BE GOING TO versus WILL/SHALL

Does Syntax Matter?

BENEDIKT SZMRECSANYI
University of Freiburg, Germany

This study offers a novel account for the variation between the two major syntactic options to express futurity in English, BE GOING TO and WILL/SHALL. The focus of attention, unlike in many previous studies, is chiefly the choice that speakers of American and British English make between future markers with reference to syntactic characteristics of the surrounding text. On the basis of an empirical analysis of spoken data, this study demonstrates that future marker distributions seem to be sensitive to four factors: (1) contexts of negation, (2) contexts of subordination, (3) IF-clause environments, and (4) sentence length. More specifically, there is a positive correlation between syntactic complexity and the likelihood of the occurrence of BE GOING TO instead of WILL/SHALL. The analysis proposes that an issue with economy and online-processing constraints might be responsible for the sensitivity of future marker distributions to syntactic context.

Keywords: *English future markers; syntax; corpus linguistics*

Thanks to the intriguing nature of the subject matter, the body of literature dealing with the two syntactic options for overtly expressing futurity in English, WILL/SHALL and BE GOING TO,[1] is truly sizable. While space limitations preclude the possibility of going into much detail here, a brief review of the extant literature reveals that previous studies primarily deal with proposals for semantic and/or pragmatic differences between BE GOING TO and WILL/SHALL, with stylistic, regional, or sociolinguistic variation, or with discussions of frequency, both synchronically and diachronically. WILL/SHALL, in all, is commonly agreed to be the unmarked or simplex future, making a "plain statement about the future" (Close 1988, 51), albeit with a possible overtone of volition or obligation (cf. Kytö 1990, 277; Wekker 1976, 40). BE GOING TO is typically taken to suggest "prior intention, imminence, or inevitability" (Nicolle 1997, 355), "dynamic current orientation" (Haegeman 1983, 157), "future culmination of present intention or cause" (Haegeman 1989, 293; similarly, Nicolle 1997, 373), immediate or proximal futurity, inceptive present, and intentionality (Binnick 1971) or, simply, that there are "indications in the present that something will happen" (Wekker 1976, 124). It has also been argued that whatever the difference between BE GOING TO and WILL/SHALL is, it must be pragmatic rather than truth-conditionally semantic (cf.

Journal of English Linguistics, Vol. 31 / No. 4, December 2003 295-323
DOI: 10.1177/0075424203257830

Haegeman 1989) or that the difference between the two coding devices should be referred to as an aspectual one (e.g., Kortmann 1991).

Previous research has established that the more informal the setting, the more speakers tend (1) to use contracted/cliticized future marker variants (such as *won't* or *'ll*; cf. Close 1988, among others) and (2) to use BE GOING TO instead of WILL/SHALL (e.g., Berglund 1999, 2000a, 2000b; Close 1988; Mair 1997b; Wekker 1976). As far as regional variation—British English versus American English—is concerned, BE GOING TO has been shown, all other things being equal, to be consistently more frequent in American English than in British English (cf. Biber et al. 1999; Hundt 1997; Mair 1997a; Tottie 2002, among others). Overall, study after study has maintained that frequency-wise, BE GOING TO—no matter what register and variety—is consistently outnumbered by WILL/SHALL, more clearly so, however, in written varieties than in spoken varieties (cf., for instance, Berglund 1997, 1999, 2000b; Biber et al. 1999; Mair 1997a; Martin and Weltens 1973; Wekker 1976). A number of scholars have also suggested that BE GOING TO has actually been spreading over time (e.g., Berglund 1999; Danchev et al. 1965; Hundt 1997; Mair 1997a, forthcoming).

Yet what is absent from the literature is a systematic investigation into how the choice of one or the other future marker paradigm might be induced by the syntactic environment as another relevant intralinguistic variable, besides semantics and pragmatics. This minimal attention to an, I believe, actually interesting question is surprising, given that linguistic common sense alone would suggest a potentially significant role of syntactic context. In fact, a number of authors are on record with claims as to the preference for certain future markers in certain syntactically subordinated clauses, albeit from a rather theoretical, intuition-based perspective (e.g., Binnick 1971; Comrie 1982, 1985; Danchev et al. 1965; Declerck 1991; Hall and Hall 1970; Wekker 1976). The bottom line of such studies is that WILL/SHALL sometimes renders temporal and conditional subclauses bad, whereas BE GOING TO is always possible where its meaning is appropriate. Suggestions along these lines, however, by and large, still lack empirical scrutiny. The only studies I am aware of that take a rigorously data-driven approach to the syntactic behavior of future markers are Berglund's (1999, 2000b) collocation pattern studies. In these, Berglund analyzed the British National Corpus and found that BE GOING TO markers are significantly more often negated than WILL/SHALL.

In the present study, I attempt to fill the gap that now exists in the literature by systematically investigating whether, and to what extent, there are correlations between future marker distributions and their syntactic environment in spoken discourse. The research questions that will guide the present study can be summarized as follows:

1. What can be said about overall frequencies of future markers? More specifically, how do frequencies established in this study tie in with previous research, and are there any major differences between American English and British English, on one hand, and between formal and informal spoken English, on the other hand?
2. Previous research (e.g., Berglund 1999, 2000b) has indicated that BE GOING TO might be preferred in contexts of negation. Can this finding be replicated with the method and the data used in this study, and are there differences between varieties and/or registers?
3. Danchev et al. (1965) have stated that whenever BE GOING TO is used in subclauses,[2] the notion of intention is less dominant than when used in main clauses. Hence, are there differences between text frequencies of BE GOING TO and WILL/SHALL depending on whether they are embedded in syntactically dependent environments or in syntactically independent environments? Again, are there register differences or differences between American English and British English?
4. Are there indeed empirically measurable restrictions on the occurrence of certain future markers in IF-clauses (as maintained by Comrie 1982, 1985, among others), and how big is this effect? In addition, is that effect—should it exist—uniform across registers and across varieties, or are there quantifiable differences?[3]
5. Is there any correlation between what I will conceptualize as "sentence length" and occurrence likelihoods of future markers?

Method and Data

Defining the Inventory of Future Markers in English

This study focuses on what Bybee, Perkins, and Pagliuca (1994) have called "primary" future markers, that is, on constructions that consist of an auxiliary verb with or without an infinitive (and contractions thereof). Full (*be*) *going to*, such as in (1), and contracted (*be*) *gonna*, such as in (2), are the two major variants of the semimodal BE GOING TO paradigm, including past tense forms of the future marker BE GOING TO, such as in (3):

(1) Do you think that'*s going to* come to anything? (BNC KB0 69)[4]
(2) I'*m gonna* sit down quietly. (BNC KB0 3038)
(3) I've forgotten what I *was gonna* say. (BNC KB0)

Some authors (e.g., Berglund 1999) would not include past tense forms as in (3) in their analyses because they are not possible with WILL/SHALL. However, were these excluded, the following uses of WILL/SHALL, by the same logic, should be excluded too: (1) “future perfect” forms such as *I will have seen it*, which are highly awkward with BE GOING TO (??*I am going to have seen it*), and (2) tag questions, which—when used with WILL/SHALL—repeat the future marker (as in *he won’t do it, will he?*) but usually reiterate the auxiliary only when used with BE GOING TO (as in *he is going to do that, isn’t he?*; cf. Tottie 2002). For the sake of inclusion and simplicity, this study includes all the aforementioned forms, where applicable. Spatial uses of *be going to* (such as in *I am going to school*) are, of course, excluded from analysis.

Full *will*, as in (4); negated contracted *won’t*, as in (5); cliticized *’ll*, as in (6); and *shall*, as in (7), are the four major realizational variants of the WILL/SHALL paradigm:

(4) People *will* be saying things aren’t they? (BNC KBL 385)
(5) I *won’t* say any more. (BNC KB0 1643)
(6) I say I*’ll* put me feet up before we wash up. (BNC KBB 7766)
(7) We *shall* have to wait and find out until May. (BNC KRT 838)

Shall has come to be somewhat marginalized in present-day spoken English (cf. Kjellmer 1988; Tottie 2002; Trudgill 1984). NOT-contracted forms of *shall* (*shan’t*) are virtually nonexistent in my data and are subsumed under figures for *shall* because of lack of relevance. For the same reason, frequencies of *shall* are not corrected for nonfuture usages of *shall*. Nonfuture usages of *will* (e.g., *This is an ultimatum, if you will*) are excluded from analysis.

For operational reasons, BE GOING TO and WILL/SHALL, as well as their realizational variants, are considered semantically interchangeable for the remainder of this study. This constitutes an abstraction in that I certainly do not mean to argue here that the choice between the two paradigms is always unconditionally optional. Yet, the assumption of general interchangeability is not unmotivated, given how frequently one encounters this claim in the literature. Palmer (1974, 163) has stated that “in most cases, there is no demonstrable difference between *will/shall* and *be going to*”; Danchev et al. (1965, 384), Hall and Hall (1970, 138), and Quirk et al. (1985, 218), just to name a few, are on record in a similar fashion.

Data

The present study will analyze three major computerized corpora of contemporary spoken English: the informal spoken and the formal spoken section of the British National Corpus (BNC), the Santa Barbara Corpus of Spoken American English (CSAE), and the Corpus of Spoken Professional American English (CSPAE). Note that the latter two corpora have not been analyzed with regard to future time reference yet in the literature.

The BNC contains a spoken section of about 10 million words. It consists of spoken English of various kinds, produced by different speakers in various situations. What is important here is that the corpus claims to be representative of contemporary British English. The spoken section of the BNC is subdivided relatively equally into a *demographically sampled* (DS) component, consisting of language in informal encounters recorded by a socially balanced sample of informants, and a *context-governed* (CG) component of formal encounters categorized into four domains. For the remainder of this study, the DS and CG sections of the BNC will be treated as separate corpora, the first of which contains informal British English and the second formal British English. Note that the original version of the BNC, released in 1995, was used in this study.

The CSAE is currently being composed at the University of California at Santa Barbara and contains, in its first installment released in 2000, fourteen conversations with fifty-one speakers. This corpus, then, is a small one (c. 61,000 words), but it is large enough for some of the purposes of this study. Moreover, it is currently the only major corpus of American English conversation accessible to the wider research community. Results obtained from the CSAE, then, may often not prove to be statistically significant (which, of course, says nothing about substantial significance). With regard to its composition, its creators claim that the CSAE can be taken to be representative of contemporary American English; this corpus will be used here to match the informal spoken DS section of the BNC.

The CSPAE, finally, is a corpus of roughly 2 million words consisting primarily of short interchanges by approximately 400 speakers that “are centered on professional activities broadly tied to academics and politics,” as the publisher asserts (http://www.athel.com/corpdes.html). The corpus is made up of press conference transcripts and transcripts from faculty meetings and other committee meetings. As these transcripts are official or semi-official (and have probably not been transcribed by linguists), *gonna* is not transcribed in the corpus, as the form is apparently deemed to be too substandard for official releases. However, because *gonna* has, with sufficient certainty, been transcribed as “going to,” this transcription practice does not pose a grave caveat to the present study, as long as it is kept in mind that this corpus is good only for measuring frequencies of the paradigm BE GOING TO and not of its variants. For this reason, frequencies of variant forms of BE GOING TO are not provided for the CSPAE in what follows, as it is impossible to determine for any given instance of transcribed “be going to” whether it was originally full *be going to* or contracted *gonna*. It should be pointed out, though, that except for *gonna*, no other future marker variant seems to be affected by inadequate transcription: *will*, *won’t*, *’ll*, and *shall* all occur in the corpus, and their frequencies seem to be neither suspiciously high nor suspiciously low when compared to the other corpora. In contrast to the CSAE and the DS section of the BNC, the CSPAE does not contain natural face-to-face conversation but more careful speech that is—

given the settings (faculty and committee meetings, the White House pressroom)—formal rather than informal in nature. For these reasons, the CSPAE is used in this study to match the formal spoken CG section of the BNC, which contains similar material.

This selection of data is an attempt to span two varieties of English (British English and American English) and two spoken registers in each variety (formal spoken English and informal, colloquial spoken English). While especially the two corpora of American English may have minor shortcomings (size as regards the CSAE and *gonna* not being transcribed in the CSPAE), it is important to note that pending the completion and/or publication of a corpus of American English that matches the size and quality of the BNC, the CSAE and CSPAE must be considered our best take on spoken American English at this point.

Research Design

Analysis was conducted as follows. To establish frequencies of future markers overall and in contexts of negation, each corpus was searched automatically by retrieval software (SARA Version 0.930 for the BNC and WordSmith Tools Version 3.0 for the CSAE and CSPAE). Frequencies thus obtained contain occurrences of spatial *going to*, such as in (8), or instances of nonfuture-marking *will*, such as in (9), for which figures must be controlled.

(8) I mean, you're *going to* Africa. (DS KDW 8177)
(9) Mark is making his *will* isn't he? (DS KCN 2126)

These nonfuture-marking forms were accounted for by different methods, depending on the corpus:

- For the (relatively small) CSAE, all future marking and non-future-marking forms were manually disambiguated, and non-future-marking forms were then removed from the counts. It turned out that of thirty-seven *going to* forms occurring in the corpus, eight (c. 21.6 percent) were nonfuture marking; of the seventy-two *will* forms occurring in the corpus, seven (c. 9.7 percent) were nonfuture. Table 1 and the following display adjusted figures.
- For the DS, CG, and CSPAE, two random samples of 300 occurrences each—one containing occurrences of transcribed *will* and the other containing instances of transcribed *be going to*—were drawn from each of these corpora. All future marking and nonfuture-marking forms were then manually disambiguated in these samples, and the resulting percent factors of non-future-marking forms were then extrapolated to the whole corpora.[5] Raw counts, as obtained by the concordancing software, therefore, were statistically adjusted by the following percent factors: (1) nonfuture-marking *will*: CSPAE 2.7 percent, DS 6 percent, and CG 5 percent and (2) nonfuture-marking *be going to*: CSPAE 1.3 percent, DS 13 percent, and CG 4 percent. Table 1 and the following display adjusted figures.

To study the behavior of future markers in specific syntactic environments, the following method was applied: two independent, random future marker samples—one of negated markers and one of nonnegated markers—were drawn from each of the four corpora and entered into a database. The total number of future marker occurrences in the database was 9,193, which constitutes a sample size large enough for generalization. Every future marker in the database was then coded manually according to the syntactic neighborhood in which it was embedded. Criteria for coding included the following: (1) is the marker embedded in a syntactically independent or dependent clause? (2) If it is embedded in a subclause or in the main clause of a subclause, of what type is the subclause (i.e., is it a IF-subclause or another subclause)?[6] Results obtained from manual coding were then entered into the database and statistically analyzed with regard to the research questions. To enable cross-corpus and cross-marker comparisons, findings from samples were standardized by weighting samples based on the overall frequency of the respective future marker form in the corpus from which the sample was drawn (this was necessary because occurrence likelihoods differ between markers and corpora). More details on the sampling method, the coding method (including a simplified coding scheme), and the result of a test of intercoder reliability can be found in Appendix A. To investigate the relationship between sentence length and future marker frequencies, a VisualBasic script was used in word-processing software. More information about the procedure is provided in the respective section.

Results of chi-square tests for statistical significance are generally provided, except for Tables 1 and 2 (which will serve as default distributions against which to test later findings) and in instances where the use of chi-square tests is not appropriate (i.e., if any one expected frequency is zero or if the expected frequency is less than five occurrences in more than 20 percent of the cells). Usually, this requirement is not met by distributions of future marker variant forms in the corpora of American English due to their comparatively small sizes.

Results and Discussion

Overall Frequencies of Future Markers

Overall future marker frequencies—corrected for nonfuture-marking homonyms—are given in Table 1 and visualized in Figure 1. Percentages in parentheses refer to the distribution of future markers within each corpus.

Differences between corpora are statistically highly significant. In what follows, I will therefore, for the most part, not report chi-square values for individual observations. As can be seen from Table 1, *gonna* is most frequent in the CSAE, where it outnumbers full *going to* by a remarkable ratio of roughly 7:1 (*gonna* is simultaneously the single most frequent future marker in the CSAE). Also, *gonna*

TABLE 1
Future Marker Forms in Corpora

	CSAE	CSPAE	DS	CG
going to	29 (5.8)	5,838 (31.1)	4,289 (9.6)	6,251 (16.9)
gonna	205 (41.3)		8,048 (18.1)	3,866 (10.5)
will	65 (13.1)	9,349 (49.9)	6,657 (15.0)	12,184 (33.0)
'll	178 (35.9)	3,137 (16.7)	19,867 (44.6)	11,847 (32.1)
won't	16 (3.2)	385 (2.1)	3,998 (9.0)	1,678 (4.5)
shall	3 (0.6)	35 (0.2)	1,657 (3.7)	1,070 (2.9)
Total	496 (100)	18,744 (100)	44,516 (100)	36,895 (100)

NOTE: Percentages in parentheses. CSAE = Santa Barbara Corpus of Spoken American English; CSPAE = Corpus of Spoken Professional American English; DS = demographically sampled; CG = context governed.

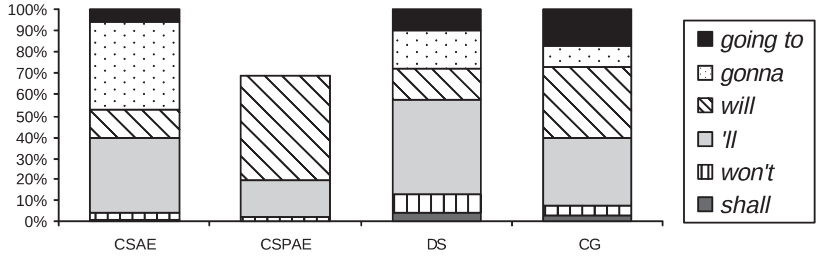


Figure 1: Future Markers in Corpora (Distributions in Percentages).

is more frequent than the full form in the DS corpus, although it is "only" roughly twice as frequent there. In the CG corpus, in contrast, the full form *going to* is more frequent than the contracted form.

Full *will* has a bigger share than its contracted variants in the formal corpora; the reverse holds for the informal corpora. Overall, the British English corpora contain a higher share of cliticized *'ll* than the American corpora. *Won't* is comparatively more frequent in the informal corpora than in the formal corpora and, overall, more frequent in the British English corpora. Compared to the other markers, though, *won't* is rather infrequent. Also note that *shall* is truly marginal, albeit lesser so in British English than in American English. At the same time, however, *shall* seems to be more frequent in informal discourse than in formal discourse in both American English and British English (significantly so only in British English, although at $\chi^2 = 39.4$, $df = 1$, $p < .01$). Table 2 conflates individual figures and gives the shares of the future marker paradigms.

In all four corpora, BE GOING TO forms are outnumbered by WILL/SHALL forms. But while the proportion is roughly 73:24 in the CG corpus, it decreases to almost 50:50 in the CSAE. In general, BE GOING TO is clearly more frequent in the American corpora than in the British corpora, and it less frequent in the formal

TABLE 2
Future Marker Paradigms in Corpora

	CSAE	CSPAE	DS	CG
BE GOING TO	234 (47.2)	5,838 (31.1)	12,337 (27.7)	10,117 (27.4)
WILL/SHALL	262 (52.8)	12,906 (68.9)	32,179 (72.3)	26,779 (72.6)
Total	496 (100)	18,744 (100)	44,516 (100)	36,895 (100)

NOTE: Percentages in parentheses. CSAE = Santa Barbara Corpus of Spoken American English; CSPAE = Corpus of Spoken Professional American English; DS = demographically sampled; CG = context governed.

data than in the informal data but statistically significantly so only in American English ($\chi^2 = 57.5$, $df = 1$, $p < .01$).

We have seen in this section that clearly, overall distributions of future markers are stratified regionally as well as stylistically. *Won't* and *shall* are not particularly frequent in the overall data (but see the next section for a detailed analysis of negated contexts only). While *shall* is rarely used overall (cf. Kjellmer 1998; Tottie 2002), it is more frequent in informal discourse than in formal discourse. This might be a residue of the greater propensity of informal conversational discourse to contain direct questions and first-person subjects (cf. Chafe 1980; Tannen 1982), contexts in which *shall* is often said to be comparatively frequent (cf. Berglund 1999; Biber et al. 1999). All other things being equal, the informal corpora contain lower percentages of the full, noncontracted future marker forms than their respective formal counterparts (cf. Close 1988). BE GOING TO, as a paradigm, is clearly more frequent in informal discourse than in formal discourse (this is consistent with, for instance, Berglund 2000b). Also, the share of BE GOING TO is higher in formal American English than in formal British English and remarkably higher in informal American English than in informal British English. This conforms with Biber et al.'s (1999) and especially Tottie's (2002) results from the *Longman Spoken American Corpus* (a corpus not accessible to the wider research domain) that BE GOING TO is considerably more common in spoken American English than in spoken British English. I found that in the CSAE, BE GOING TO and WILL/SHALL have almost equal shares—which is, to my knowledge, the biggest share of BE GOING TO that has, to date, been measured in a quantitative study of stratified corpus data.

Future Markers in Contexts of Negation

Negation, in this study, will be primarily understood as the function of the word *not* (or a contraction thereof). Frequencies for future markers that are negated by *not*, such as in (10), or by a NOT-contracted auxiliary, such as in (11) or (12), as well as figures for *won't* are presented in Table 3 and visualized in Figure 2.[7]

TABLE 3
NOT-Negated Future Markers in Corpora

	CSAE	CSPAE	DS	CG
not going to	1 (1.9)	725 (51.5)	385 (7.1)	565 (18.3)
not gonna	23 (54.1)		816 (15.1)	390 (12.6)
will not	3 (6.4)	297 (21.1)	104 (1.9)	389 (12.6)
'll not	0 (0)	0 (0)	91 (1.7)	51 (1.7)
won't	16 (37.6)	385 (27.3)	3,998 (74.1)	1,678 (54.3)
shall not	0 (0)	2 (0.1)	5 (0.1)	16 (0.5)
Total	43 (100)	1,409 (100)	5,399 (100)	3,089 (100)
Chi-square	NA	2.819.3, *df* = 4, *p* < .01	15,797.4, *df* = 5, *p* < .01	9,755.4, *df* = 5, *p* < .01

NOTE: Percentages in parentheses. CSAE = Santa Barbara Corpus of Spoken American English; CSPAE = Corpus of Spoken Professional American English; DS = demographically sampled; CG = context governed; NA = not applicable.

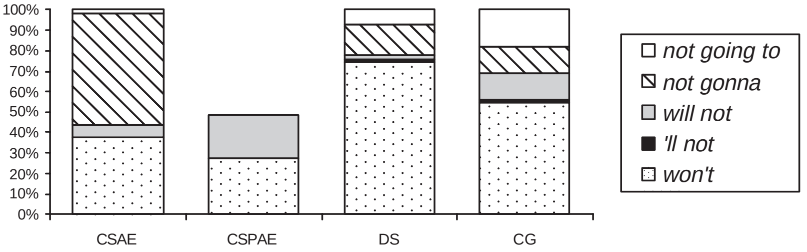


Figure 2: Negated Future Markers in Corpora (Distributions in Percentages).

(10) Those ministers from the South who *will not* be conducting morning worship tomorrow . . . request whoever is conducting worship to use this in the service in the prayers! (DS KBK 481)

(11) Neil Kinnock will tell you what the Conservatives *aren't gonna* do. (DS KCF 241)

(12) Cos the walls *ain't gonna* be done, I'll get back and get a tub next week. . . . (DS KB6 49)

As can be seen from the chi-square values given, for three of the four corpora, distributions of NOT-negated future markers are significantly different from the overall distributions of future markers. Two observations with regard to these figures are particularly noteworthy: first, *won't*, as in (13), is the single most frequent negated marker in the British English data.

(13) You guys *won't* believe what happened to us in the parking lot of the mall the other day . . . some guy came out and he he was, he was trying to sell us Cologne. (CSAE AD)

TABLE 4
NOT-Negated Future Marker Paradigms in Corpora

	CSAE	CSPAE	DS	CG
BE GOING TO	24 (55.8)	725 (51.5)	1,201 (22.2)	955 (30.9)
WILL/SHALL	19 (44.2)	684 (48.5)	4,198 (77.8)	2,134 (69.1)
Total	43 (100)	1,409 (100)	5,399 (100)	3,089 (100)
Chi-square	1.2, *df* = 1, *p* = .28	246.1, *df* = 1, *p* < .01	72.9, *df* = 1, *p* < .01	17.4, *df* = 1, *p* < .01

NOTE: Percentages in parentheses. CSAE = Santa Barbara Corpus of Spoken American English; CSPAE = Corpus of Spoken Professional American English; DS = demographically sampled; CG = context governed.

Won't is not, however, the single most frequent future marker in the American English data—in other words, there is a strong regional differentiation of negated future marker distributions. Second, it is striking that *'ll not*, such as in (14), has a frequency of zero in both corpora of American English.

(14) Mum says . . . that's the main part of the bedroom int it? And you*'ll not* see it next to Granddad. The wallpaper. (DS KBC 1059)

Judging from the data, *'ll not* is a form absent from both corpora of American English and infrequent in British English. This issue, though interesting, cannot be discussed in much detail here. Suffice it to say that previous research has claimed that expressions such as *I'll not do it* are characteristic of the north of England (Trudgill 1984; Kjellmer 1998) and of Scottish English (Aitken, 1984). Indeed, an investigation into the regional distribution of *'ll not* in the BNC suggests that the form is primarily produced by speakers from the north of England and Scotland.

As Table 4 makes clear, BE GOING TO is the preferred paradigm to be negated in the data of American English, while it is WILL/SHALL in the data of British English. Except for the CSAE, the distribution of paradigms in negated slots is statistically significant when compared to the overall distribution.

In sum, though *won't* is rather infrequent overall, *won't* has a considerable though regionally stratified share in negated contexts. *Won't*, which is, after all, a completely irregular, opaque grammatical morpheme, could be a case in point for the often asserted claim (cf. Hofland and Johansson 1982; Hundt 1997) that analogical pressures are stronger in American English than in British English. In this view, it would not be surprising that systemically regular expressions, such as NOT BE GOING TO, are preferred to less regular options, such as *won't*, in American English. It is also striking that apart from *won't*, WILL/SHALL markers appear to be hardly negated at all, according to my data. An extreme case is negated cliticized *'ll*, which is infrequent in British English and not existent in my data of American English. To conclude, contexts of negation clearly have a significant impact on future marker distributions.

Syntactically Independent/Dependent Environments

I will now discuss correlations between future marker frequencies and syntactically independent and dependent environments. To illustrate, (15) is an example of *gonna* occurring in a syntactically dependent environment (in this case, a relative subclause), while (16) is an example of *will* occurring in a syntactically independent environment (in this case, in the main clause of a subclause of time):

(15) You need somebody who*'s gonna* work with him every day and with an individual programme and you just can't offer that in a class. (DS KBG 60)

(16) Do they look nice? Mm, they're alright, they *will* do when they're, when they grow big. (DS KC2 3282)

In what follows, I will not, for reasons of space, differentiate between negated and nonnegated markers (because, on average, only roughly one in ten future markers is negated, the impact of negation can be considered negligible here). Table 5 gives the distributions, contrasting independent with dependent environments.[8]

As can be seen from the last row in Table 5, differences in distribution between dependent slots and independent slots are statistically highly significant in the corpora of British English; in the corpora of American English, there is a trend pointing in the same direction. In a nutshell, then, both variants of BE GOING TO are comparatively more frequent in dependent environments than in independent environments; the opposite is true for most variants of WILL/SHALL. To illustrate differences further, Figure 3 displays the differences between independent and dependent slot types (with numbers, mathematically, equaling *frequencies in dependent slot types* minus *frequencies in independent slot types*). To enhance clarity, *shall* is excluded from Figure 3.

While full *will*—just as both BE GOING TO variants—is slightly but uniformly overrepresented in dependent slot types, the variants *won't* and, in particular, *'ll* are dramatically less frequent in dependent slots than in independent slots. In fact, *'ll* appears to be the marker most strongly affected by the dichotomy dependent-independent. Also note that future marker distributions in the American English corpora generally seem to be less sensitive to whether slots are embedded in syntactically dependent or independent contexts than in the British English corpora.

These results strongly suggest that, all other things being equal, distributions in dependent clause slots have higher percentages of BE GOING TO markers than distributions in main clause slots. This specific finding regarding the paradigms—which is uniform across all corpora and statistically highly significant except in the CSAE—is illustrated in Figure 4.

TABLE 5
Distribution of Future Markers in Independent/Dependent Syntactic Environments

	CSAE		CSPAE		DS		CG	
	Independent	Dependent	Independent	Dependent	Independent	Dependent	Independent	Dependent
going to	17 (5)	10 (9)	408 (27)	277 (37)	192 (8)	94 (17)	333 (14)	231 (28)
gonna	127 (38)	54 (48)		384 (16)		166 (30)	214 (9)	108 (13)
will	40 (12)	21 (19)	741 (49)	389 (52)	360 (15)	88 (16)	738 (31)	314 (38)
'll	134 (40)	27 (24)	317 (21)	67 (9)	1,128 (47)	171 (31)	904 (38)	132 (16)
won't	13 (4)	1 (1)	45 (3)	15 (2)	240 (10)	22 (4)	119 (5)	8 (1)
shall	3 (1)	0 (0)	1 (0)	1 (0)	96 (4)	12 (2)	72 (3)	33 (4)
Total	334 (100)	113 (100)	1,512 (100)	749 (100)	2,400 (100)	553 (100)	2,380 (100)	826 (100)
Chi-square	NA		NA		134.1, *df* = 5, *p* < .01		206.2, *df* = 5, *p* < .01	

NOTE: Percentages in parentheses. CSAE = Santa Barbara Corpus of Spoken American English; CSPAE = Corpus of Spoken Professional American English; DS = demographically sampled; CG = context governed; NA = not applicable.

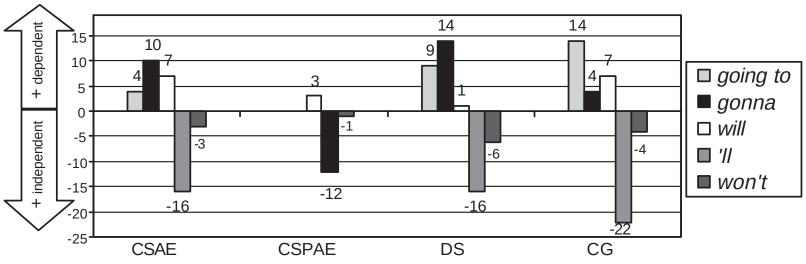
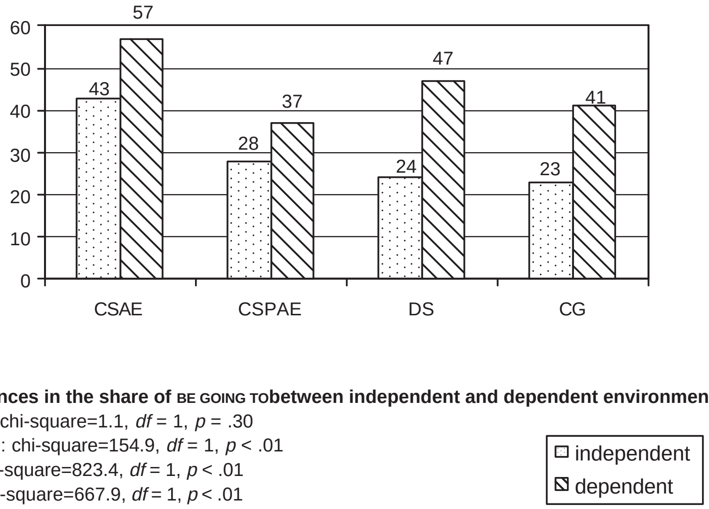


Figure 3: Distributional Differences between Syntactically Independent and Dependent Environments (in Percentage Points).

NOTE: CSAE = Santa Barbara Corpus of Spoken American English; CSPAE = Corpus of Spoken Professional American English; DS = demographically sampled; CG = context governed.



Differences in the share of BE GOING TObetween independent and dependent environments:
CSAE: chi-square=1.1, *df* = 1, *p* = .30
CSPAE: chi-square=154.9, *df* = 1, *p* < .01
DS: chi-square=823.4, *df* = 1, *p* < .01
CG: chi-square=667.9, *df* = 1, *p* < .01

Figure 4: Shares of BE GOING TO in Syntactically Independent and Dependent Environments (in Percentages).

NOTE: CSAE = Santa Barbara Corpus of Spoken American English; CSPAE = Corpus of Spoken Professional American English; DS = demographically sampled; CG = context governed.

That BE GOING TO is more frequent in syntactically dependent contexts than in independent contexts could be due to the fact that BE GOING TO contains more lexical and morphological material than WILL/SHALL, in that BE GOING TO obligatorily involves an auxiliary that inflects for person, number, and tense. Because simple clauses are typically shorter than complex clauses, speakers might—all other things being equal—be more likely to employ the more compact paradigm (i.e., WILL/SHALL) in simple and/or main clauses. In this context, note that in my data, *'ll*—which, as a clitic, is the shortest marker of all—is most frequent in syntactically independent environments. Inversely, the longer, more complex paradigm (i.e., BE GOING TO) might be preferred in more complex clause structures (cf. Rohdenburg's [1996] "complexity principle"). This point will be further elaborated on in the conclusion of this study.

Future Markers in IF-Clauses

In this section, I will contrast future marker distributions in IF-subclauses with distributions of future markers in main clauses of IF-subclauses. Example (17) exemplifies an instance of *gonna* occurring in an IF-subclause:

(17) And if he*'s gonna* walk to Tenby they could be starting when he's in Tenby. (DS KCN 3375)

In (18), *will* occurs in the main clause of an IF-subclause:

(18) Or do you just want to take the pages out? Er it's up to you. I *will* do if you want to. (DS KB9 3778)

Table 6, then, gives the shares of individual markers in IF-subclauses. When differences between these shares and the overall shares, as given in Table 1, exceed 5 percentage points, these differentials (in percentage points) are given in parentheses. In addition, results of a test for statistical significance for any such differential are provided if the differential exceeds 5 percentage points.

Figure 5 illustrates that BE GOING TO is more frequent in IF-subclauses than expected in all four corpora subject to analysis here. In British English specifically, distributions are significantly skewed toward BE GOING TO, with BE GOING TO being used in a striking 89 percent of all IF-subclauses in the CG corpus (formal British English). In informal British English and in informal American English, there is still a clear, though less overwhelming, preference for BE GOING TO in IF-subclauses.

Distributions in main clauses of IF-clauses are given in Table 7. Except for the distribution in the CSAE, distributions in main clauses of IF-subclauses are quite clearly skewed toward WILL/SHALL, with BE GOING TO being underrepresented.

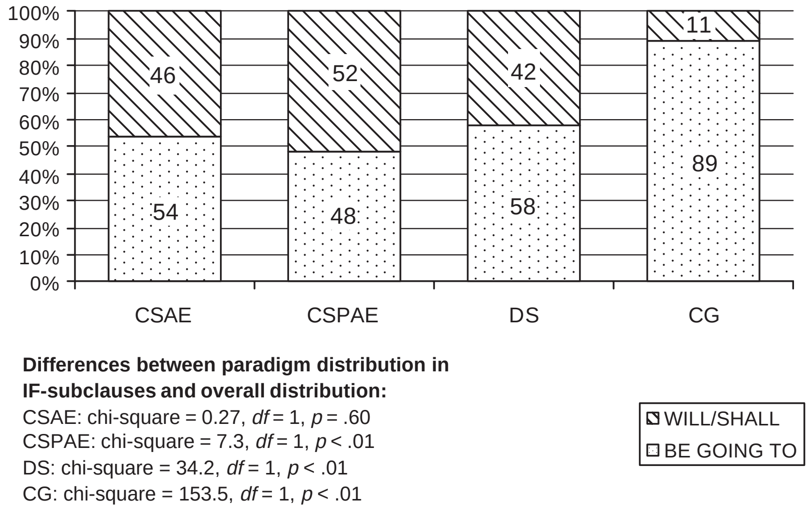
Hence, unlike in IF-subclauses, WILL/SHALL tends to be the preferred syntactic option in main clauses of IF-subclauses, as Figure 6 illustrates.

The analysis in this section has suggested that distributions of future markers are indeed strikingly sensitive to IF-clause environments. In main clauses of IF-subclauses, WILL/SHALL tends to be the clearly preferred paradigm. BE GOING TO, in sharp contrast, is much more frequent in IF-subclause slots than one would expect, knowing this paradigm's overall frequencies in the corpora (and not knowing the literature). Consider, in this context, formal British English, where BE

TABLE 6
Distributions of Future Markers in IF-Subclauses

	CSAE	CSPAE	DS	CG
going to	1 (7)	26 (48; +16.9 pts.)	16 (21; +11.4 pts.**)	42 (52; +35.1 pts.**)
gonna	7 (47; +5.7 pts.)		27 (37; +18.9 pts.**)	31 (38; +27.5 pts.**)
will	5 (33; +19.9 pts.)	27 (49)	10 (13)	7 (9; –24.0 pts.**)
'll	2 (13; –22.9 pts.)	2 (3; –13.7 pts.*)	21 (28; –16.6 pts.)	0 (0; –32.1 pts.**)
shall	0 (0)	0 (0)	0 (0)	0 (0)
won't	0 (0)	0 (0)	1 (1; –8.0 pts.*)	1 (2)
Total	15 (100)	55 (100)	75 (100)	81 (100)
Chi-square	NA	NA	36.3, *df* = 5, *p* < .01	162.0, *df* = 5, *p* < .01

NOTE: Percentages and percentage points in parentheses. CSAE = Santa Barbara Corpus of Spoken American English; CSPAE = Corpus of Spoken Professional American English; DS = demographically sampled; CG = context governed; NA = not applicable.
**p* < .05. ***p* < .01 (obtained by testing individual marker shares against the shares given in Table 1).



Differences between paradigm distribution in IF-subclauses and overall distribution:
CSAE: chi-square = 0.27, *df* = 1, *p* = .60
CSPAE: chi-square = 7.3, *df* = 1, *p* < .01
DS: chi-square = 34.2, *df* = 1, *p* < .01
CG: chi-square = 153.5, *df* = 1, *p* < .01

Figure 5: Future Marker Distributions in IF-Subclauses (in Percentages).
NOTE: CSAE = Santa Barbara Corpus of Spoken American English; CSPAE = Corpus of Spoken Professional American English; DS = demographically sampled; CG = context governed.

GOING TO occurs in 89 percent of all IF-subclauses. My corpus evidence therefore clearly supports Comrie's (1982, 1985) claim that while the use of WILL/ SHALL can be grammatically problematic in conditional protases, periphrastic BE GOING TO is, at least grammatically, always possible. It is an interesting finding that while both American English and British English speakers prefer BE GOING TO over WILL/SHALL in IF-subclauses, this tendency is much more pronounced in British English than in American English. This phenomenon might be due to the fact that prescriptivist traditions are more influential for speakers of British English (particularly in formal situations) than for speakers of American English.

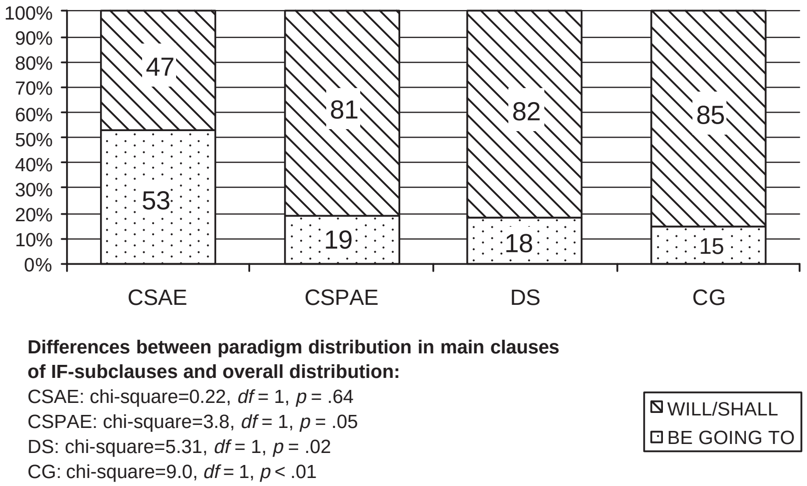
Sentence Length

I have established earlier that frequencies of future markers are sensitive to whether their slots are embedded in syntactically dependent or independent environments. I will now turn to a discussion of another characteristic of the surrounding text, which I will conceptualize here as *sentence length* (in words) of sentences that contain future markers. This variable, of course, is related to the slot feature *plus/minus dependent* in that higher degrees of subordination will usually yield longer sentences than will simple clause structures, all other things being equal; if grammatical complexity contributes to sentence length, then structures of a higher degree of subordination will yield longer sentences. Unlike the variable *plus/minus dependent*, sentence length (though correlated to the former) is a criterion that is not binary and more gradual.

TABLE 7
Distributions in Main Clauses of IF-Subclauses

	CSAE	CSPAE	DS	CG
going to	2 (12; +6.2 pts.)	11 (19; –21.1 pts.)	4 (4; –5.6 pts.)	12 (10; –6.9 pts.)
gonna	7 (41)		16 (14)	6 (5; –5.5 pts.)
will	2 (12)	29 (53)	18 (16)	50 (43; +10.0 pts.)
’ll	4 (24; –11.9 pts.)	12 (22; +5.2 pts.)	67 (59; +14.4 pts.)	42 (36)
shall	0 (0)	0 (0)	1 (1)	1 (1)
won’t	2 (12; +8.8 pts.)	3 (6)	7 (6)	6 (5)
Total	17 (100)	55 (100)	113 (100)	117 (100)
Chi-square	NA	NA	14.2, *df* = 5, *p* = .02	11.9, *df* = 5, *p* = .03

NOTE: Percentages and percentage points in parentheses. CSAE = Santa Barbara Corpus of Spoken American English; CSPAE = Corpus of Spoken Professional American English; DS = demographically sampled; CG = context governed; NA = not applicable.
**p* < .05. ***p* < .01 (obtained by testing individual marker shares against the shares given in Table 1).



Differences between paradigm distribution in main clauses of IF-subclauses and overall distribution:
CSAE: chi-square=0.22, *df* = 1, *p* = .64
CSPAE: chi-square=3.8, *df* = 1, *p* = .05
DS: chi-square=5.31, *df* = 1, *p* = .02
CG: chi-square=9.0, *df* = 1, *p* < .01

Figure 6: Future Marker Distributions in Main Clauses of IF-Subclauses (in Percentages).
NOTE: CSAE = Santa Barbara Corpus of Spoken American English; CSPAE = Corpus of Spoken Professional American English; DS = demographically sampled; CG = context governed.

Note that using sentence length as a criterion has certain drawbacks (which, it should be pointed out, pertain more or less to *all* syntactic analyses of spoken data). Crystal (1980) has shown how indeterminate connectivity, intercalated structures, and ellipsis can blur syntactic boundaries to such an extent that an identification of sentence units is difficult. In some cases, the decision of how to classify phrasal material into “sentences” is at the discretion of the transcribers. Nonetheless, I will use the notion of “sentence” as a unit of measure here for three reasons: (1) In the data used in this study, there is reason to assume that transcription methods were not entirely arbitrary, so that the method can be expected to return a fairly systematic measure. (2) While “sentence” might be a controversial unit in the analysis of spoken language, it has been used in the literature: Chafe and Danielewicz (1987, 103), for instance, point out that often, speakers use intonation to indicate that they have arrived at the end of some coherent structure. Chafe and Danielewicz, in their study, are happy to work with the notion of “sentence” in spoken language, and so am I. (3) In the sections leading up to this one, I have presented a fine-grained analysis on the basis of manually coded, exceedingly reliable data. This section will complement what has been established before by a more quantitative analysis of large amounts of data. That the findings obtained through both methods point in the same direction contributes to the robustness of the results presented in this section.

Technically, the notion of “sentence” was conceptualized as comprising material between two punctuation marks (full stops, question marks, or exclamation marks but *not* commas), thus adopting the concept of “orthographic” sentences (cf. Greenbaum 1980, 26). Admittedly, this procedure necessitates some confidence in the transcribers' ability to adequately translate nonorthographic, intonational devices used in actual speech to indicate coherent sentence and/or meaning units into orthographic punctuation. With regard to the material used in this study, I believe, this confidence is warranted.[9] Thus, a "sentence" in this section may be any of the following: (1) a simple sentence, (2) a complex sentence, (3) a compound sentence, (4) an elliptical sentence, or (5) any combination of the former four. Examples (19) to (21), which are actually part of a coherent conversation and would count as one sentence each in this section, will illustrate.

(19) But anyway, we get these horse hooves, from this one cannery, they they have to go, a long ways to go get em, like back East somewhere, to get these horse hooves.
(20) For the college.
(21) . . . So we have this frozen horse hoof, that we have to start out on, cause you don't want to cripple up a really good horse, and like, my first hoof, that horse would have been, lame, like crazy. (CSAE AB)

Instances of nonverbal material such as *um*, *uh*, *oh*, and *er* were not counted. Likewise, future markers themselves were excluded from the word count because they differ in length (cliticized *'ll* is not a word of its own, while full *be going to* consists of three words). Including them in the count would have skewed results in favor of BE GOING TO. Results thus have to be interpreted to indicate the quantity of the material adjacent to any given future marker slot, excluding the slot itself. To exemplify, while (22) was analyzed as having a length of seven words, (23) counted for three words:

(22) . . . Um, they *were gonna* go out, because they felt called. (CSAE TL)
(23) It *won't* last long. (CSAE AB)

In addition, sentences containing two or more instances of the same future marker form, like in (24), were included in the count only once in order not to skew results.

(24) Er, stop here, we*'ll* we*'ll* cross here look. (DS KB8 512)

In all, the data analyzed in this section comprise around 860,000 words and include the following:

- the entire CSAE (c. 61,000 words), which was analyzed as one text;
- a random sample drawn from the CSPAE (consisting of the texts WH6, MCM597, UNC95, and RC696; c. 500,000 words); and
- a random sample drawn from the DS section of the BNC (consisting of the texts KBW, KDM, and KDW; c. 300,000 words).

TABLE 8
Average Length (in Words) of Sentences Containing Future Marker Slots

	CSAE		CSPAE Sample		BNC Sample (DS Section)	
	n	Average Length	*n*	Average Length	*n*	Average Length
going to	29	20.4	1,151	31.4	341	20.6
gonna	194	18.1			458	20.5
will	59	18.0	1,739	31.6	369	20.2
'll	154	18.2	586	29.5	1,133	18.8
won't	15	13.4	61	13.1	241	17.5
shall	3	20.3	19	24.4	135	14.8
BE GOING TO	223	18.4	1,151	31.4	799	20.5
WILL/SHALL	231	17.8	2,405	30.6	1,878	18.6
ANOVA statistics	$F = 1.1$, $df(n) = 5$, $df(d) = 448$, $p = .36$		$F = 110.9$, $df(n) = 4$, $df(d) = 3{,}551$, $p < .01$		$F = 2.1$, $df(n) = 5$, $df(d) = 2{,}671$, $p = .06$	

NOTE: CSAE = Santa Barbara Corpus of Spoken American English; CSPAE = Corpus of Spoken Professional American English; BNC = British National Corpus; DS = demographically sampled.

Table 8 displays average sentence lengths, consolidated according to future marker variant. In addition, average sentence lengths for both paradigms, BE GOING TO and WILL/SHALL, are given.

It is clear from Table 8 that there is a robust, uniform trend for sentences containing BE GOING TO to be longer than sentences containing WILL/SHALL. In the CSAE, the difference is, on average, 0.6 words (equaling a relative difference of c. 3 percent); it is 0.8 words (c. 3 percent) in the CSPAE sample and 1.9 words (c. 9 percent) in the BNC-DS sample. While these differences in length may not, prima facie, seem tremendous and reach statistical significance only in CSPAE (with p = .06, figures obtained from the DS fail to reach statistical significance by a very slight margin only),[10] it should be pointed out that the phenomenon seems to be quite significant substantially: not only does it occur in data from three different corpora (cf. Table 8), but it is also verifiable in all individual texts except one (DS KDW) that make up the CSPAE and BNC-DS text samples. The robustness of the phenomenon in the data thus strongly suggests that we must be dealing with a systematic issue here. It should also be remembered that, as has been said before, future markers themselves were excluded from the count. Because, however, instances of BE GOING TO are typically longer than ones of WILL/SHALL, sentences containing BE GOING TO will actually be even longer than the numbers in Table 8 indicate.

Let me summarize the importance of these findings. I have argued before that slots in syntactically dependent environments increase the chance that speakers will use BE GOING TO. The results presented in this section bear clear evidence that there is, in addition, a measurable correlation between sentence length and the future marker employed. This means that the longer and, by inference, the more syntactically complex any given environment is, the more likely it is for BE GOING TO to be used instead of WILL/SHALL.

Summary and Conclusion

In this study, I have focused on a kind of variation in the overt expression of futurity in English that has, I believe, received less empirical attention so far than it deserves: sensitivity to syntactic context. Assuming that the choice between BE GOING TO and WILL/SHALL is, at base, an optional one between two interchangeable patterns, I have added to current theory on the expression of futurity in English by presenting quantitative evidence that, in addition to pervasive stylistic and regional patterns of variation, there is also significant syntactic stratification involved.

First, with regard to overall frequencies of future markers, the patterns of stylistic and regional stratification I detected dovetail nicely with what has been established in previous research. Second, my analysis suggests that contexts of negation have a significant impact on future marker distributions. *Won't* is—unsurprisingly—far from infrequent in negated contexts. Except for opaque *won't*, however, I have shown that WILL/SHALL is rarely explicitly negated. Third, I have demonstrated that BE GOING TO is much more frequent in syntactically dependent contexts than it is in independent contexts, while the reverse holds for WILL/SHALL. Fourth, I have presented evidence that, much as hypothesized by extant scholarship (e.g., Comrie 1982, 1985; Declerck 1991), WILL/SHALL is overrepresented in main clauses of IF-subclauses, while BE GOING TO is overrepresented in IF-subclauses. I also found that this effect is more marked in British English than in American English. Finally, I measured average sentence length in words of sentences with a future marker slot and showed that sentences that contain a slot for BE GOING TO are longer than sentences that contain a slot for WILL/SHALL. I took this finding—which I could verify independently in data from three different corpora—to reinforce my earlier observation that BE GOING TO is preferred by speakers in syntactically dependent environments.

In conclusion, this study would seem to suggest that the longer, the "more subordinated," and the more "syntactically complex" any given syntactic environment is, the more speakers tend to use BE GOING TO instead of WILL/SHALL. Note, now, that we might be dealing here with an issue of *economy* and *expressivity.* Hopper and Traugott (1993, 65) have argued that BE GOING TO "is more substantive (phonologically longer) and therefore more accessible to hearers than, e.g., *'ll* or even *will*."

In a similar vein, albeit tentatively, I would like to propose that BE GOING TO could be more frequent in grammatically dependent and syntagmatically more complex environments because of an issue with cognitive economy and, relatedly, online processing constraints. The point here is that by virtue of BE GOING TO being the more expressive, phonologically longer, and thus more marked syntactic option, there are two incentives for speakers to incur the costs of having to be more explicit: (1) by using the longer paradigm, BE GOING TO, speakers can stall for planning time. Planning time is a particularly scarce resource in syntactically complex environments, the hierarchical processing of which is more demanding in terms of processing resources on the speakers' side. (2) Because BE GOING TO typically contains more material than WILL/SHALL, it provides a sort of redundancy that will ease online processing for hearers by making the predication more accessible.

The observation, for instance, that BE GOING TO is the preferred paradigm in contexts of overt negation (with the special case of comparatively frequent *won't* in British English) ties in nicely with the above hypothesis: negation, by adding morphological material and reversing truth conditions, makes any future predication more complex to process, which is why payoffs (1) and (2) would apply here too. Somewhat paradoxically, therefore, BE GOING TO might be the more resource-optimizing and more economic syntactic option in syntactically complex environments precisely because it is phonologically and morphologically richer than WILL/SHALL.

APPENDIX A
Sampling Method for Manual Coding of Syntactic Context

For each of the six relevant future marker forms (*going to*, *gonna*, *will*, *'ll*, *won't*, *shall*), the typical procedure was to draw two randomly stratified samples of 400 instances from each of the four corpora, one consisting exclusively of nonnegated forms and the other exclusively of NOT-negated forms of the same marker. Exceptions: (1) if a corpus contained less than 400 instances of the relevant form, all forms the corpus contained were studied; (2) for *won't*, only one sample of 400 forms (or less, if the corpus as a whole contained less forms) has been drawn from each corpus, as *won't* is a negated form already; and (3) *shall*, because of lack of practical relevance, was studied using samples of only 100 (or less, if the corpus as a whole contained less forms) randomly drawn forms. In all, 39 samples totaling 9,193 primary future marker forms have been analyzed and coded with regard to their syntactic neighborhoods. To enable cross-corpus and cross-marker comparisons, figures from individual samples were standardized by weighting intrasample distributions by the overall frequency of the respective future marker form in the respective corpus. This was necessary because future markers have different overall likelihoods to occur in any given corpus. Table 9 gives an overview over sample sizes.

TABLE 9
Sample Sizes

	CSAE	CSPAE	DS	CG	Total
going to	28[a]	400	400	400	1,228
gonna	191[a]		400	400	986
will	62[a]	400	400	400	1,262
'll	178[a]	400	400	400	1,378
won't	16[a]	385[a]	400	400	1,201
shall	3[a]	35[a]	100	100	238
not going to	1[a]	400	378[a]	400	1,179
not gonna	19[a]		400	318[a]	737
will not	3[a]	305[a]	111[a]	400	819
'll not	0[a]	0[a]	91[a]	51[a]	142
shall not	0[a]	2[a]	5[a]	16[a]	23
Total	496	2,327	3,085	3,285	9,193

NOTE: CSAE = Santa Barbara Corpus of Spoken American English; CSPAE = Corpus of Spoken Professional American English; DS = demographically sampled; CG = context governed.

a. Sample contains all instances of the relevant form in the corpus.

Coding Scheme

Each occurrence of a future marker was coded manually according to a specific procedure set forth in a coding scheme, a simplified version of which reads as follows: (1) is the future marker integrated in a clause structure that can be characterized as a *simple* (*main*) *clause*, a *coordinated clause*, a *complex clause structure with only one finite verb*, or a *tag question*? (2) If not, is the future marker embedded in the *main clause* or the *dependent clause* of a complex clause? (3) If the future marker is embedded in a complex clause structure, by which of the following six subclause types can the subclause in the complex clause structure be characterized: (a) interrogative clauses or nominal relative clauses, (b) complement clauses, (c) restrictive or nonrestrictive relative clauses, (d) IF-clauses, (e) time clauses, or (f) cause clauses?

Intercoder Reliability

Determining intercoder reliability of the manual coding procedure served five primary goals: (1) to assess robustness of findings that derive from codings, (2) to bound error levels and to facilitate interpretation of study results within the research domain, (3) to enhance confidence in results, (4) to clarify generalizability to other samples, and (5) to improve likelihood of accurate replication of the coding system (on which findings are dependent). To assess intercoder reliability, the procedure laid out in Orwin (1994) was followed and Cohen's kappa (k) was computed, measuring the proportion of the best possible improvement over chance. Intercoder reliability between the researcher and a second trained coder proved to be very satisfactory. The stratified random sample used to assess intercoder reliability consisted of six sets of 50 future marker variants each (i.e., 300 future markers in all, totaling c. 3.3 percent of the data analyzed in this study), which was drawn from the DS corpus. The sample was first coded independently by the researcher and the second coder; results were then compared, and Cohen's kappa was computed. The second coder was a native speaker of English with a diploma in linguistics who was instructed to follow the procedures laid down in the coding scheme. The comparison of the two independent codings yielded a simple agreement rate of almost 90 percent and a Cohen's kappa (*k*) value of .79 (see Orwin 1994 on how exactly to interpret these values). Typically, any $k \geq .75$ is interpreted to indicate excellent reliability.

APPENDIX B

Text identifiers for the texts that make up the CSAE:

Actual Blacksmithing	(AB)	Tell the Jury That	(TT)
Lambada	(LB)	Zero Equals Zero	(ZZ)
Conceptual Pesticides	(CP)	Bank Products	(BP)
Raging Bureaucracy	(RB)	Letter of Concerns	(LC)
A Book about Death	(BD)	This Retirement Bit	(TR)
Cuz	(CZ)	American Democracy Is Dying	(AD)
A Tree's Life	(TL)	Appease the Monster	(AM)

Text identifiers for the texts that make up the CSPAE:

MathCommitteeMeeting5/97	(MCM597)	MathCommitteeMeeting6/97	(MCM697)
MathCommitteeMeeting7/97	(MCM797)	MathCommitteeMeeting8/97	(MCM897)
ReadingCommittee6/96	(RC696)	ReadingCommittee6/97Part2	(RC697)
ReadingCommittee/97	(RC797)	FacultyMeetingOctober101997	(UNC97)
WhiteHousepressbriefing1	(WH1)	WhiteHousepressbriefing2	(WH2)
WhiteHousepressbriefing3	(WH3)	WhiteHousepressbriefing4	(WH4)
WhiteHousepressbriefing5	(WH5)	WhiteHousepressbriefing6	(WH6)
UniversityofNorthCarolinaFaculty CouncilMeetings:1995	(UNC95)		
UniversityofNorthCarolinaFaculty CouncilMeetings:1996	(UNC96)		

Notes

1. In this study, I will distinguish between realizational variants such as *gonna* or *'ll* and paradigms such as BE GOING TO and WILL/SHALL.

2. I will treat the terms *subclause*, *subordinate clause*, and *dependent clause* as synonymous in this study, referring to nonmain clauses that are morphologically marked so that they cannot stand by themselves.

3. Ideally, one would also have wished to examine time clauses (such as *be nice when you'll be able to go*, British National Corpus [BNC] KC2 4079). Unlike IF-clauses, however, time clauses with overt future marking are exceedingly rare in

my data, so that subject to the limits of the data source, no reliable analysis of this environment can be made.

4. In this study, quotes from the BNC will be identified by the respective text's identifier plus line; quotes from the other corpora will be identified by the text identifiers as defined in Appendix B.

5. This approach used to eliminate nonfuture tokens is one of the reasons why figures for the demographically sampled (DS) and context-governed (CG) corpora in this study differ slightly from those given by, for instance, Berglund (1999), who used another approach.

6. IF-subclauses were taken to include all subclauses that are subordinated by *if* or by related subordinators such as *unless*, *provided*, and *as if*. Although this procedure adopts standard practice in empirical studies involving manual coding (consider, e.g., Beaman 1984), some of the subclauses that I conceptualize as IF-subclauses here would not qualify as conditional protases in Comrie's (1982, 1985) or Declerck's (1991) sense, but rather as "*interrogative yes/no questions*" (Quirk et al. 1985, 737)—for example, subclauses such as *I wonder if he'll sell* (DS KC1 1925). Because only a negligible minority of IF-subclauses (approximately 6 percent) in my data are not conditional protases, however, I will not distinguish between conditional protases and interrogative yes/no questions in what follows.

7. In Table 3 and those that follow, column chi-square statistics were obtained by testing column distributions against overall distributions of future markers in the respective corpus. This means that in Table 3, columns were tested against columns in Table 1.

8. In Table 5, column chi-square statistics were obtained by testing within-corpus distributions in dependent slots against within-corpus distributions in independent slots.

9. It is certainly warranted for the Santa Barbara Corpus of Spoken American English (CSAE), which has been transcribed in an exceedingly accurate and precise fashion for conversational analysis purposes. That results obtained from the other corpora point in the same direction enhances confidence in the way the notion of "sentence" is conceptualized here.

10. The last row in Table 6 was obtained by analyses of variance (ANOVAs) on sentence length (in words) between future marker variants.

References

Primary Sources

British National Corpus (BNC I). Distributed by Oxford University Computing Services, http://www.natcorp.ox.ac.uk/.

Corpus of Spoken Professional American English. Distributed by athelstan, http://www.athel.com/cspa.html.

Santa Barbara Corpus of Spoken American English, Part I. Distributed by LDC, http://www.ldc.upenn.edu/Projects/SBCSAE/.

Secondary Sources

Aitken, Adam J. 1984. Scottish Accents and Dialects. In *Language in the British Isles*, edited by Peter Trudgill, 94-114. Cambridge, UK: Cambridge University Press.

Beaman, Karen. 1984. Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse. In *Coherence in Spoken and Written Discourse*, edited by Deborah Tannen, 45-80. Norwood, NJ: Ablex.

Berglund, Ylva. 1997. Future in Present-Day English: Corpus-Based Evidence on the Rivalry of Expressions. *ICAME Journal* 21:7-20.

———. 1999. Utilising Present-Day English Corpora: A Case Study Concerning Expressions of Future. *ICAME Journal* 24:25-63.

———. 2000a. *Gonna* and *Going to* in the Spoken Component of the British National Corpus. In *Corpus Linguistics and Linguistic Theory*, edited by Christian Mair and Marianne Hundt, 35-49. Amsterdam: Rodopi.

———. 2000b. "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 *PALC'99: Practical Applications in Language Corpora*, edited by Barbara Lewandowska-Tomaszcyk and Patrick James Melia, 161-92. Frankfurt am Main, Germany: Peter Lang.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow, UK: Longman.

Binnick, Robert I. 1971. *Will* and *be going to*. In *Papers from the Seventh Regional Meeting of the Chicago Linguistics Society*, 40-53. Chicago: Chicago Linguistic Society.

Bybee, Joan L., Revere D. Perkins, and William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press.

Chafe, Wallace L. 1980. The Deployment of Consciousness in the Production of a Narrative. In *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*, edited by Wallace Chafe, 9-50. Norwood, NJ: Ablex.

Chafe, Wallace L., and Jane Danielewicz. 1987. Properties of Written and Spoken Language. In *Comprehending Oral and Written Language*, edited by Rosalind Horowitz and S. Jay Samuels, 83-113. New York: Academic Press.

Close, R. A. 1988. The Future in English. In *Kernprobleme der englischen Grammatik: sprachliche Fakten und ihre Vermittlung*, edited by Wolf-Dietrich Bald, 51-66. München, Germany: Langenscheidt-Longman.

Comrie, Bernard. 1982. Future Time Reference in the Conditional Protasis. *Australian Journal of Linguistics* 2:143-52.

———. 1985. *Tense*. Cambridge, UK: Cambridge University Press.

Crystal, David. 1980. Neglected Grammatical Factors in Conversational English. In *Studies in English Linguistics for Randolph Quirk*, edited by Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik, 153-66. London: Longman.

Danchev, A., A. Pavlova, M. Nalchadjan, and O. Zlatareva. 1965. The Construction *going to* + inf. in Modern English. *Zeitschrift für Anglistik und Amerikanistik* 13:375-86.

Declerck, Renaat. 1991. *Tense in English: Its Structure and Use in Discourse*. London: Routledge Kegan Paul.

Greenbaum, Sidney. 1980. The Treatment of Clause and Sentence in *A Grammar of Contemporary English*. In *Studies in English Linguistics for Randolph Quirk*, edited by Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik, 17-29. London: Longman.

Haegeman, Liliane. 1983. *Be going to*, *gaan*, and *aller*: Some Observations on the Expression of Future Time. *International Review of Applied Linguistics in Language Teaching* 21:155-57.

———. 1989. *Be going to* and *will*: A Pragmatic Account. *Journal of Linguistics* 25:291-317.

Hall, R. M. R., and Beatrice L. Hall. 1970. A Note on *will* vs. *going to*. *Linguistic Inquiry* 1:138-39.

Hofland, Knut, and Stig Johansson. 1982. *Word Frequencies in British and American English*. London: Longman.

Hopper, Paul J., and Elizabeth C. Traugott. 1993. *Grammaticalization*. Cambridge, UK: Cambridge University Press.

Hundt, Marianne. 1997. Has British English Been Catching Up with American English over the Past Thirty Years? In *Corpus-Based Studies in English: Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora*, edited by Magnus Ljung, 135-49. Amsterdam: Rodopi.

Kjellmer, Goran. 1998. On Contraction in Modern English. *Studia Neophilologica* 69:155-86.

Kortmann, Bernd. 1991. The Triad “Tense-Aspect-Aktionsart”: Problems and Possible Solutions. *Belgian Journal of Linguistics* 6:9-30.

Kytö, Merja. 1990. *Shall* or *will*? Choice of the Variant Form in Early Modern English, British and American. In *Historical Linguistics 1987: Papers from the 8th International Conference on Historical Linguistics*, edited by Henning Andersen and Konrad Koerner, 275-88. Amsterdam: John Benjamins.

Mair, Christian. 1997a. 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*, edited by Raymond Kickey and Stanslaw Puppel, 1537-43. Berlin: Mouton de Gruyter.

———. 1997b. Parallel Corpora: A Real-Time Approach to the Study of Language Change in Progress. In *Corpus-Based Studies in English: Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora*, edited by Magnus Ljung, 195-209. Amsterdam: Rodopi.

———. Forthcoming. *Corpus Linguistics and Grammaticalisation Theory: Beyond Statistics and Frequency?*

Martin, Willy, and Jan Weltens. 1973. A Frequency-Note on the Expression of Futurity in English. *Zeitschrift für Anglistik und Amerikanistik* 21:289-98.

Nicolle, Steve. 1997. A Relevance-Theoretic Account of *be going to*. *Linguistics* 33:355-77.

Orwin, Robert G. 1994. Evaluating Coding Decisions. In *The Handbook of Research Synthesis*, edited by Harris Cooper and Larry V. Hedges, 139-62. New York: Russell Sage Foundation.

Palmer, Frank R. 1974. *The English Verb.* London: Longman.

Quirk, Randolph, Sidney Greenbaum, Geoffry Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Rohdenburg, Günther. 1996. Cognitive Complexity and Increased Grammatical Explicitness in English. *Cognitive Linguistics* 7:149-82.

Tannen, Deborah. 1982. Oral and Literate Strategies in Spoken and Written Narratives. *Language* 58:1-21.

Tottie, Gunnel. 2002. Non-Categorical Differences between American and British English: Some Corpus Evidence. In *Studies in Mid-Atlantic English*, HS Institutionens Skriftserie 7, edited by Marko Modiano, 37-58. Gävle: University of Gävle Press.

Trudgill, Peter. 1984. Standard English in England. In *Language in the British Isles*, edited by Peter Trudgill, 32-44. Cambridge, UK: Cambridge University Press.

Wekker, Herman. 1976. *The Expression of Future Time in Contemporary British English: An Investigation into the Syntax and Semantics of Five Verbal Constructions Expressing Future Time.* Amsterdam: North-Holland.

Benedikt Szmrecsanyi is a member of the staff in the English Department at the University of Freiburg. He is currently involved in the composition of the Freiburg English Dialect Corpus*, which is part of the research project, English Dialect Syntax from a Typological Perspective, funded by the German Research Foundation. His main research interest lies with contextual determinants of grammatical variation.*