8	14	0	236
G	205	7	26	10	0	248
H	157	0	95	4	0	256
J	299	0	60	5	0	364
K	96	66	**27**	16	0	205
L	52	78	11	20	0	161
M	20	10	1	5	0	36
N	81	156	15	26	2	280
P	131	137	41	32	0	341
R	46	11	3	2	0	62
sum	2326	505	354	175	2	3362
INFO (A-J)	1900	47	256	74	0	2277
IMAG (K-R)	426	458	98	101	2	1085

* From the ICAME CD (1999) using WordSmith Tools, version 3.

Irrelevant instances (*will* not used as a modal verb, *going* followed by the preposition *to*, past tense instances of *going to*) were manually identified and excluded.

Table A.2. *Expressions of future in the FLOB corpus (genres, hyper-categories and total)**

	will	***'ll***	***shall***	***going to***	***gonna***	**sum**
A	361	21	3	22	0	407
B	259	24	16	7	0	306
C	48	0	1	13	0	62
D	58	1	6	0	0	65
E	238	20	2	6	1	267
F	324	10	8	6	0	348
G	157	2	26	6	0	191
H	240	0	43	4	0	287
J	303	6	40	5	0	354
K	66	60	6	17	0	149
L	58	72	11	18	1	160
M	14	10	4	6	0	34
N	68	77	3	25	0	173
P	101	88	30	16	1	236
R	29	12	1	6	1	49
sum	2324	403	200	157	4	3088
INFO (A-J)	1988	84	145	69	1	2287
IMAG (K-R)	336	319	55	88	3	801

* From the ICAME CD (1999) using WordSmith Tools, version 3.

Irrelevant instances (*will* not used as a modal verb, *going* followed by the preposition *to*, past tense instances of *going to*) were manually identified and excluded.

Table A.3. *Expressions of future in the Brown corpus (genres, hyper-categories and total)*[*]

	will	***'ll***	***shall***	***going to***	***gonna***	**sum**
A	391	34	5	9	0	439
B	238	5	19	16	0	278
C	63	1	2	2	0	68
D	64	1	17	1	0	83
E	275	34	5	5	1	320
F	169	16	12	8	0	205
G	237	14	32	2	0	285
H	240	0	99	1	0	340
J	335	2	41	0	1	379
K	60	33	1	14	1	109
L	36	83	4	26	0	149
M	18	11	3	5	0	37
N	64	91	11	16	6	188
P	63	96	4	20	4	187
R	20	9	2	4	0	35
sum	2273	430	257	129	13	3102
Info (A-J)	2012	107	232	44	2	2397
Imag (K-R)	261	323	25	85	11	705

[*] From the ICAME CD (1999) using WordSmith Tools, version 3.

Irrelevant instances (*will* not used as a modal verb, *going* followed by the preposition *to*, past tense instances of *going to*) were manually identified and excluded.

Table A.4. *Expressions of future in the Kolhapur corpus (genres, hyper-categories and total)*[*]

	will	*'ll*	*shall*	*going to*	*gonna*	**sum**
A	225	0	12	4	0	241
B	227	0	4	9	0	240
C	37	0	0	3	0	40
D	43	0	10	1	0	54
E	166	2	9	3	0	180
F	155	0	6	4	0	165
G	157	0	28	5	0	190
H	315	0	162	12	0	489
J	203	1	42	2	0	248
K	200	94	35	19	0	348
L	97	63	17	7	0	184
M	19	2	18	1	0	40
N	40	22	11	1	0	74
P	66	31	7	10	0	114
R	35	17	3	5	0	60
sum	1985	232	364	86	0	2667
Info (A-J)	1528	3	273	43	0	1847
Imag (K-R)	457	229	91	43	0	820

* From the ICAME CD (1999) using WordSmith Tools, version 3.
Irrelevant instances (*will* not used as a modal verb, *going* followed by the preposition *to*, past tense instances of *going to*) were manually identified and excluded.

Table A.5. *Expressions of future in the LLC*[*]

	will	*'ll*	*shall*	*going to*	*gonna*	**sum**
sum	933	1111	218	579	14	2855

* From the ICAME CD (1999) using WordSmith Tools, version 3.
Irrelevant instances (*will* not used as a modal verb, *going* followed by the preposition *to*, past tense instances of *going to*) were manually identified and excluded.

Table A.6. *Expressions of future (raw frequencies) in the BNC: full corpus, written and spoken parts, Imaginative (written) domain, CG and DS (spoken) components*[*]

	will	*'ll*	*shall*	*going to*	*gonna*	**sum**
BNC (written)	235602	17476	40737	20069	548	314432
BNC (written imaginative)	31193	29131	5333	11347	323	77327
BNC (spoken)	25049	31714	2873	9441	11914	75170
BNC (spoken CG)	13034	10904	995	5194	3611	33712
BNC (spoken DS)	8984	19146	1637	2730	8036	40419

* From the first release of the British National Corpus, using BNCweb.

Table A.7 *Expressions of future in the BNC Sampler. Spoken part and CG and DS components*[*]

	will	***'ll***	***shall***	***going to***	***gonna***	**sum**
Sampler (spoken)	2402	3328	308	805	1171	8014
Sampler (CG)	1345	1049	122	533	327	3376
Sampler (DS)	1057	2279	186	272	844	4638

* From the BNCSampler CD, using WordSmith Tools, version 3.

Table A.8. *Expressions of future in monologue and dialogue texts in the spoken CG and DS components of the BNC*[*]

	CG monologue	**CG dialogue**	**DS monologue**	**DS dialogue**
will	3649	9385	115	8741
'll	2587	8317	220	18577
shall	334	661	12	1584
going to	1140	4054	51	2663
gonna	890	2721	122	7777
sum	8600	25138	520	39342

* From the first release of the British National Corpus, using BNCweb.

Table A.9. *Expressions of future used by male and female speakers in the CG and DS components of the BNC*[*]

CG	***will***	***'ll***	***shall***	***going to***	***gonna***	**sum**
male	6580	6599	441	2620	2140	18380
female	1508	1592	155	966	343	4564
sum	8088	8191	596	3586	2483	22944
DS	***will***	***'ll***	***shall***	***going to***	***gonna***	**sum**
male	3003	6547	525	678	2844	13597
female	4891	10402	953	1752	4072	22070
sum	7894	16949	1478	2430	6916	35667

* From the first release of the British National Corpus, using BNCweb.

Table A.10. *Expressions of future used by speakers of different age groups in the CG and DS components of the BNC*[*]

CG	***will***	***'ll***	***shall***	***going to***	***gonna***	**sum**
-14	36	98	5	10	33	182
15-24	198	321	21	102	95	737
25-34	830	927	80	446	288	2571
35-44	766	801	43	364	197	2171
45-59	2229	2702	116	1269	717	7033
60+	382	638	54	200	80	1354
sum	4441	5487	319	2391	1410	14048
DS	***will***	***'ll***	***shall***	***going to***	***gonna***	**sum**
-14	864	1648	204	253	1034	4003
15-24	829	1917	144	172	1159	4221
25-34	1556	3669	281	394	1541	7441
35-44	1620	3489	236	461	1400	7206
45-59	1551	3304	240	470	994	6559
60+	1320	2636	245	597	674	5472
sum	7740	16663	1350	2347	6802	34902

[*] From the first release of the British National Corpus, using BNCweb.

Table A.11. *Expressions of future used by speakers of different social classes in the DS component of the BNC**

DS	***will***	***ll***	***shall***	***going to***	***gonna***	**sum**
AB	1228	2684	286	550	1076	5824
C1	909	2136	152	244	810	4251
C2	1176	2343	189	200	1147	5055
DE	479	1219	66	79	436	2279
sum	3792	8382	693	1073	3469	17409

* From the first release of the British National Corpus, using BNCweb.

Table A.12. *Personal pronoun subjects used with FUT in LOB*[*]

	will	***'ll***	***shall***	***going to***	***gonna***	**sum**
I	94	197	88	45	0	424
you	189	122	6	36	2	355
he	108	32	2	6	0	148
she	33	18	0	3	0	54
it	218	17	8	13	0	256
we	37	81	110	10	0	238
they	99	20	3	6	0	128
sum	778	487	217	119	2	1603

[*] From the ICAME CD (1999) using WordSmith Tools, version 3.

Table A.13. *Personal pronoun subjects used with FUT in FLOB**

	will	***'ll***	***shall***	***going to***	***gonna***	sum
I	87	172	72	35	0	366
you	131	74	0	18	1	224
he	86	35	0	9	0	130
she	25	9	0	3	0	37
it	185	15	0	10	0	210
we	76	49	76	15	0	216
they	107	25	0	6	1	139
sum	697	379	148	96	2	1322

* From the ICAME CD (1999) using WordSmith Tools, version 3.

Table A.14. *Personal pronoun subjects used with FUT in LLC**

	will	***'ll***	***shall***	***going to***	***gonna***	**sum**
I	111	495	120	102	2	830
you	111	149	4	91	2	357
he	35	52	0	30	2	119
she	4	20	0	7	0	31
it	54	91	2	51	1	199
we	35	166	79	72	1	353
they	63	60	1	36	1	161
sum	413	1033	206	389	9	2050

* From the ICAME CD (1999) using WordSmith Tools, version 3.

Table A.15. *Personal pronoun subjects used with FUT in Sampler CG**

	will	***'ll***	***shall***	***going to***	***gonna***	**sum**
I	106	362	61	67	54	650
you	121	132	2	76	55	386
he	21	9	0	5	7	42
she	8	11	0	13	0	32
it	151	78	0	56	28	313
we	138	332	42	109	52	673
they	79	58	0	30	29	196
sum	624	982	105	356	225	2292

* From the BNCSampler CD, using WordSmith Tools, version 3.

Table A.16. *Personal pronoun subjects used with FUT in Sampler DS*[*]

	will	***'ll***	***shall***	***going to***	***gonna***	**sum**
I	152	1049	117	74	226	1618
you	152	209	1	46	159	567
he	55	117		16	45	233
she	47	107		8	39	201
it	140	168		20	64	392
we	59	341	58	33	88	579
they	106	120		22	56	304
sum	711	2111	176	219	677	3894

* From the BNCSampler CD, using WordSmith Tools, version 3.

Table A.17. *Infinitival collocates in the BNC, spoken part*[*]

	gonna	***going to***	**sum**
VVI (lexical verb)	7293	5703	12996
VBI (*be*)	1944	2123	4067
VHI (*have*)	1029	622	1651
VDI (*do*)	882	738	1620
sum	11914	9441	21355

* From the first release of the British National Corpus, using BNCweb.

Table A.18. *Infinitival collocates in Sampler CG*[*]

verb	***will***	***'ll***	***shall***	***going to***	***gonna***	**sum**
be	391	171	25	144	86	817
have	73	101	6	29	37	246
get	33	41		30	17	121
do	32	45	8	32	14	131

* From the BNCSampler CD, using WordSmith Tools, version 3.

Table A.19. *The most frequent infinitival verbs used with expressions of future in the spoken part of the BNC. Ranking (1-13) and absolute frequency.*[*]

ranking	***will***	**25049**	***'ll***	**31714**	***shall***	**2873**	***going to***	**9441**	***gonna***	**11914**
1	be	6534	be	5327	be	395	be	2122	be	1943
2	have	1382	have	3909	have	269	do	737	have	1029
3	do	778	get	1848	go	173	have	621	do	882
4	get	660	go	1507	do	151	get	516	get	869
5	go	657	do	1444	say	133	go	339	say	586
6	come	523	see	1058	put	99	say	290	go	571
7	take	441	give	844	get	96	take	209	take	286
8	give	337	take	725	give	60	put	179	put	261
9	see	302	come	703	sell	60	give	158	come	220
10	say	272	tell	686	tell	58	come	142	give	173
11	know	270	put	615	take	56	make	137	make	152
12	make	226	say	500	see	47	happen	135	ask	123
13	need	220	find	500	start	34	ask	122	happen	114

[*] From the first release of the British National Corpus, using BNCweb.