Utilising Present-day English corpora: A case study concerning expressions of future

Ylva Berglund
Uppsala University

1 Introduction

During 1999, two new resources became available to corpus linguists. In the spring, the BNC Sampler CD was released, and the autumn saw the distribution of the new ICAME CD. These resources have added a new dimension to corpus-based studies, as they make it possible to use a large number of valuable corpora and search programs on an ordinary, stand-alone PC with even rather modest performance at a very reasonable cost. Some of the corpora have been available before, distributed, for example, by ICAME, and the programs are also familiar to many. It is, however, the combination of the corpora and the programs on a single CD-ROM that is new, and which will undoubtedly make these resources more easily available, and more readily accessed.

This paper presents a study based on four corpora of British English from the new CDs. The primary aim of the study is to describe how a set of expressions of future is used today, and to see to what extent this usage can be seen to vary with the features of time, genre/text category, and medium (written/spoken). The distribution of the expressions across the corpora is studied, as well as collocations and frequent clusters where the expressions occur. A secondary aim of the study is to see how the resources on the new corpus CDs can be exploited for a study of this kind. Some experiences of using the tools and corpora are discussed in the addendum section at the end of the paper.

1.1 Presentation and comparison of the corpora

The primary data for this study have been drawn from two corpora of written British English: the Lancaster-Oslo/Bergen Corpus of British English (LOB) and the Freiburg-LOB Corpus of British English (FLOB), and two corpora of spoken British English: the London-Lund Corpus of Spoken English (LLC) and the spoken component of the British National Corpus Sampler (Sampler). Several search and retrieval programs have been used in the process of this study, including WordSmith Tools version 3 (all corpora), SARA version 0.931 (Sampler), and Qwick version 1.0 (FLOB and Sampler). Unless otherwise stated, the figures presented have been obtained by using the WordSmith Tools program on the corpora as they are found on the CDs (for further comments, see the addendum section).

1.1.1 Written corpora

The LOB corpus contains written material from 1961, divided into 15 different text categories, nine of which are informative (for example press texts (three categories), popular lore, learned and scientific writings) and six are imaginative (such as romance and love story, science fiction, adventure and western fiction). The FLOB corpus is modelled on the LOB categories and texts, but contains material from 1991. Both corpora contain about one million words. The information about the corpora is summarised in Table 1:

Table 1: Written corpora. Lancaster-Oslo/Bergen Corpus (LOB), Freiburg LOB Corpus (FLOB)

	LOB	**FLOB**
Medium	Written	Written
Regional variety	British English	British English
Number of words	1,000, 000	1,000,000
Genres	15 categories	15 categories
Size of texts	2,000 words	2,000 words
Texts published	1961	1991

1.1.2 Spoken corpora

The London-Lund Corpus and the spoken component of the BNC sampler differ in several ways. The Sampler contains one million words, and the LLC approximately 500,000.[1] The Sampler consists of roughly equal proportions (just under 500,000 words each) of context-governed material and demographically sampled data, collected from across the UK. The LLC is composed of a number of different text types, such as conversations between equals or disparates in a face-to-face situation or on the telephone, radio discussions, spontaneous or prepared oration, and commentary. The LLC texts are all about 5,000 words long, while the Sampler contains both longer and shorter samples.

The time of recording of the materials differs somewhat. The demographically sampled material was collected in 1991 and 1992, while the texts in the context-governed component were recorded over a longer period, from 1982 to 1994.[2] The LLC texts were collected between 1953 and 1988 (the bulk of the data is from the mid-1960s to the mid-1970s). Although there is some overlap, the LLC texts are generally somewhat earlier than the texts in the Sampler.

The speakers in the LLC are adults, mostly 30–50 years old, with a large proportion of what is labelled 'academics' or people found in an academic context (university secretaries, prospective students, researchers, computer experts, etc). Information about age, sex, and occupation is available for the speakers in the LLC. The demographically sampled texts in the Sampler consist of spontaneous conversations between demographically selected respondents and people they talk to, while the context-governed data consist of recording from 'more formal encounters', where, for example, the proportion of male speakers is larger than that of female speakers (see Burnard 1995). The amount of information given about the Sampler speakers varies. A large number of the speakers are unknown, that is, no information about their age, sex, social class, etc is available. For a small number of speakers (most of the respondents in the demographically sampled component) information about a number of extra-linguistic parameters is supplied.

Information about the speakers is more generally available for the LLC than for the Sampler. The information reveals that the range of age and social status of the speakers is narrower in the LLC than in the Sampler (see Table 2):

Table 2: Spoken corpora. London-Lund Corpus (LLC), BNC Sampler context-governed component (CG), and BNC Sampler demographically sampled component (DS)

	LLC	**CG**	**DS**
Medium	Spoken	Spoken	Spoken
Regional variety	British English	British English	British English
Number of words	500,000+	496,852	493,852
Settings	mixed	'more formal contexts'	'more informal'

Genres	mixed (twelve categories)	four context-governed categories	spontaneous conversation
Year of recording	1953–88	1982–94	1991–92
Size of texts	5,000 words	appr. 4,000–16,000 words	appr. 4,000–12,000 words
Speakers	academics, adults	mixed, primarily adult men	demographically sampled
Speaker information	age, sex, occupation available for all speakers	varies from none to very detailed	varies; for some speakers information about age, sex, social class, accent, relationship between speakers etc is available

For the remainder of this study, the two Sampler components will be regarded as two different corpora, referred to as DS (the demographically sampled component) and CG (context-governed component). This means that the study comprises three spoken corpora of approximately the same size. The DS corpus is then a corpus of spontaneous conversation from 1991–1992, while the CG corpus contains data that have been classified as emerging from ‘more formal encounters’ (Aston and Burnard 1998:31), collected at roughly the same time as the DS corpus. The LLC contains a more varied set of data: spontaneous conversations, prepared oration and other kinds of texts, recorded slightly earlier than the Sampler material.

1.2 Expressions of future included in the study

Much has been written about expressions of future in English, often with emphasis on semantic aspects (for example Wekker 1976; Leech 1987). Many authors have dealt with comparisons between different ways to express future reference, in particular *will/shall* vs *BE going to* (for example Aijmer 1984, Collins 1987, Haegeman 1989), and the historical background of the expressions has also been described, for example by Bybee (1987) and Poplack and Tagliamonte (forthcoming).

The question whether English has a future tense or not has been discussed by a number of authors. Despite differences of opinion concerning this question, it is generally agreed that there are a number of ways to express future time in English; several sources list five main means: *will/shall*+infinitive, *BE going to*+infinitive, present progressive, simple present, and *will/shall*+progressive infinitive.[3] In the present study, the focus will be on the constructions consisting of an auxiliary verb used with or without an overt infinitive: *will, 'll, shall, BE going to,* and *(BE) gonna*. These will be referred to as the expressions of future, and they are presented further below (sections 1.2.1–1.2.5).

The reason for not including the simple present and present progressive in this study is that, arguably, the future reference in those constructions primarily lies in what Biber et al (1999:455) refer to as 'grammatical contexts'. In their corpus-based grammar, the authors state that the '[s]imple present tense is also used *in special cases* to refer to either past events or future events' (my emphasis), and that '[n]early all occurrences of present tense referring to future time occur ... either with an accompanying time adverbial that explicitly refers to the future, or in a conditional or temporal adverbial clause that has future time reference' (1999:455). It is, thus, only by reference to the context where the construction is found that it can be decided whether it is referring to the future or not. A similar argument can be put forward as to the present progressive. Leech (1987:63) exemplifies this, and states that '[t]he following sentences without adverbial modification are in fact ambiguous out of context, as they may be given either a present (limited duration) or future (imminent) interpretation: I'*m taking* Mary out for a meal. We*re starting* a bridge club. ...'.

To be able to include the instances of simple present or present progressive with future reference in the present study, it would be necessary to manually disambiguate all the occurrences to find the relevant instances. This would be a task too time-consuming under the circumstances, even if the constructions could be retrieved easily.[4]

1.2.1 Will/won't

The homograph *will* can be a noun, a main verb and an auxiliary verb, but it is only the auxiliary verb that is of interest to this study. Since not all corpora in this study are tagged with Part-of-speech (POS) information, the figures presented have all been obtained by searching for the word *will* and then deleting the instances where *will* is not an auxiliary verb. This process has been performed semi-manually, by combining the ‘sort’ function in WordSmith with a certain amount of manual scanning. Although this method does not identify all unwanted instances of *will*, it has been considered accurate enough for the present study, especially since valuable time could thus be saved. The result of the semi-automatic identification was compared to the result of searching tagged versions of the corpora, when these were available, and the difference was found to be negligible (2,326 instances in the untagged LOB semi-manually identified, and 2,316 instances tagged as modal verbs in the tagged version). [5]

The form *won’t* is infrequent in the written corpora (108 in LOB, 97 in FLOB). In the smaller spoken corpora, however, the *won’t* construction is found to a greater extent, 120 times in the LLC, 144 times in the CG and 453 times in the DS corpus. In this study the frequency for *won’t* has been included in the figures for *will*.

1.2.2 ’ll

It is generally accepted that *’ll* is the contracted, or reduced, variant form of *will.* It is, however, sometimes argued that the expression is also found as a variant of *shall*, for example by Leech (1987:57): ‘The full auxiliary forms *will* and *shall* are frequently contracted in speech (…) to the form written *’ll’*. In this paper I will not take a stand in this issue, but choose instead to treat *’ll* as a variant expression of future, along with *will, shall, BE going to,* and *(BE) gonna.* It has been shown that the use of this variant varies across time (Axelsson 1998), which would further motivate this approach.

1.2.3 Shall/shan’t

It has been claimed that *shall* is only used for future reference with first person subjects, and that when used with other subjects, the construction expresses obligation rather than prediction.[6] In the new corpus-based English grammar (Biber et al 1999), *shall* is, however, listed among the modals with volition/prediction meaning and not obligation/necessity (section 6.6). In agreement with Svartvik and Sager (1977:42B), it is, in this paper, considered that the question of how the instances of *shall* are to be interpreted ‘often lacks practical relevance’ (my translation). All instances of *shall* have been included in this study. Included in the figures presented are also the instances of the form *shan’t* (five instances in LOB, three in the FLOB, eight in the LLC, one in the CG and 15 in the DS corpus).

1.2.4 Going to

The expression *BE going to* differs from the *will/'ll/shall* expressions in that it can appear with an auxiliary marked for the present or the past tense (*We are/were going to…*). Used with *BE* in the past tense, the expression 'marks reference to a projected future time dating from some point in the past…' (Biber et al 1999, section 6.2.1.3). This is a feature which is not shared by the other expressions in this study. To enable comparison with the other expressions, the instances of *going to* used with past tense forms of *BE* were excluded from the study. For obvious reasons, instances of *BE going to* where *to* is a preposition (such *as 'He is going to London*') were also excluded from the study. The identification of the relevant cases was made semi-automatically. The instances of *BE going to* will, henceforth, be referred to by the shorter form *going to.*

1.2.5 Gonna[7]

It has been shown that *gonna* and *going to* are so similar in their collocational behaviour that they can be considered variants of one expression (Berglund forthcoming b). The data for the *going to* and *gonna* variants will, however, be presented separately in this study, in analogy with the treatment of *will/'ll/shall.* The interest in the variation between the *gonna* and *going to* forms that has been expressed in a number of recent studies is a further motivation for this approach (see, for example, Krug 1998/99; Berglund forthcoming a; Poplack and Tagliamonte forthcoming).

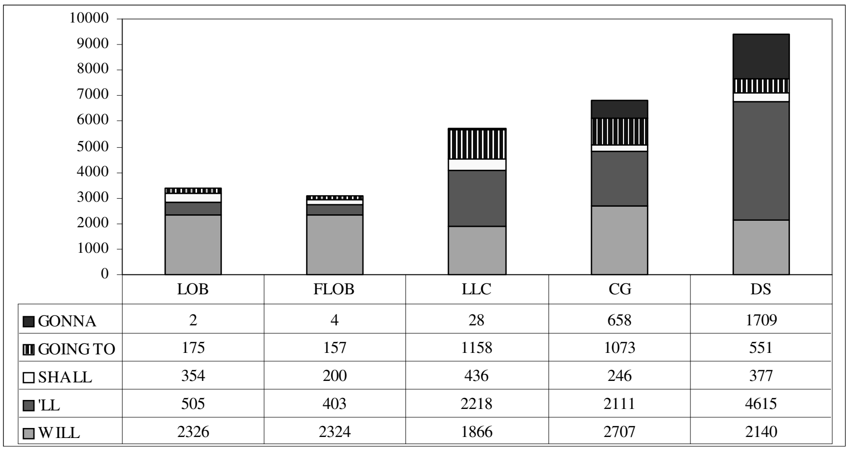
The *gonna* expression is very infrequent in the written material in this study, but is found to a considerable extent in the spoken data. Instances of *gonna* used with a past tense form of *BE* have been excluded from the study.

2 Frequency survey

2.1 Frequency overview

This section is introduced by a brief survey of the frequency of the different expressions of future in the corpora. A substantial difference between corpora or text categories in the frequency of the expressions of future may point to differences in the corpus set-up that would be relevant to take into consideration when evaluating the results of the study.

Figure 1 illustrates the frequency (per million words) of the expressions of future across the corpora in this study:



	LOB	FLOB	LLC	CG	DS
■ GONNA	2	4	28	658	1709
▥ GOING TO	175	157	1158	1073	551
□ SHALL	354	200	436	246	377
■ 'LL	505	403	2218	2111	4615
▨ WILL	2326	2324	1866	2707	2140

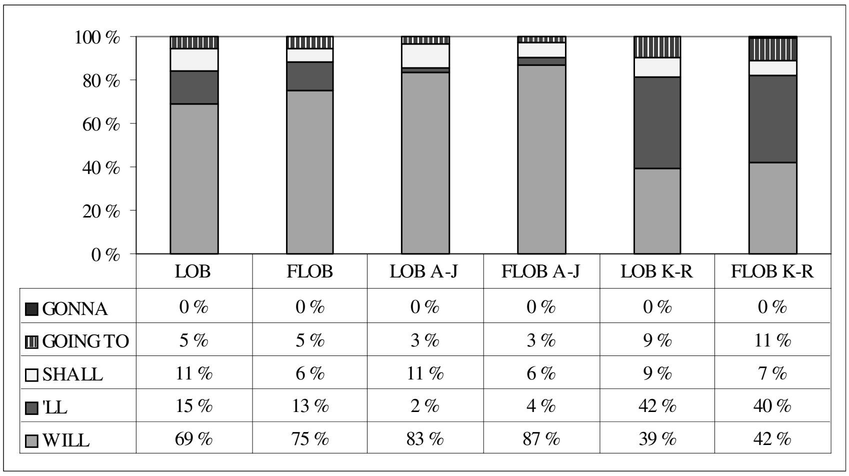
Figure 1: Frequency (per million words) of expressions of future in the five corpora

The two matched corpora, LOB and FLOB, are rather similar, in particular regarding the total frequency of the expressions of future. However, the proportions of the different expressions vary somewhat, which will be studied further below (sections 2.2 and following).

In the spoken corpora (LLC, DS, and CG), there is a considerably higher frequency of the expressions of future than in the written. The variation between the spoken corpora is larger than for the written, both with regard to the frequency and the proportions of the expressions.

2.2 Written corpora: LOB and FLOB

As seen above, there are no great differences between the corpora as far as the frequency of the expressions of future is concerned. Figure 2 illustrates the proportions of expressions of future in the corpora as a whole and in two sub-sets, or hyper-categories, in the corpora.[8] The proportions are given as percentages of the combined frequency of all expressions in each corpus or hyper-category.



	LOB	FLOB	LOB A-J	FLOB A-J	LOB K-R	FLOB K-R
■ GONNA	0 %	0 %	0 %	0 %	0 %	0 %
▥ GOING TO	5 %	5 %	3 %	3 %	9 %	11 %
□ SHALL	11 %	6 %	11 %	6 %	9 %	7 %
■ 'LL	15 %	13 %	2 %	4 %	42 %	40 %
▨ WILL	69 %	75 %	83 %	87 %	39 %	42 %

Figure 2: Proportions of expressions of future in different parts of the LOB and FLOB corpora (for frequencies, see Figure 1)

Totally speaking, the proportion of *will* is higher in the FLOB corpus (75% vs 69% in LOB), while the proportion of *'ll* is lower (13% vs 15% in LOB). The proportion of *shall* is considerably lower in FLOB (6% vs 11%), while the proportion of *going to* is the same in both corpora (5%). To judge from these figures alone, it seems as if the major development across the 30-year span from 1961 to 1991 is that the use of *shall* has decreased while the use of *will* has increased. This is not a surprising development, considering that *shall* has been said to be used less frequently today. The slight decrease in the proportion of the contracted form *'ll* is more surprising, on the basis of other investigations which report an increased use of this form during this period (for example Axelsson 1998). As those studies show, however, it is not enough to look at frequency alone as far as this development is concerned, as other factors are found to be significant.

A previous study (Berglund 1997) showed that the variation within a corpus in the distribution of expressions of future is greater than the variation between comparable corpora of different regional varieties of English. The present study shows that the differences between the earlier LOB corpus and the later FLOB also seem to be smaller than the variation between the Imaginative and Informative hyper-categories within the corpora.

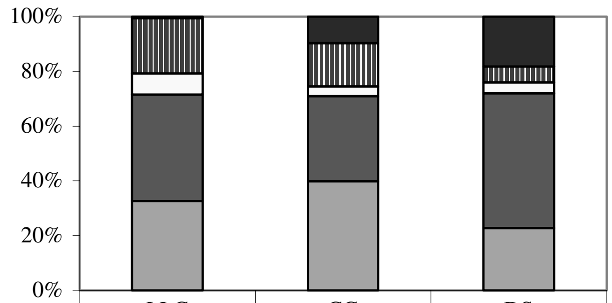
In the Informative hyper-category (text categories A–J), *will* is the expression used most in both corpora, constituting 83 per cent of all expressions of future in the LOB and 87 per cent in the FLOB corpus. The second most frequent expression, *shall,* is used more in the LOB corpus (11%) than in FLOB (6%), while the infrequent expression *'ll* is found slightly more in the later FLOB corpus. *Going to* is used to the same, very low, extent in both corpora (3%), while *gonna* is found only once (in FLOB).

In the Imaginative hyper-category (text categories K–R), *will* and *'ll* are found to a similar extent in both corpora, and the two expressions together make up 81 per cent and 82 per cent of all the expressions of future in the LOB and FLOB corpora, respectively. Contrary to the Informative hyper-category, *'ll* appears more in the earlier LOB corpus. The expression *shall* occurs less in the FLOB corpus than in LOB in the Imaginative hyper-category, which is the same pattern as found for the Informative hyper-category. *Going to* is found considerably more in the Imaginative hyper-category, and is also found to be more frequent in the FLOB corpus than in LOB, eleven per cent and nine per cent of all expressions of future, respectively. There are only two instances of *gonna* in LOB and four in FLOB in the Imaginative hyper-category.

To sum up, the Informative hyper-categories in both corpora are characterised by a high proportion of *will*, and a very low proportion of *going to* and *'ll,* while the distribution of *will* and *'ll* is more even in the Imaginative category, where there is also a higher proportion of *going to*. *Shall* is found more often in the earlier LOB corpus in both hyper-categories.

2.3 Spoken corpora: LLC and Sampler

The frequency and proportions of the expressions of future vary considerably between the three spoken corpora in this study, as illustrated in Figure 3:



	LLC	CG	DS
■ GONNA	0%	10%	18%
▥ GOING TO	20%	16%	6%
□ SHALL	8%	4%	4%
■ 'LL	39%	31%	49%
▨ WILL	33%	40%	23%

Figure 3: Expressions of future in spoken corpora. Proportions (for frequencies, see Figure 1)

A striking difference between the spoken corpora is that the proportion of *gonna* is considerably higher in the two Sampler components, ten per cent in the CG and 18 per cent in the DS, compared with the virtual non-existence of the expression in the LLC. Although the proportion of *going to* is higher in the LLC than in the other two corpora, the combined proportion of *gonna+going to* is lower, just over 20 per cent, to be compared to 26 per cent and 24 per cent in the CG and DS corpora.

Will was the most frequent expression in the written data. In the spoken corpora, the proportions of *will* vary greatly between the corpora; 23 per cent in the DS, 33 per cent in the LLC, and 40 per cent in the CG corpus. A similar variation is also found for the expression *'ll,* which is most frequent in the corpus with the lowest proportion of *will* (DS), and least frequent where the highest proportion of *will* was found, in the CG component. The combined proportions of *will* and *'ll* are the same in all spoken corpora (71% – 72%); overall lower than in the written data.

A common feature of the Sampler components is that *shall* occurs less there than in the slightly earlier LLC. It also seems to be the case that *going to* and *gonna* appear more in the CG and DS corpora than in the LLC. The combined proportion of *will* and *'ll* is the same in the three spoken corpora, but the proportions of the full and contracted forms vary. The difference is largest between the more formal CG and the conversational DS, with the LLC corpus somewhere in between.

Gonna is found almost exclusively in the Sampler data. One explanation could be that the use of the expression has increased recently, since the collection of the LLC. That would, however, not explain the difference between the two Sampler components that are from, approximately, the same time. The level of formality may be an important factor here. A further explanation may lie in differences in transcription practices, something that has been discussed previously (Krug 1998/1999, Berglund 1999).

3 Collocations and clusters

In the previous section, the focus was on the distribution of expressions of future across the text categories and corpora. It was found that the expressions of future are used more frequently in spoken data, that the proportions of the expressions differ when the spoken and written corpora are compared (see Figure 1) and also that there are differences between the genres in both written and spoken corpora in this regard (different proportions of the expressions in the Imaginative/Informative hyper-categories in the written corpora and in the spoken DS and CG components). In what follows, the emphasis will be on the linguistic context of the expressions, studied in the form of collocations and clusters where the expressions occur.

3.1 Collocations: 'One can tell a word from the company it keeps' (WordSmith Tools)

3.1.1 Introduction

Collocations have been defined as 'associations between two words, so that the words co-occur more frequently than expected by chance' (Biber and Conrad 1999:183). Such associations can be studied to discern differences and similarities between texts or expressions. Biber and Conrad add that '[c]orpus-based analysis shows that words with similar meaning are often distinguished by their preferred collocations' (1999:183). In this study collocations are studied as a means to identify similarities or differences between the expressions of future.

The value of studies of collocations has been recognised by a number of authors, and most corpus handling tools now contain a function to identify collocations. WordSmith can find the frequency for all words in the positions L25 to R25 (25 words on either side of the search word), and also provides a presentation of the most frequent collocates in each position.[9] In SARA, it is possible to obtain the frequency of a collocation where the collocates are given by the user. The Qwick program gives the user the option of choosing between a number of ways to compute collocational strength, and then identifies the collocations and their collocational strength.[10]

As WordSmith is the only program of the ones listed that can be used on all corpora in this study, this program was also used for the collocation study. Due to problems with the version of the program available, it was not possible to obtain reliable data for all expressions automatically. That means that the collocations had to be identified and counted manually. To be able to do this within the scope of the present study, only the most frequent items could be studied. The items chosen for this part of the study are high frequency main verbs and personal pronouns used as subjects of the expression.

3.1.2 Main verbs

When expressions of future are discussed, they are often referred to as *will/ (BE)going to* (etc) + infinitive. The infinitival verb is seen as part of the construction. The extent to which the different expressions can be found to collocate with certain frequent verbs is examined below. The main verbs chosen for this part of the study were selected on the basis of overall frequency, and thus the verbs *be, have, do* are found among them. In addition, the verb *get* was studied, as well as *go*. Other high-frequency verbs that were considered were *see, say* and *know.* In the following section, the main focus is on the collocations with the most frequent verb, *be.* The collocates with the other verbs have been examined, but will only be referred to in certain cases, where particularly interesting features were found.

In the corpora in this study, there are about 7,000 instances of *be* per one million words*,* which means that the proportion of *be* in the corpora is less than one per cent. Among the infinitival verbs in the CG corpus, *be* constitutes about 15 per cent, *have* about seven per cent, and *do* about five per cent.[11] That means that just over one fourth (27%) of all infinitival verbs in that corpus are one of the verbs *be, have,* or *do.* The proportion of the three verbs is higher among the infinitives used with the expressions of future than in the corpus as a whole. This is illustrated in Table 3a. Table 3b shows the main verbs used with *will* in the different corpora. The proportions refer to the proportions of the expressions of future that are found with one of these verbs.[12]

Table 3a: Distribution of infinitival *be, have, do* in the CG corpus. Proportions of the expressions of future used with the verbs

(raw frequencies)	**CG (22,558 infinitives)**	***will* (1,345)**	***'ll* (1,049)**	***shall* (122)**	***going to* (533)**	***gonna* (327)**	**Total (3,376)**
	%						
be	15	29	16	20	27	26	24
have	7	5	10	5	5	11	7
do	5	2	4	7	6	4	4
Total	27	36	30	32	38	41	35

Table 3b: *Will* used with *be, have, do* in the different corpora (percentages)

WILL	**LOB**	**FLOB**	**LLC**	**CG**	**DS**
be	34	34	27	29	21
have	5	6	8	5	3
do	1	1	1	2	5
Total	40	41	36	36	29

Table 3a shows that there is an unusually high proportion of the verb *be* co-occurring with the expressions of future. *Be* constitutes 15 per cent of the infinitival verbs in the corpus, but 29 per cent of the instances of *will,* for example, are used with *be.* Table 3b shows that the proportion of *be* used as main verb with an expression of future is high in all corpora. In order to evaluate to what extent the proportion differs from that in the corpora as a whole, it would be necessary to manually disambiguate all the instances of *be, have,* and *do* to decide how many are infinitives and how many are not, which was not possible to do within the scope of this study. [13]

Among the instances of expressions of future + *be*, a small proportion is followed by a present progressive form, as in [1]:

[1] Up and down the country husbands *will be saying* they would never behave like that. (LOB B04 123–124)

The proportion of this progressive infinitive with *will* is higher in the FLOB than in the LOB corpus, about eight per cent and five per cent of the *will+be* occurrences. In the spoken material, the construction is used more frequently, at least in the CG and LLC corpora, where it is found after about 13 per cent and 14 per cent of the instances of *will+be*. Leech (1987, section 107) suggests that 'one reason why the *will/shall*+Progressive usage has become quite common in everyday speech is that it is often a more polite and tactful alternative to the non-progressive form'. On the basis of this, it is somewhat surprising to find that the progressive infinitive is found with only about five per cent of the instances of *will+be* in the DS corpus of spontaneous conversation.

The proportion of the progressive infinitive is higher for *'ll* than *will*; 13 per cent of *'ll be* in the LOB and 17 per cent in the FLOB corpus are used with a present progressive form. Leech (1987) notes that the construction 'has become quite common in everyday speech' (section 107), which could also explain the relatively high proportion of the construction found to collocate with another construction known to be used in speech: *'ll.* It should be noted, however, that the raw frequencies of *'ll* are low in the written corpora, which means that the proportions vary greatly also with small differences in number.

The progressive infinitive with *shall* is not very frequent in absolute numbers. It is, for example, found only eight times in the LOB corpus and five times in FLOB (which would correspond to seven per cent and 15 per cent of the occurrences of *shall+be*).

The second most frequent verb used with the expressions of future is *have.* It collocates with about six per cent of the expressions of future in the corpora (slightly more in the LLC, less in the DS corpus). Of the instances of *have* in this position, about one third is the construction *have to*+infinitive, as in:

[2] Where all benefit, all *will have to* contribute. (FLOB J58 139–140)

The most frequent verbs found to follow the expressions of future, apart from *be* and *have,* are generally the same in both the written and spoken corpora, and with practically the same, rather low, frequencies. These collocates are verbs that otherwise are relatively frequent in the corpora, such as *get, take, see, give, make, come.* It is, however, only the verbs *be* and *have* that are found with more than five per cent of the instances of *will* (the figures for the other more frequent verbs start at around two per cent).

3.1.3 Personal pronouns

All the expressions are used with personal pronouns as subjects to a great extent. There are some differences between the corpora, both with regard to the proportions of personal pronouns as subjects and with regard to the choice of pronoun. It seems, however, that the difference between the corpora is smaller than the difference between the various expressions of future in this respect. Table 4 summarises the proportions of personal pronouns found with the five expressions in the different corpora:

Table 4: Expressions of future with personal pronouns as subjects (* no figures given for raw frequencies under 20)

	will	***'ll***	***shall***	***going to***	***gonna***	**Total**	**Personal pronouns in the corpora (raw figures per million words)**
	%						
LOB	33	96	61	68	*	48	41,600
FLOB	30	94	74	61	*	43	39,745
LLC	44	93	94	67	*	72	118,680
CG	46	94	86	67	69	68	91,130
DS	67	93	95	81	80	84	140,868

There is some variation between the corpora in the total proportion of personal pronouns used with the expressions of future. The proportions are lower in the written corpora and higher in the spoken. This, to some extent, reflects the fact that there are more personal pronouns in the spoken corpora overall. It is, however, not the case that the proportion of expressions of future used with a personal pronoun is directly proportional to the number of personal pronouns in the corpora.

The difference between the written and spoken corpora in the proportion of personal pronoun subjects is found primarily for the expression *will,* and to a somewhat smaller extent *shall.* For *'ll* and *going to*, however, the proportions of personal pronoun subjects are approximately the same in all corpora. The corpus with the highest overall frequency of personal pronouns is the DS corpus, and that is also where there is the highest proportion of personal pronoun subjects used with all expressions (except *'ll,* which has a very high proportion overall).

The most frequent expression in the study, *will,* is used least with personal pronoun subjects. Only about one third of the instances in the written corpora are used with this kind of subject. The proportions in the spoken corpora are higher, but still considerably lower than for any other expression. The highest proportions of personal pronoun subjects appear with the expression *'ll*, 93–96 per cent. This is hardly surprising against the background of what has been said about the subject, or host of contracted forms in general by, for example, Axelsson (1998) and Kjellmer (1998). *Shall* is also used with a personal pronoun as subject to a great extent, especially in the spoken corpora (where the expression is infrequent). *Going to* is used with personal pronoun subjects to a similar extent in all corpora except the DS, where the proportion is high (81%).

The different personal pronouns occur to a varying degree with the different expressions. Figure 4 illustrates the proportions of personal pronouns used with the various expressions in the written corpora (*gonna* is not included in the figure since the expression is so infrequent):

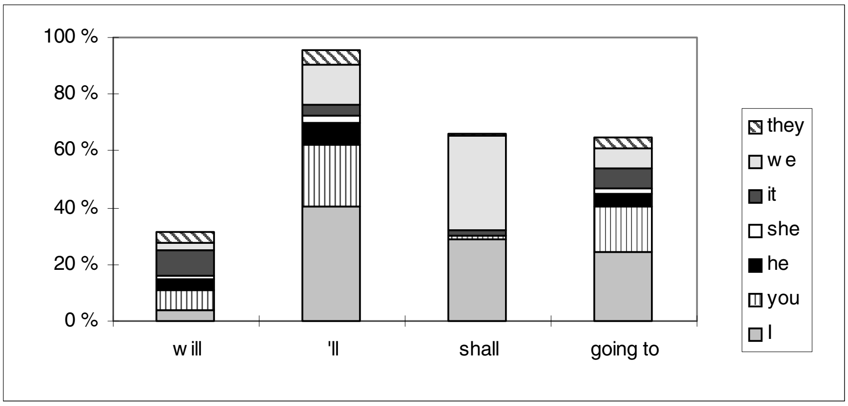


Figure 4: Proportions of personal pronouns used with the various expressions in the written corpora

It appears that *will* differs from the other expressions not only because it is used less with this kind of subject. It also appears that the expression is used considerably more with *it* than are the other expressions. The proportion of first person singular *I* used with *will* is considerably lower than the proportion found with the other expressions. As expected, *shall* occurs almost exclusively with the first person pronouns *I* and *we.* The two expressions *going to* and *'ll* show similar patterns of co-occurrence with the personal pronouns, with the exception that *we* is used more with *'ll,* while *it* occurs more with *going to*.

Figure 5 illustrates the distribution of expressions of future used with personal pronouns in the spoken corpora:

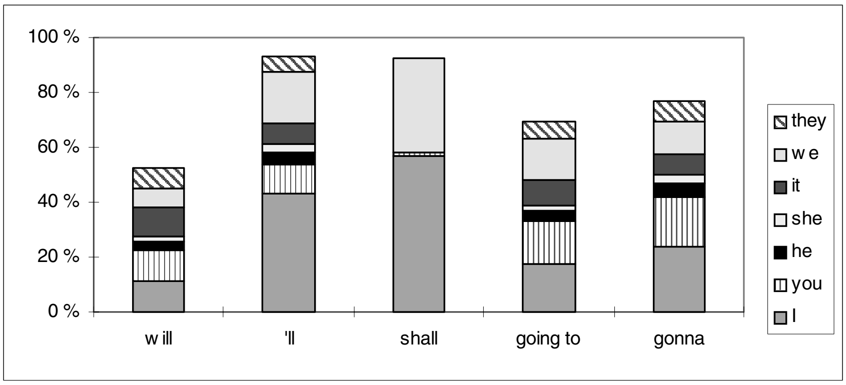


Figure 5: Proportions of personal pronouns used with the various expressions in the spoken corpora

As in the written corpora, *will* is used with *it* to a large extent. The proportion of *I* used with *will* is greater in the spoken corpora. That is also the case for *shall,* which is almost exclusively used with the first person pronouns in both the written and the spoken corpora. Unlike the other expressions of future, *shall* is often used in questions, and consequently followed by its subject, in questions such as:

[3] '*Shall* we?' he asked (LOB N29 120)

This is found in the spoken corpora in particular, and to a higher degree for the plural pronoun in all cases. In the written corpora, LOB in the first place, the use of *shall* with subjects other than pronouns often coincides with the use of *be*+past participle. These instances are often passives of a seemingly 'prescriptive' kind, giving orders or recommendations:

[4] The present Agreement *shall* be ratified and the instruments of ratification *shall* be exchanged at London as soon as possible. (LOB H14 90–91)

Going to and *'ll* were similar in the written corpora as far as the proportions of the different pronouns were concerned. The same holds for the spoken corpora. Axelsson (1998:164) notes that the first and second person pronouns '… act as contraction-promoting factors', so that the contracted variants are more frequent when the host is a first or second person pronoun, which is the case for the written data.

Table 5: *Will+'ll* with personal pronouns as subject. Proportion of *'ll* (percentages)

	LOB	**FLOB**	**CG**	**DS**	**LLC**
I	68	66	77	87	82
you	39	36	52	58	57
he	23	29	30	68	60
she	35	26	58	69	83
it	7	8	34	55	63
we	69	39	71	85	83
they	17	19	42	53	49
All personal pronouns	38	35	61	75	71

In Table 5, the frequencies for *will* and *'ll* used with the different personal pronouns have been combined, and the proportion of the instances of the contracted variant *'ll* is given. In the two written corpora, about two thirds of the occurrences of *I* are found with *'ll* rather than *will,* while only seven per cent and eight per cent of the instances of *it* are used with the contracted variant. In the spoken corpora, however, the differences are not as great. It is nevertheless clear that the first person pronouns *I* and *we* are found with *'ll* considerably more often than with *will,* while *it* and *they* are used less with *'ll.*

If the frequency of the expression *'ll* varies with the frequency of the different personal pronouns in a text or text type, as suggested by Axelsson (1998), the high frequency of *'ll* in fictional texts (see section 2.2) can be explained by a high frequency of the contraction-promoting first and second person pronouns.

The difference between the personal pronouns with regard to the extent to which they occur with the full form (*will*) or the contracted form (*'ll*) is not found for the expressions *going to* and *gonna.* Table 6 illustrates this:

Table 6: *Gonna* and *going to* used with personal pronouns as subject. Proportion of *gonna* (percentages; * = raw frequency for either form is under 20)

	CG	DS
I	45	75
you	42	78
he	*	*
she	*	*
it	33	76
we	32	73
they	49	72
Total	39	76

The main difference observed is that between the two corpora (CG and DS), where *gonna* is more frequent with all pronouns in the DS corpus. Thus it does not seem to be the case that first or second person pronouns promote the choice of the reduced form *gonna.*

3.2 Clusters

Clusters are a form of collocations where the collocating words are always found directly adjacent to each other in the same position. WordSmith Tools has a function named cluster, which finds clusters of a pre-defined length. In this paper, that function has been used to identify frequent clusters automatically.[14] However, the function does not work well on the LLC, as the texts contain mark-up that is not identified as such by WordSmith. Instances of a phrase including mark-up are not found to be similar to one without mark-up. As an illustration of the problem can be mentioned that the WordSmith cluster function identifies *it will be* as the most frequent cluster with *will*, occurring 14 times. A search for ‘it will be’ in the LLC, however, yields only nine hits, while a manual inspection of the concordance lines for *will* identifies 16 occurrences of the phrase. Although one of the aims of this study is to see what can be done with the corpora and tools as they are found on the CDs, it became obvious that the manual retrieval of possible clusters in the LLC would be too time-consuming to be performed within the scope of the present study. A stripped version of the corpus was therefore created and used for the identification of LLC clusters.[15]

Biber and Conrad (1999) introduce the term 'lexical bundles' defined as 'the most frequent recurring lexical sequences; … usually *not* complete structural units, and usually not fixed expressions' (1999:183). In their study of lexical bundles in conversation and academic prose, Biber and Conrad study clusters consisting of at least four words (where constructions such as *won't* are regarded as one word). To be considered a lexical bundle, the four-word combination has to occur at least 20 times per million words. That means that in this study, clusters found at least 20 times in the written corpora or ten times in the spoken corpora meet that frequency limit.

Among the clusters with expressions of future, the more frequent ones are, with a few exceptions, formed with a personal pronoun and one of the most frequent verbs, usually *be*. It is natural that frequent expressions and frequent pronouns and verbs are found in clusters more than the less frequent expressions. It is, however, not always the case that only the most frequent words are found in the clusters, as will be further illustrated below.

3.2.1 Will

It was shown above that the most frequent expression, *will,* often collocates with *be* and *have,* and that it frequently occurs with personal pronouns as subjects (even though the other expressions are used with personal pronouns more). There are great similarities between the two written corpora in what clusters are the most frequent, even though the frequencies differ slightly. The most frequent clusters with each expression are given in Table 7:

Table 7: Clusters with *will* in the different corpora (raw frequencies)

LOB	**FLOB**	**LLC**	**CG**	**DS**
it will be (87)	*it will be* (63)	*it will be* (21)	*it will be* (48)	*it will be* (20)
there will be (40)	*there will be* (38)	*there will be* (16)	*there will be* (21)	*there will be* (10)
will have to (30)	*will have to* (34)		*that will be* (20)	*I won't be* (12)
will not be (29*)*	*will not be* (38)		*will be able to* (12)	*won't be able to* (12)

will be a (22)	*will be the* (20)			*it won't be* (11)
	will be able (22)			*no I won't* (10)
	will be able to (20)			

In the written corpora, only *it* and *there* are found preceding *will* in the frequent clusters. *It* is the most frequent pronoun used with *will*, but for example *you* and *they,* which do not occur in the frequent clusters, are more frequent than *there,* which occurs in the clusters. The reason why these more frequent pronouns are not found in the high-frequency clusters could be that they are used with the less frequent, lexical verbs, while *it* and *there* are used to form semantically lighter constructions that can appear in a variety of contexts.

In the spoken data, there are similar clusters with *will* as in the written data, although the frequency varies between the corpora. It seems that the most frequent clusters are the same in the written and the spoken material, with the noticeable difference that the DS corpus also contains a number of clusters with *won't.* The construction *won't* is considerably more frequent in the DS corpus than in the other corpora, appearing as often as 453 times. This can be compared to the frequency of 144 instances in the CG, 107 instances in the LOB and 97 in the FLOB corpus. The corresponding non-contracted form *will not* occurs 105 times in the LOB, 111 times in the FLOB and 29 times in the CG corpus, compared with only six occurrences in the DS corpus. The extent to which negated forms, contracted or non-contracted, are used differently in different kinds of text is an interesting issue which, unfortunately, cannot be pursued further within the scope of the present study.

It is interesting that, although the expression *will* is rather frequent in these data, there are only two lexical bundles, as defined above. *Will be able to* is found 20 times in the FLOB and twelve times in the CG corpus, while the negated form *won't be able to* is the only lexical bundle in the DS corpus, where it occurs twelve times.

3.2.2 'll

The expression *'ll* differs from the other expressions in the study not only in that it is almost invariably found with a (personal) pronoun subject, but also in that the subject and *'ll* can only appear in one order: the pronoun directly followed by *'ll*. It is not surprising then that a number of clusters include the pronominal host, the expression, and the very frequent verb *be.* If constructions such as *I'll* are counted as one word, in accordance with Biber and Conrad (1999), there are not many lexical bundles, or four-word clusters in these data occurring more than 20 times per million words. In the CG corpus, only *we'll have a look* comes close to meeting the prerequisite, with nine occurrences in the 500,000 word corpus, and in the LLC the cluster *I'll give you a*, also occurs nine times. In the DS corpus, however, there are two lexical bundles: *I'll tell you what* (17 instances) and *I'll go and get* (15 instances). A common feature of these clusters is that none of them contain the most frequent verb *be,* but are used with other verbs. These verbs are common in the corpus, but far from the most frequent.

In the written corpora, there are no four-word clusters with *'ll* occurring more than three times. There are not even any frequent three-word clusters in the written data. Thus it seems that, although the expression *'ll* is used with such a relatively limited set of subjects, it does not form long high-frequency clusters but is found in a number of different constructions.

Table 8: Clusters with *'ll* in the different corpora (raw frequencies)

LOB	**FLOB**	**LLC**	**CG**	**DS**
you'll have to (11)	*I'll tell you* (10)	10 three-word clusters occurring >10 times	4 three-word clusters occurring >10 times	18 three-word clusters occurring >10 times
		I'll try and (18)	*we'll have a look* (9)	*you'll have to* (51)
		I'll give you (17)		*we'll have to* (34).
		I'll tell you (12)		*I'll tell you what* (17)
		I'll give you a (9)		*I'll go and get* (15)

In the spoken data there are a number of three-word clusters, in particular in the DS corpus. The *'ll* expression is about twice as frequent in the DS corpus as in the other spoken corpora. There are four times as many instances of *'ll* in the DS as in the written corpora, which are twice the size. The frequency of the expression can, to some extent, explain the large number of clusters in the DS corpus. That explanation cannot, however, be applied to the difference between the LLC and the CG corpus, which both contain about the same number of instances of *'ll*.

3.2.3 Shall

Shall is used with a limited number of subjects, as is also the case for *'ll*. As seen above, *'ll* clusters with the personal pronouns, but there were few long clusters (four words or more), especially in the written data. The expression *'ll* is fairly infrequent in the written data, which to some degree could explain the absence of long clusters. *Shall* is even less frequent than *'ll,* so if frequency alone were a determining factor for the number of clusters found, there would not be many clusters with *shall.* In the written corpora, *shall* is not found often enough in clusters of at least four words to qualify as a lexical bundle, as defined above.

Table 9: Clusters with *shall* in the different corpora (raw frequencies)

LOB	**FLOB**	**LLC**	**CG**	**DS**
we shall have to (8)	*the requested party shall* (8)	*I shall be* (18)	*I shall be* (5)	*I shall be* (10)
I shall be (16)	*I shall be* (9)	*I shall have to* (9)	*that the evidence I shall give* (12)	*shall I do* (9)
shall not be (12)			*shall be the truth* (12)	*I shall have to* (5)
we shall have (11)				
we shall see (11)				

None of the most frequent three-word clusters with *shall* in the written corpora occur more than 20 times. When the clusters found in the FLOB corpus are examined, it is clear that a great proportion of the instances of *shall* is found in a single text (an agreement between two countries). The highly specific phrase *the requested party shall* is found eight times. In addition to that, *I shall be* appears almost as often in the FLOB (9 times) as in the LOB (10 times). The other, relatively frequent clusters with *shall* found in the LOB are not found in the FLOB corpus.

In the spoken data, *I shall have to* is the most frequent four-word cluster in the DS and the LLC, occurring only five and nine times, respectively. The three-word clusters in the DS corpus are similar to those in the written data, for example, *I shall be.* In the LLC, *I shall be* is comparatively frequent, occurring 18 times. In the CG corpus, however, the cluster, which was relatively common in the other corpora, occurs only five times. There are, however, two clusters that are frequent enough to qualify as lexical bundles according to the frequency and size definition given in Biber and Conrad (1999) referred to above. The clusters are *that the evidence I shall give* (12) and *shall be the truth* (12). Even before consulting the actual texts where the clusters are found, it is obvious that these two clusters are highly specialised. They occur in two texts in the CG corpus, JJV and JJW, both recordings of a court case (O'Halloran vs Chief Constable of Bedfordshire). The clusters form a part of the oath sworn by people on the case, and the high frequency is a result of the fact that the phrases are first read by one person, (preceded by 'Take the book and repeat ...') and then repeated by another. There is thus one more reason not to consider these clusters lexical bundles, or '... sequences of words that commonly go together *in natural discourse'* (Biber and Conrad 1999:184, my emphasis).

3.2.4 Going to

If *going to* and orthographic units such as *I'm* are counted as one word, there are no particularly frequent clusters with *going to* in the written data in this study. Clusters are more frequent in the spoken corpora where the expression occurs more, particularly in the LLC and CG, as can be seen in Table 10:

Table 10: Clusters with *going to* in the different corpora (raw frequencies)

LOB	**FLOB**	**LLC**	**CG**	**DS**
is going to be (9)	*is going to be* (10)	*is going to be* (31)	*is going to be* (28)	*I'm going to get* (9)
are you going to (11)		*it's going to be* (27)	*it's going to be* (22)	*it's going to be* (9)
		are going to be (27)		*I'm not going to* (7)

It seems as if the distribution of the clusters varies between the spoken corpora. The expression *'ll* was frequently found in clusters in the DS corpus, while there are few clusters with *going to* in that corpus. Clusters with *going to*, on the other hand, are found more in the CG corpus and the LLC, where the *'ll* clusters were fewer. There are no lexical bundles with *going to* in this material.

3.2.5 Gonna

Gonna is most frequent in the DS corpus where the expression occurs in some lexical bundles. *I'm gonna have to* and *we're gonna have to* appear 13 times each. In the CG corpus, the slightly odd four-word cluster *party and I'm gonna* is found nine times. A closer study of the cluster shows that it is actually the longer *I'm going to the party and I'm gonna take.* All the examples, and some variants with *going to* instead of *gonna,* are found in one text that, among other things, includes the playing of a game where the participants are to say this particular phrase. That means that they can hardly be considered lexical bundles, since they neither occur in 'natural discourse' nor in 'multiple texts' (see section 3.2 above).

3.3 Summary – clusters and collocations

This study has shown that the five expressions investigated are similar in that they often collocate with personal pronouns and infinitival *be.* There are, however, great differences between the expressions. *Will* is used less with personal pronouns than the other expressions, while *'ll* collocates with personal pronouns most, in over 90 per cent of all cases. *Will* is used more with *it,* while *shall* almost exclusively appears with *I* and *we*. There are also differences between the corpora, so that the proportion of personal pronouns is higher in the spoken data, and collocates more with the expressions in the spoken corpora. Apart from *be* and *have,* no other verbs collocate with the expressions in more than two per cent of the cases. The expressions are thus used to a great extent with the so-called function verbs *be* and *have*; they also occur with a considerable number of different lexical verbs. It could be claimed that the expressions are to some extent specialised (used much with a limited number of verbs), but that they are also non-specialised and used in a number of different collocations.

The fact that personal pronouns and the verb *be* are frequent collocates of the expressions of future is to some extent reflected in the clusters where the expressions occur. The number of clusters seems to vary with the frequency of the expression and the other words in the clusters, but it is not always the case that the most frequent collocates are also found in the most frequent clusters. As examples of this has been mentioned that *will* clusters more with *there* than with the more frequent pronoun collocate *you,* and that the clusters with *'ll* in the LLC contain some, comparable, infrequent main verbs. This suggests that, although the number of clusters co-varies with the frequency of the clustering items in the corpora, that is not the only influential factor.

The most frequent clusters that the expressions occur in are often similar in all corpora. The frequencies differ considerably, both between the corpora and, in particular, between the expressions. It seems that the frequency of the expression is reflected in the clusters to a great extent: a frequent expression is found in many clusters, while infrequent expressions cluster less. However, there does not appear to be an absolute relationship between the frequency of an expression in a corpus and the number of clusters where it occurs. It may be interesting to note that the main verbs co-occurring with the expressions in the clusters differ. *Will* and *going to* cluster with *be,* but only *will* is found with *be able. Shall* clusters more with *have*, while *'ll* occurs with a greater variety of verbs than the other expressions.

The study of collocations and clusters points to interesting features of variation between the expressions and corpora that merit further exploration. In the following section, the focus will be on variation with the three features of time, medium, and genre.

4 Discussion

4.1 Time

According to the present study, the major development concerning these expressions of future during the latter part of the 20th century is that the use of *shall* has decreased. The proportion of the expression is lower in the FLOB corpus from 1991 than in LOB from 1961. In the spoken corpora, the proportion of *shall* is twice as high in the LLC (8%) as in the other two corpora (4%). This could further support the interpretation that there has been a decrease in the use of the expressions over time, as the expression is found more in the earlier LOB corpus and also in the LLC, where the texts are somewhat earlier than in the Sampler. It is not only the frequency of *shall* that has decreased; it also seems that the use of the expression has changed. In the FLOB corpus in particular, the expression is primarily found in a few texts, and it is often used in quoted contexts. This could be interpreted as the expression having become more marked, or less general. It is also interesting to note that the expression occurs to a relatively high degree in clusters in FLOB, a further indication that the expression is not generally used but found primarily in specialised contexts or constructions.

A possible further difference between the earlier data (LOB, LLC) and the later (FLOB, Sampler) is that the use of *gonna* appears to have increased over time. The difference is very small between the written corpora, at least in absolute numbers, while the difference between the spoken LLC and the Sampler corpora is enormous: <one per cent in the LLC to be compared to 18 per cent of all expressions of future in the DS. As discussed above (section 2.3), this could point to a recent increase in the use of the reduced form, but it must be noted that this difference, at least in part, is due to differences in transcription practices or the level of formality in the texts.

It has been argued that the *going to* construction is spreading, '...as a substitute for the pure future, pushing out the forms with *shall* and *will'* (Danchev et al 1965:380). There is no clear evidence of this in the written data, where the *going to* expression is used to a similar, low, extent in the two corpora. A possible tendency might be seen in the Imaginative hyper-category, where the proportion of *going to* is higher in the FLOB than in the LOB corpus. The combined proportion of *going to* and *gonna* is also lower in the LLC than in the Sampler, which may indicate that if the expression is spreading, it is first noticed in the spoken and more speech-like data, such as fiction.

4.2 Medium

It has repeatedly been pointed out above that there are great differences between the spoken and written corpora where the use of expressions of future are concerned. The expressions are considerably more frequent in the spoken data. There is more *going to* and *gonna* in the spoken corpora, and also a considerably larger proportion of *'ll*. The proportion of *will* is, consequently, lower in the spoken corpora than in the written. No consistent difference between the spoken and written corpora can be found where the expression *shall* is concerned. The proportion is lower in the LLC than in LOB, but higher than in FLOB. When the earlier LOB and LLC are compared to each other, as well as the later FLOB and Sampler corpora, a pattern of variation can be discerned, indicating that *shall* is used less in the spoken corpora from the same time.

When the collocations and clusters with the various expressions are studied, some similarities between the written and spoken data emerge. The most frequent clusters with *will* in the written data*,* for example, are also found to about the same extent in the spoken CG component. The proportion of *'ll* used with personal pronouns is similar in the two media, although the proportion of the expression varies greatly.

4.3 Genre

When the distribution of the expressions of future is studied across the genres, it is apparent that the difference is greater between the hyper-categories in a corpus than between corpora of the same medium. The frequency and proportions of the expressions differ considerably between the Imaginative and Informative hyper-categories in the written data, and between the CG and DS components of the spoken Sampler. This would indicate that the use of expressions of future, to a great extent, is decided by the context where the expression is used: *will* is used most in written, Informative text, while *'ll* is found more in the Imaginative writing. *Gonna* is found primarily in spontaneous conversation, where*'ll* is also frequent, while *will* occurs more in the more formal spoken component.

Shall occurs to a great extent in specialised genres, and even in specific texts. The text category with the highest proportion of *shall* in both corpora is category H, Miscellaneous. In LOB, 95 out of 354 (27%), and in FLOB 43 of 197 (22%) of the instances are found in Category H (Miscellaneous). Within that category in FLOB, the majority of the instances of *shall* occur in one single text, a text from an agreement between the governments of the United Kingdom and Italy.[16] In that text, *shall* is used with third person subjects, such as *the (requesting/requested) party* (34 instances), *this Agreement, this Article.* In the LOB corpus, *shall* is particularly frequent in a couple of the texts in category H, with third person subjects such as *the Agreement, Council,* and *this section.* The proportion of *shall* in Category P, Romance and love story, is equally high in both corpora, while it is lower in most other categories in the FLOB corpus. One explanation of this lies in the content of the texts in the category. Some of the text extracts are from romantic novels that are set in earlier times than the present, and the characters are portrayed as using language that may seem dated to modern readers.

[5] 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)

Ten of the 30 instances in category P are from the text P01, (*Sweet sacrifice,* by Elizabeth Bailey), in sentences such as the following:

[6] .. Still, you may wear the sprigged walking dress and the blue pelisse. Murray *shall* lend you my chinchilla muff, and (FLOB P01 148–150)

[7] Oh, tush, Southern! I was only trying to divert her. Dear Clementina, you *shall* tell us nothing at all if you dont wish to. (FLOB P01 86–88)

5. Final remarks

This study had the primary aim to study expressions of future. The results of the study suggest that the use of expressions of future varies with medium (written/spoken) and genre (Informative/Imaginative) to a great extent. The variation with time is less noticeable, except for the expression *shall,* the use of which seems to have decreased in frequency. The expression *gonna* is found more in the later, spoken corpora, which might be an indication that the use of the expression is increasing, even if it cannot be excluded that the increased frequency can be explained by other factors. The study has also shown that the expressions of future are similar in that they all frequently co-occur with personal pronouns and the main verb *be.* There are differences between the corpora in this respect, but the main difference seems to be between the expressions. *Will* is used less with personal pronouns than the other expressions. When used with such a pronoun, the collocate is often *it. Shall* is almost exclusively used with the first person pronouns *we* and *I.* The proportion of other subjects used with *shall* is larger in the written corpora. The expression *'ll* is used with personal pronouns to a very large extent, and most frequently with *I.* The proportion of expressions used with *we* is generally highest in the CG corpus.

A secondary aim of the study has been to see to what extent the new corpus CDs can be used for a study of this kind. The study has shown that the resources on the BNC Sampler CD (1999) and ICAME CD (1999) can be combined, and that the combination can be exploited successfully for a study of variation in corpora from different times and of different mediums. However, it is necessary to be aware of, and be able to compensate for differences in the corpus formats. Studies such as this, involving corpora of different kinds that are searched with software not created for use with the particular corpora, would benefit greatly from some pre-editing of the corpus files. The amount of manual work would have been considerably less if the corpora had been converted into a format where all text not forming part of the actual corpus text (such as corpus headers, line references, prosodic annotation) had been removed or suitably tagged in a format recognisable by WordSmith Tools.

An addendum: experiences of using the tools and corpora

Tools

This study has been based on new and easily available resources: the ICAME CD and the BNC Sampler CD. The two CDs contain both corpora and corpus handling software. On the Sampler CD four programs are included; SARA created specifically to be used with the BNC, WordSmith Tools, Qwick and Corpus Work Bench. The Corpus Work Bench is for use in a UNIX environment, and has not been used for the present study. Similarly to the Sampler, the ICAME CD contains WordSmith and Qwick (the same versions as on the Sampler CD). A number of other programs can also be found on the ICAME CD; Lexa, Linguafont, TACT, and WordCruncher. The latter have, however, not been used for the present study.

To get an indication of one aspect of how the search programs on the CDs differ, a simple test was carried out. The expression *shall* was searched for in the different corpora by the various programs. The result is that, as far as this simple function is concerned, there are no differences between the programs (see Table 11).

Table 11: Results of searches for 'shall' in different corpora by different programs

– = cannot be searched with the program
( ) = can be searched with the program but the user has to identify which files are written/spoken, CG/DS etc

	WS	**Qwick**	**SARA**
LOB	349	–	–
FLOB	197	197	–
LLC	210	–	–
DS	(171)	171	171
CG	(120)	120	120
Sampler	291	291	291

The usefulness of a corpus-handling tool, of course, does not lie only in its ability to find the instances of a word or phrase, as there are other features that are equally important. As mentioned above, all the programs have features that make possible the identification of collocations. The concordance lines can be sorted, unwanted examples can be deleted and various other functions can provide further information. As these functions differ between the programs, a combined use of more than one program can be a useful way to solve different tasks. It is, for example, much easier to make searches in the CG and DS components separately in Qwick than in WordSmith and/or SARA, but the text references are easier to interpret in WordSmith and SARA. The collocation function in SARA is only available for collocates specified by the user, but the program does allow the user to search for utterances made by socio-linguistically defined speakers (for example of a particular age or social class). Unfortunately, not all the programs on the CDs can be used with all corpora. SARA can be used only with the Sampler corpus. The Sampler can also be searched with Qwick, and so can the FLOB corpus. LOB and the LLC have not been indexed to be used with Qwick at the time of writing (early 2000), but WordSmith can be used on these corpora as well as on the FLOB and Sampler corpora. That means that of the three corpus handling tools listed here, only WordSmith can be used on all the corpora in the study.

Corpora formats

It may be interesting to note that, even when the same program is used, the search and retrieval process has to be varied depending on which corpus is being searched. That is not a feature of the tool, but lies primarily in the formats of the corpora. It is important to be aware of what the texts in the corpora look like, and how the corpora differ with regard to how certain words and phrases are rendered. In LOB and FLOB, for example, *won't* is typed ‘won't’ while it in the LLC is found as ‘won`t’, with the accent ‘`’ rather than the apostrophe ‘’’. In the Sampler, the construction is regarded as two words, *wo n't.* Not realising this means that a search may fail. A search for ‘won't’ in the LLC, for example, will return 0 hits, while a search for ‘won`t’ finds 119 instances.

The LOB and FLOB corpora are available in a format where the line reference is provided in plain text with the corpus text (see [8]).

[8] C01 8 |^While never minimising the immensity of her work, it lifted the

C01 9 saintly halo which usually surrounds her name to reveal a warm,

C01 10 dedicated person who accomplished most by perseverance and hard work. (LOB C01 8–10)

These references are thus recognised as part of the text by WordSmith. This causes problems, for example when the search word is a phrase (such as *going to*). If the program simply were asked to find ‘going to’, instances with interfering line references would be lost. Instead the search has to be formulated as ‘*going* followed by *to* within three words’, to make sure that the instances interrupted by line references (counted as two words by WordSmith) are also found. This means, of course, that all instances of *going* followed by *to* within three words are found, also those where the phrase is not interrupted by a line reference, such as:

[9] ... he could hear him *going* on speaking *to* her,... (LOB K06 25)

These instances have to be removed before the analysis of the relevant examples can commence.

The line references also cause problems in another respect, as can be seen when the cluster function is used. One of the most frequent four-word clusters with *going to* in LOB, for example, is *n # # going to* (where the numeric line reference is substituted by # # by the program). That the cluster is identified is not a big problem, as it will be ignored by anyone evaluating the display. What is a greater problem is, of course, that potential clusters are missed when the line reference is interpreted as part of the text.

Similar problems are found when the LLC is searched. Here, however, it is not only line references that cause the problem, but also the speaker reference codes. The prosodic annotation of the spoken corpus is a further problem, as it is also found within the words. In [10], for example, the phrase *I should like to* occurs, but due to the prosodic annotation within 'like', the phrase would not be found if the search word were 'like':

```
[10]  3 1 a  2  170 1 1 A    11 1I should ^l\/ike to#               /
      3 1 a  2  180 1 1 A    11 1^\anyhow#                  /       (LLC 31a:170,
      180)
```

This means that even if a particular word or phrase is in the corpus, a search for it may not identify it. For the present study, searches have been made for substrings of the expressions studied, to identify possible variant renderings. The concordances were scanned manually in order to identify the collocating items.

The Sampler texts are annotated with SGML mark-up. [11] shows a part of the Sampler corpus:

[11] <s n=0004 p=Y><w AT>The <w NN1>DAX <w NN1>index <w IO>of <w NP1>West <w NP1>Germany<w GE>'s <w MC>30 <w JJ>leading <w NN2>shares <w VVD>surged <w MC>33.73 <w NN2>points<c YCOM>, <w CC>or <w MC>2.3 <w NNU>per cent<c YCOM>, <w II>to <w MC>1,496.69 <w II>in <w MC1>one <w IO>of <w AT>the <w JJT>strongest <w JJ>daily <w NN2>peformances <w II>in <w JJ>recent <w NNT2>months<c YSTP>. </s> (BNC Sampler A87: 0004)

To benefit maximally from the extensive annotation of the corpus, the SARA program should be used for searches in the Sampler. It is, however, also possible to search the corpus with the other two programs. Qwick needs a specifically indexed version of the corpus (provided with the program on the CD), while WordSmith can be used on the original text. The SGML tags do not seem to cause any problems when WordSmith is used, as they are ignored by the program by default.[17] If one wishes to search the corpus with the tags, this can be done by disabling the 'ignore tags' option. This means that also the brackets, '< >', and text inside the brackets will be read as text and analysed by the program.

One problem with the Sampler, which is apparent when WordSmith is used on the corpus, is that each text includes a header with an identical, fairly extensive section containing information about the corpus, user restrictions, etc (about 250 words). This is rendered as text, and as such searched by WordSmith, included in frequency counts, concordances, word lists etc. As the corpus information is the same in all texts, any expression or phrase found there in one text can be found also in the others, which means, for example, that very frequent, long clusters will be identified. An illustration of this is that the word 'project', which occurs 300 times in the spoken component of the BNC Sampler. Of the concordances, two occur 98 times each: *British National Corpus project…*, and *See the project description in the corpus header....* A number of the instances are also found in the descriptions/titles of individual files in the headers, which means that more than two thirds of the identified instances of *project* are actually NOT used in spoken language, although retrieved by a search in the spoken corpus.

The FLOB corpus contains some SGML mark-up, although to a much smaller extent than the Sampler. The LOB corpus does not contain SGML mark-up as such, but the brackets '< >' have been used to mark, for example, titles, as in example [12]:

[12] R01 20 *<*1Foreign contacts*>

R01 21 |^*0When Jones goes abroad, he does not go as a member of any

R01 22 group, delegation or coach party. ^He goes alone. (LOB R01 20–22)

If the default option 'ignore tags' is not de-activated, WordSmith will not retrieve instances of a search word found within the brackets, so the number of identified instances will vary depending on whether the 'ignore tags' option is activated or not.

A further feature of the LOB corpus that may cause problems, or at least confusion, is illustrated in [12]. The words *Foreign* (line 20) and *When* (line 21) are immediately preceded by 1 and 0, respectively, and as the figures are identified as part of the word, searches for 'foreign' and 'when' will not identify these instances. The way to find these are by searching for '*foreign' and '*when', which will then identify also any other words including the string 'foreign' and 'when'. This is not a big problem quantitatively. Among the 2,326 occurrences of *will* in the LOB corpus, for example, there are 20 instances that are not found by a search for 'will' (no additional instances of *'ll*, *shall*, *going*, or *gonna* are found by adding the wild card * to the search word).

Overall, the availability of WordSmith makes it possible to search and compare the ICAME and Sampler corpora relatively easily. However, as this overview has shown, it is important to be aware of potential pit-falls caused by features of the corpora. Some of the problems encountered in this study could be solved quite easily by editing the corpus files, removing or marking headers (Sampler), line and speaker references (LOB, FLOB, LLC), and prosodic annotation for studies not concerned with pronunciation or prosody (LLC).

Notes

1. The total number of words exceeds 500,000, although it is difficult to estimate by how much. The exact number of words in the corpus is not given in the manual, where it is stated that '[i]n cases where one or more participants had knowledge of the recording (and had the task of keeping the conversation going), ... [t]heir contributions to the conversation have not been included in the total 5,000 words of each text..' (LLC manual).
2. The texts in the context-governed component of the BNC Sampler are from 1982 (one text), 1986 (two texts), 1988 (one text), 1990 (one text), 1992 (seven texts), 1993 (27 texts), and 1994 (four texts). For eight of the 51 texts in the component, the date of recording is unknown.
3. See, for example, Wekker (1976), Quirk et al (1985), Leech (1987).
4. As not all corpora in the study are tagged so that the simple present forms can be identified, there is no way to identify them all, other than by searches for the individual lexical items. The present progressive forms can be found by searching for '-ing', but that search over-generates data as it identifies all words with the ending.
5. In this context, it can be interesting to note that there is a difference between the tagged and untagged versions of LOB, and that the automatic POS-tagging of text is usually about 96–98 per cent correct. According to Knut

Hofland, who kindly answered my cries for help, there is also a slight difference in the frequencies of *will* in the two versions of LOB. 2,271 instances are found in the untagged version, compared to 2,382 in the tagged.

6. Some illustrative quotes:
   'The infrequent modal *shall* is used (…) to indicate futurity, but only with a first person subject' (Quirk et al 1985:4.42).
   '... used with *I* and *we* to express future tense, and with other words in promises or statements of obligation' (*The Oxford guide to the English language* 1984:498).
   'In the second and third person, *shall* can be used to express the speaker's **promise, threat, consent,** etc. ... This usage is unusual in modern English' (Svartvik and Sager 1977: 44:A, my translation).
7. 'Gonna' is in this paper used to refer to instances of *gonna* occurring with or without the auxiliary *BE.*
8. The hyper-categories are formed on the basis of purely extra-linguistic considerations and are, as such, to be considered *genres* rather that *text types* in the tradition following Biber (1988). The Imaginative hyper-category is formed of the text categories K–R, and the Informative of categories A–J.
9. Unfortunately there seem to be some problems with this function in version 3, which means that the number of collocations found is not always correct. Mike Scott informs me that this problem will be solved in version 4 of the program.
10. Collocational strength can be measured in different ways, which have been found to have their flaws and merits (see, for example, Stubbs 1995). Usually the measurement is based on the frequency of an item in the corpus as a whole, put in relation to the frequency of that item co-occurring with another item. A high value means high collocational strength: the items are found together to a greater degree than what could be expected from their frequency in the corpus.
11. These proportions have been retrieved for the CG corpus as an illustration. The Sampler texts are tagged so that infinitival forms of verbs can be distinguished, but the corresponding proportions cannot be found in LOB and LLC without a considerable amount of manual work.
12. The proportions are based on the total frequencies of the expressions of future in the corpora. No adjustment has been made to compensate for the extent to which the expressions occur without an overt infinitive verb, as in 'One thing that's certain is that you *won't*' (LOB A26 124).

13. There are about 35,000 instances of *be, have,* and *do* in the LOB, FLOB, and LLC corpora (which are not tagged so that infinitival verbs can be distinguished).
14. The frequency of the clusters given by WordSmith does not always match the frequency that is found if the cluster is entered as a query. Where there are discrepancies, the frequency quoted is the one obtained by searching for the cluster, not the one calculated automatically.
15. It is possible to create a stripped version of the corpus relatively easily by using some of the functions in, for example, WORD. I am, however, grateful to Klas Prütz for providing me with PERL scripts and UNIX assistance, thereby making the process much quicker and simpler.
16. *Agreement between the government of the United Kingdom of Great Britain and Northern Ireland and the government of the Italian republic concerning mutual assistance in the relation to traffic in narcotic drugs or psychotropic substances and the restraint and confiscation of the proceeds of crime.* Rome, 16 May 1990. London: HMSO. 1991:3–7.
17. The default option in WordSmith is that tags of the format '<*>' are ignored.

References

Primary sources
Corpora

British National Corpus Sampler. Distributed by BNC Sales/HCU, http://info.ox.ac.uk/bnc/getting/sampler.html

Freiburg LOB Corpus of British English (FLOB). Distributed by ICAME, http://www.hd.uib.no/icame/newcd.htm

Lancaster-Oslo/Bergen Corpus of British English (LOB). Distributed by ICAME, http://www.hd.uib.no/icame/newcd.htm

London-Lund Corpus of Spoken British English (LLC). Distributed by ICAME, http://www.hd.uib.no/icame/newcd.htm

Software

Qwick, version 1.0; http://www.clg.bham.ac.uk/QWICK/

SARA, version 0.931; http://info.ox.ac.uk/bnc/sara/index.html

WordSmith Tools, version 3.0; http://www.liv.ac.uk/~ms2928/index.htm

Secondary sources

Aijmer, Karin. 1984. *Go to* and *will* in spoken English. In H. Ringbom and M. Rissanen (eds) *Proceedings from the Second Nordic Conference for English Studies*, 141–157. Åbo: Åbo Akademi.

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

Axelsson, Margareta Westergren. 1998. *Contractions in British newspapers in the late 20th century.* Studia Anglistica Upsaliensia 102. Uppsala: Acta Universitatis Upsaliensis.

Berglund, Ylva. 1997. Future in Present-day English: Corpus-based evidence on the rivalry of expressions. *ICAME Journal* 21:7–19.

Berglund, Ylva. 1999. Exploiting a large spoken corpus: An end-users way to the BNC. *International Journal of Corpus Linguistics* 4,1:29–52.

Berglund, Ylva. Forthcoming a. *Gonna* and *going to* in the spoken component of the British National Corpus.

Berglund, Ylva. Forthcoming b. '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.

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

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

Biber, Douglas and Susan Conrad. 1999. Lexical bundles in conversation and academic prose. In H. Hasselgård and S. Oksefjell (eds) *Out of corpora. Studies in honour of Stig Johansson*, 181–190. Amsterdam: Rodopi.

Burnard, Lou (ed). 1995. *Users' reference guide to the British National Corpus*. Oxford: Oxford University Computing Service.

Bybee, Joan L. 1987. The evolution of future meaning. In A. G. Ramat, O. Carruba, and G. Bernini (eds) *Papers from the 7th International Conference on Historical Linguistics*, 109–122. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Collins, Peter. 1987. *Will* and *shall* in Australian English. In W. Bahner, J. Schildt, and D. Viehweger (eds) *Proceedings of the Fourteenth International Congress of Linguistics II*, 181–199. Berlin: Akademie-Verlag.

Danchev, A., A. Pavlova, M. Nalchadjan and O. Zlatareva. 1965. The construction *going to + inf.* in Modern English. *Zeitschrift für Anglistik und Amerikanistik* 13:375–386.

Haegeman, Liliane. 1989. *Be going to* and *will*: A pragmatic account. *Journal of Linguistics* 25: 291–317.

Johansson, Stig and Knut Hofland. 1989. *Frequency analysis of English vocabulary and grammar.* Vol. 12. Oxford: Clarendon Press.

Kjellmer, Göran. 1998. On contraction in Modern English. *Studia Neophilologica* 69:155–186.

Krug, Manfred G. 1998/1999. *Emerging English modals*. Unpublished PhD dissertation, Albert-Ludwigs-Universität, Freiburg (Wintersemester 1998/ 1999).

Leech, Geoffrey. 1987. *Meaning and the English verb.* London/New York:Longman.

The Oxford guide to the English language. 1984. London:Guild Publishing.

Poplack, Shana and Sali Tagliamonte. Forthcoming. The grammaticization of *going to* in (African American) English.

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

Stubbs, Michael. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language* 2, 1:23–55.

Svartvik, Jan and Olof Sager. 1977. *Engelsk universitetsgrammatik.* Almqvist & Wiksell Förlag AB.

Wekker, Herman Ch. 1976. *The expression of future time in contemporary British English.* Amsterdam: B.V. Noord-Hollandsche Uitgeversmaatschappij.