Choices over time: methodological issues in investigating current change[1]

Bas Aarts,[a] Joanne Close[b] and Sean Wallis[a]
[a]University College London
and [b]University of Leeds

1 Introduction

The fact that English is changing is immediately apparent to a modern reader of, say, 18th or 19th century literature, or indeed to a teenager speaking to an elderly relative. However, as Mair (2006) points out, anecdotal evidence for linguistic change is unreliable. The systematic study of language change requires large, evenly balanced, and reliably annotated corpora with texts sampled over a period of time. These considerations are accepted by many linguists working on current changes in English. However, with regard to methodology we observe that within the field of diachronic corpus linguistics there are still a number of issues that generate a certain amount of discussion and debate.

One of these concerns the issue of variability. Bauer (1994: 19) highlights the importance of this concept in studies of language change when he states that "change is impossible without some variation". Variation within a set of linguistic choices, including the idea that there may be 'competition' between these variants, is fundamental to studies in current change. In this paper we will argue that an important methodological task for corpus linguists studying language change is to focus on linguistic variation *where there is a choice*. Many factors are likely to influence the use of particular words, phrases or constructions. If we wish to study and explain variation found in a corpus as being the result of factors affecting variation over time, then we need to eliminate as many potential alternative sources of variation as possible. This, we contend, calls for a restricted definition of the variants involved in a perceived change, and a consideration of any 'knock-out' contexts, i.e. contexts where variation may be impossible, or constrained in a different manner to the general case.

We use the *Diachronic Corpus of Present-day Spoken English* (DCPSE) as a database. This corpus is unique in two important respects: it exclusively contains spoken English and is fully parsed, and as such is suitable for studying current change in English from the late 1950s to the early 1990s. It complements other resources, including major historical corpora of writing, notably *A Representative Corpus of Historical English Registers* (ARCHER)[2] which contains written texts sampled from the late 17th to the late 20th century, as well as corpora of earlier speech derived from written sources such as *A Corpus of English Dialogues* (CED; Kytö and Culpeper 2006) and the *Old Bailey Corpus* (Huber 2007).[3] In the next section we briefly present the functionality of DCPSE.

2 The Diachronic Corpus of Present-Day Spoken English

The *Diachronic Corpus of Present-Day Spoken English* (DCPSE) is a diachronic corpus with a difference: it spans a time period of approximately thirty years and is composed of material from spoken English. DCPSE is composed of speech samples collected between the late 1950s and the early 1990s, and it allows us to monitor grammatical changes during this period. In this paper we will present data on the alternation between *shall* and *will* and the increasing use of the progressive construction, with a focus on the methodological issues raised by these studies. Before showing how this can be done with DCPSE we will discuss a few general features of the corpus.

DCPSE was released by the Survey of English Usage (SEU) in 2006. It contains 464,074 words of orthographic (word-for-word) transcriptions of English speech taken from the *London-Lund*

[1] Early versions of this paper were presented at the symposium *Current Change in the English Verb Phrase* celebrating the fiftieth anniversary of the Survey of English Usage (July 2009), and at the workshop *Corpus-based advances in historical linguistics* organised by the Philological Society, University of York (November 2009). We are grateful to a number of colleagues for comments: Geoffrey Leech, Manfred Krug, Jill Bowie and an anonymous reader. We acknowledge with gratitude the support of the Arts and Humanities Research Council under grant AH/E006299/1.
[2] See: www.llc.manchester.ac.uk/research/projects/archer.
[3] See: www.uni-giessen.de/oldbaileycorpus.

Corpus (LLC),[4] and 321.362 words of spoken data from the *British Component of the International Corpus of English* (ICE-GB, Nelson,Wallis and Aarts 2002).[5] These are sampled in matching text categories, so there is approximately the same quantity of face-to-face conversation (for example) in both portions ('subcorpora'). Two caveats are in order. The LLC subcorpus is distributed over a longer period of time (20 years) than the ICE-GB subcorpus (three years), and texts are not evenly distributed by year.

DCPSE includes mostly spontaneous spoken English, such as face-to-face conversations, telephone conversations, various types of discussions and debates, legal cross-examinations, business transactions, speeches and interviews. As it is generally assumed that changes in English propagate themselves in the first instance through spontaneous discourse, we would argue that DCPSE is ideal for the study of current change. Whereas written corpora contain text genres which allow for editorial correction, DCPSE consists entirely of orthographically transcribed utterances. Immediate self-correction is explicitly marked, so that repetitions and word partials can be excluded from searches. The small proportion of scripted speech that is included is transcribed, rather than the script reproduced.

The spoken transcription is divided into putative 'sentence' utterances, termed 'text units'. Every text unit is then given a full grammatical analysis in the form of a phrase structure tree using a grammar based on Quirk *et al.* (1985) and exemplified in Figure 1. DCPSE contains over 87,000 such fully analysed text units. These trees were produced by automatic and manual parsing methods and were then extensively cross-checked. Parsing natural language is notoriously difficult, and naturally occurring spoken English especially so. There was a substantial manual editing effort, which raises the issue of consistency (Wallis 2003). To deal with this we employed extensive cross-checking in the construction of DCPSE, applying our experience with the annotation of the ICE-GB corpus to the LLC subcorpus. The result is a corpus of spoken English which allows a high degree of confidence in the reliability and completeness of the grammatical analysis.

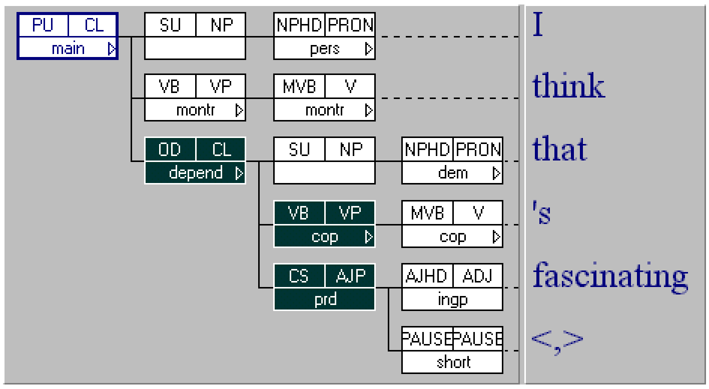

Figure 1: An example tree diagram, *I think that's fascinating* [DI-A02 #28].[6]

The question arises of how to search this forest of over 87,000 trees. Our ICECUP software (*International Corpus of English Corpus Utility Program*; Nelson, Wallis and Aarts 2002) is designed as a platform for exploring the corpus and obtaining results. Linguists can search for lexical strings,

[4] The LLC is the spoken part of the *Survey of English Usage Corpus*, founded by Randolph Quirk in 1959. It contains 510,576 words of 1960s spoken English, is prosodically annotated, and has been used — and continues to be used — by many scholars for their research.

[5] ICE-GB is composed of both spoken and written material from the 1990s. It contains textual markup, and is fully grammatically annotated. All the sentences/utterances in the corpus are assigned a tree structure.

[6] In this tree diagram each lexical item, phrase and clause is associated with a node which contains function information (top left), form information (top right), as well as features (bottom portion). Trees can be drawn in a number of orientations. Here we use a left-to-right visualization for space-efficiency reasons. PU=parse unit, CL=clause, *main*=main, SU=subject, NP=noun phrase, NPHD=NP head, PRON=pronoun, *pers*=personal, VB=verbal, VP=verb phrase, MVB= main verb, V=verb, *montr*=monotransitive, OD=direct object, *depend*=dependent, *dem*=demonstrative, *cop*=copular, CS=Subject Complement, AJP=adjective phrase, *prd*=predicative, AJHD=adjective phrase head, *ingp*=*–ing* participle.

wild cards, etc., and – importantly in *grammatical* studies of current change – tree patterns. ICECUP contains a powerful query system, termed *Fuzzy Tree Fragments* (FTFs). FTFs are 'sketches' of grammatical constructions that can be applied to the corpus to obtain an exhaustive set of matching cases. Figure 2 shows an example of an FTF which matches all instances of a VP followed by a subject complement (CS).[7] This FTF matches the three nodes highlighted in Figure 1 above.

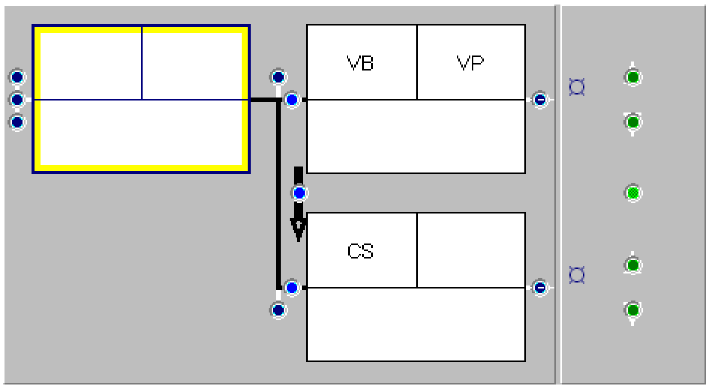

Figure 2: An FTF created with ICECUP, matching the highlighted nodes in Figure 1.

Respecting the fact that linguists disagree about grammar, ICECUP allows users to experiment with the best way of retrieving the grammatical phenomena they are interested in, using the Quirk-style representation in the corpus. The interface is designed to let linguists construct FTFs, apply them to the corpus, identify how they match cases in the corpus, and refine their queries. One can also select part of a tree structure and construct an FTF query from that fragment in order to find how a particular lexical string is analysed, and then seek all similar analyses.

ICECUP offers a range of search tools based around this idea of an abstract 'FTF' query, including a lexicon and 'grammaticon'. DCPSE is an unparalleled resource for linguists interested in short-term changes in spoken English, and in this paper we will demonstrate its value in studies of current change using the examples of the progressive and the *shall* vs. *will* alternation.[8]

3 Focusing on true alternation: the progressive

For decades, research in the field of sociolinguistics has highlighted the importance of the linguistic variant (see Labov 1969). This impetus has percolated into historical studies of language, but is often overlooked in corpus linguistics. Many studies on current change that have been carried out using corpora have collected frequencies for lexical items or grammatical constructions, but often without considering these frequencies alongside the variants of these patterns as part of a 'bigger picture'. In the next three sections we look at a number of methodologies for exploring change. First we look at an approach which measures change in the progressive construction using normalised frequency counts. In section 3.2 we then look at a measure which investigates frequency changes as a percentage of the total number of VPs. Section 3.3 considers changes within a set of variants.

3.1 Changes in frequency per million words

Leech (2003) and Smith (2003) both investigate changes in the modal system of English. They carry out a series of independent log-likelihood 'goodness of fit' tests for the item,[9] in this case a modal auxiliary, against the number of words in the corpus, using a method owing to Rayson (2003). This tests whether a perceived difference in a distribution *d* is too large to be explained by accident.

[7] While the grammar that underlies the ICE-GB parsing (Quirk *et al.* 1985) conceives of Verb Phrases as only containing verbs (see Figure 2), in this paper the focus will be on the 'extended VP', i.e. a verb + dependents, as we discuss later.

[8] For more information see www.ucl.ac.uk/english-usage, Aarts *et al.* (1998), and Nelson *et al.* (2002).

[9] Log likelihood (G^2) is best thought of as a different χ^2 test. It employs a different formula but obtains a similar result. See Appendix 1, http://ucrel.lancs.ac.uk/llwizard.html and www.ucl.ac.uk/english-usage/statspapers/2x2chisq.xls.

First we will apply Rayson's method to progressive VPs, which can be easily identified in DCPSE (cf. Aarts, Close and Wallis 2010).[10] The method compares the distribution in Column A with that of Column B in Table 1.

	A: item VP(prog)	B: words	C: rate per million words	D: increase $d^{\%}$ (LLC = 100%)
LLC (1960s)	2,973	464,063	6,406 (0.64%)	+22.13% ±5.48%
ICE-GB (1990s)	3,294	420,986	7,824 (0.78%)	
TOTAL	6,267	885,049	7,081 (0.71%)	

Table 1: Change over time of 'VP(prog)' as a proportion of the number of words. In this table and the tables below, *d* is cited with a 95% confidence interval indicated by the '±' value.

We compare Column A with B using the goodness of fit log-likelihood test. This attempts to see if the ratio between LLC and ICE-GB frequency counts in Column A is 'similar enough' (as defined by the test) to the ratio between the same counts in Column B. In this instance the results *are* significant at an error level of $p<0.05$. The observed increase in Column C (from 6,406 to 7,824) is likely to represent a *real* (non-zero) increase in the population of comparably sampled English utterances.

We can also measure the percentage difference $d^{\%}$ between the rate for ICE-GB and that found in the LLC subcorpus (column D). We apply the following formula:

(E1) percentage swing $d^{\%} = \dfrac{p_2 - p_1}{p_1}$,

where p_1 represents the probability of selecting a given item (in this example the main verb in a progressive context), at random from the first subcorpus (LLC), and p_2 the same probability in the second subcorpus (ICE-GB). Note that we could substitute any normalised frequency rate – per word, per thousand words, or per million words – for 'probability' here. We can also compute a Gaussian (Normal) confidence interval (Wallis 2010) for $d^{\%}$ at a given error level. This obtains $d^{\%}$ = +22.13 ±5.48 percent at the 0.05 level, or, to put it another way, there is a 19 out of 20 chance that the observed increase $d^{\%}$ is between 16.65 percent and 27.61 percent.

3.2 Changes in frequency as a percentage of the total number of VPs

In a POS-tagged corpus, normalising frequency counts by reporting frequencies per million words is a perfectly reasonable procedure, and obtaining word counts for subcorpora is a simple operation. However, not all words are equally substitutable with the object of study (our Column A). Language is not, to misquote Elbert Hubbard, "just one damn word after another" and corpora are not a random sample of words (Wallis 2010).

In addition, speakers and text genres may vary in how 'verbal' they are. 'Verb phrase density' may be uneven. Bowie, Wallis and Aarts (forthcoming) show that VP density varies substantially in DCPSE in two important ways: *by genre* – between 110,000 and nearly 160,000 VPs per million words in various genres – and, in some genres, *over time*. This variation is obscured by the fact that, averaged over the LLC and ICE-GB subcorpora, VP density does not change.

In formal face-to-face conversations VP density increases over time by between 8.66 and 15.60 percent (at a 95 percent confidence interval). However, in informal conversations and telephone calls, VP density does not significantly increase between the 1960s and 1990s. Therefore if the progressive is used in certain genres more frequently than others, *the opportunity to use the progressive* must also vary, simply due to this variation in VP density.

[10] This method also picks up 'tag question' progressives such as *Burning the candle at both ends… are you?*$_{\text{[DI-A18 \#162]}}$. The preceeding VP has the feature 'prog' although the auxiliary may not be included in the VP.

When we evaluate rates of progressive VP use, it is more accurate to consider changes in the rate per VP than in the rate per *n* lexical words. By taking this step we remove this VP density variation, and thereby eliminate the possibility that an observed change could be due to changes in VP density. The revised calculation looks something like the following.

	A: item VP(prog)	B: VP	C: rate *p* (proportion)	D: increase $d^{\%}$ (LLC = 100%)
LLC (1960s)	2,973	63,314	4.70%	+21.36% ±5.46%
ICE-GB (1990s)	3,294	57,801	5.70%	
TOTAL	6,267	121,115	5.17%	

Table 2: Change over time of 'VP(prog)' as a proportion of the number of VPs.

Note that we have replaced citations per million words in Column C with the simple proportion *p* (this does not affect the overall calculation). Our results obtain a similar increase ($d^{\%}$) to Table 1, but we have eliminated the possibility that variation in VP density accounted for our results.

Changing the baseline frequency from words to an overarching grammatical class (such as VPs) can have a dramatic effect on results. For example, Aarts, Wallis and Bowie (forthcoming) plotted $d^{\%}$ values for modal auxiliaries *can*, *may*, etc. from DCPSE on a per million word and per modal basis and showed that results differed markedly – *can* rose as a proportion of all modals, but did not change significantly with respect to word frequency; *could*, *would* and *should* all fell with respect to word frequency, but this fall could not be distinguished from an overall decline in modal use.

3.3 Changes in one choice out of a set of alternants

Ideally, we wish to evaluate how the progressive changes over time *where the speaker has the option of using this construction*. The aim should be to focus our experiment on the set of true alternants to which the item in Column A belongs by removing as many distracting factors as possible. In this set of alternants, variation can be hypothesised to take place *between* members of the set, i.e. such that they compete and substitute for one another over time (Wallis 2003).

A study of modal auxiliaries should ideally therefore distinguish between semantic subcategories (deontic, epistemic, etc.) to identify the particular set of alternants at any given juncture. It could also take into account other competing variants to modals, such as semi-modals or adverbial expressions. 'Drilling down' to sets of true alternants can be onerous if particular distinctions (e.g. modal semantics) are not represented in the corpus (see, for example, Close and Aarts 2010 on the modal *must*). We return to this question in Section 4.

Identifying a set of true alternants is often easier said than done. In Aarts, Close and Wallis (2010) we investigated DCPSE to show that the use of the progressive is increasing. The first step, that of focusing on VPs, is easily achieved (see above). Isolating variants is less straightforward. The optimum alternation pattern is between verb phrases which are progressive and those *that could plausibly be turned into* a progressive form (but were not) (Figure 3). We might call the resulting ideal set 'the set of progressivisable VPs'. It is simple to obtain the set of progressive VPs from DCPSE using an FTF which searches for all VPs marked with the progressive feature ('VP(prog)'). The crucial step is to identify this 'progressivisable' subset of VPs. Smitterberg (2005: 45-8) identifies a number of contexts in which verb phrases cannot be progressivised, including imperatives, non-finite VPs, and the *be going to* future construction.

Finally it is possible that the set *itself* may vary over time. Language may contain new innovations, and therefore new alternants, so linguists should ideally incorporate new alternants into their class of 'progressivisable VPs' at the point of their first citation, although these novel cases are unlikely to be sufficiently common to make a difference to an experimental outcome.[11]

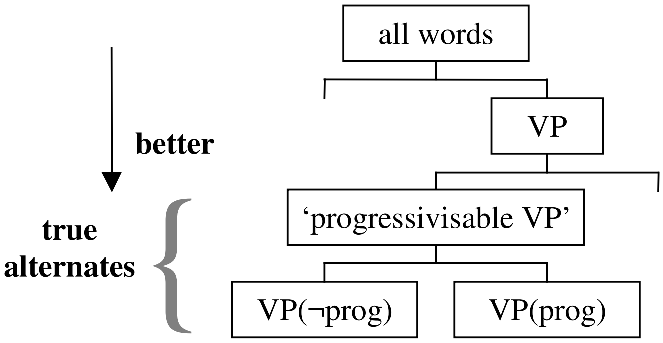

Figure 3: Descending the space of possible choices to focus on true alternants.

As should be clear from the foregoing, the process of identifying variants is, in part, subjective, and hence an approximation, and any experimentalist engaging in excluding material must explicitly state their assumptions. For Smitterberg, removing 'knock-out' contexts from the dataset was not straightforward, so his final calculation of the progressive to non-progressive ratio, which he refers to as the *S-coefficient*, is "a percentage of all finite non-imperative verb phrases (excluding *be going to* + infinitive constructions with future reference) that are in the progressive" (Smitterberg 2005: 48). [12] In Aarts, Close and Wallis (2010) we excluded only imperatives and instances of the *be going to* future. Note that, subject to the limitations of available data, it is entirely legitimate to subdivide an experiment into a series of sub-experiments in order to investigate the rising use of progressive in stative situations only, explore interrogatives only, and so forth.

In Column A we have simply retrieved all cases of VPs marked as progressive ('VP(prog)'). After elimination of Smitterberg's 'knock-out' factors, with the exception of non-finite VPs for reasons discussed in Aarts, Close and Wallis (2010:156, fn8),[13] we narrow down the scope of Column B to this restricted set of progressivisable VPs (indicated by 'VP(+prog)') in Table 3.

Focusing on alternants allows us to estimate the *true rate* of use (Column C) more precisely, and therefore the trend identified is more meaningful. Again, this is not simply a repeat of the previous result. We have eliminated a potential alternative hypothesis remaining from the previous table, namely that the observed increase in the progressive (as a proportion of all VPs) is explained by a corresponding decline in the proportion of 'knock-out' factors.

	A: item VP(prog)	B: VP(+prog)	C: rate *p* (proportion)	D: increase $d^{\%}$ (LLC = 100%)
LLC (1960s)	2,973	62,879	4.73%	+20.95% ±5.31%
ICE-GB (1990s)	3,294	57,599	5.72%	
TOTAL	6,267	120,478	5.20%	

[11] Geoff Leech (personal communication) asks whether this amounts to accepting that the class of 'progressivisable VPs' cannot be built into a model of language change. Our response would be to say that the concept of 'progressivisable VPs', i.e. VPs that could be given a progressive form without violence to their meaning, must necessarily be *defined* in some way – by enumerating types to either include (cf. *will/shall* in the next section) or exclude ('knock-out factors'). If a new 'type' is found in the future we must decide whether or not to include this in our definition.

[12] Non-finite verb phrases were also excluded by Smitterberg because they were difficult to retrieve automatically and because it is possible that there are other factors that constrain variation in non-finite VPs.

[13] In Aarts, Close and Wallis (2010: 156) stative situations were included on the grounds that it is possible for some stative verbs to be progressivised in present day English[0] (see Smith 2000:96). We also included copula constructions and non-finite verb phrases, because checking individual cases would be needed for accuracy as both can be progressivised in some (but not all) instances (compare: *Joan is (*being) tall* and *Joan is (being) friendly; she pretended to (be) sleep(ing)* and *she continued to (?be) sleep(?ing))*.

Table 3: Change over time of 'VP(prog)' as a proportion of the number of progressivisable VPs ('VP(+prog)').

In the case of the progressive, our three baselines turn out to be closely aligned over time. However this does not discount the importance of focusing the experiment as far as possible on the choice. Focusing eliminates the possibility that other sources of variation (e.g. between text genres, or sampling variation) that have an impact on higher order elements in Figure 3, are causing an observed trend, or indeed, as we shall see, obscuring a trend that might be revealed. Smitterberg (2005) found that focusing on progressivisable VPs obtained a different rank order of progressive use between written text genres. Bowie, Wallis and Aarts (this volume) found that the subclass of tensed, past-marked VPs provide a more meaningful baseline for a study of the perfect construction than all VPs.

In identifying semantic alternants we may aggregate grammatically disparate terms. Close and Aarts (2010) investigate the decline of *must* by comparing the frequency of *must* against the frequency of the semi-modals *have to* and *have got to*.[14] In what follows we carry out a quite different case study of linguistic alternation and demonstrate that these same principles apply.

4 A case study: the alternation *shall* versus *will*

4.1 Background

Modal verbs have attracted a lot of attention in the current change literature and *shall* and *will* are no exception. In 1964 Charles Barber wrote:

> [T]he distinctions formerly made between *shall* and *will* are being lost, and *will* is coming increasingly to be used instead of *shall*. One reason for this is that in speech we very often say neither [will] nor [shall], but just ['ll]: *I'll see you to-morrow, we'll meet you at the station, John'll get it for you*. We cannot use this weak form in all positions (not at the end of a phrase, for example), but we use it very often; and, whatever its historical origin may have been (probably from *will*), we now use it indiscriminately as a weak form for either *shall* or *will*; and very often the speaker could not tell you which he had intended. There is thus often a doubt in a speaker's mind whether *will* or *shall* is the appropriate form; and, in this doubt, it is *will* that is spreading at the expense of *shall*, presumably because *will* is used more frequently than *shall* anyway, and so is likely to be the winner in a levelling process. So people nowadays commonly say or write *I will be there*, *we will all die one day*, and so on, when they intend to express simple futurity and not volition. (Barber 1964: 134)

Similarly, David Denison has remarked that:

> During the latter part of our period [1776-present day] ... in the first person SHALL has increasingly been replaced by WILL even where there is no element of volition in the meaning. (Denison 1998: 167)

Comments such as these may lead us to expect that investigating the trajectory of such a change is straightforward. However, from these two quotations alone a number of interrelated issues arise. These are: (i) the status of the variants; (ii) their syntactic behaviour; and (iii) the intended meaning of the clause. In the following discussion, we will address each of these issues.

[14] The authors did not include *gotta* as there were no examples in the corpus and only one example of *got to*. This could be due to the way the corpus was transcribed. As Geoff Leech has pointed out to us (personal communication): *gotta* is not a linguistically well-defined entity. An investigation of the same phenomena in American varieties would undoubtedly include *gotta* in the pool of variants. The same may be necessary of British varieties in future if *gotta* spreads in use.

4.2 Mair and Leech’s work on written English

Recently, Mair and Leech (2006: 327) reported frequency statistics for the perceived decline of the use of *shall*. Their data are based on raw frequency statistics of *shall* and *will* in written British and American English (henceforth BrE and AmE, respectively) from the 1960s and 1990s using the ‘Brown family’ of written English corpora (LOB, F-LOB; Brown, Frown; see Smith and Leech this volume). Counts include verb and negative contractions: e.g., *won’t* and *’ll* are included under *will*.

	British English				US English		
	1960s	**1990s**	$d^{\%}$		**1960s**	**1990s**	$d^{\%}$
will	2,798	2,723	-2.7%	*will*	2,702	2,402	-11.1%
shall	355	200	-43.7%	*shall*	267	150	-43.8%
Total	3,153	2,923	-7.3%	**Total**	2,969	2,552	-14.0%

Table 4: Decline in the use of *shall* in written corpora, LOB/F-LOB and Brown/Frown. (After Mair and Leech 2006.)

This table shows that, comparing four one million word corpora, the frequency of *will* appears to decrease by 2.7 and 11 percent in the BrE and AmE corpora, respectively, and the use of *shall* by almost 44 percent overall in both BrE and AmE corpora. Mair and Leech employ a goodness of fit log-likelihood test comparing absolute frequencies against the overall word count (see Section 3.1) to confirm that this fall in *shall* is statistically significant.

However, as we have noted, this statement simply tells us that *shall* is significantly less frequent as a proportion of words in the later dataset. This is not particularly instructive, not least because there may be many causes of this particular decline. It is possible that the opportunity for speakers to utter *shall* changed (for example, due to variation between text samples), rather than that *shall* declined in use when speakers had the opportunity. What we ideally wish to know is whether *will* is replacing *shall* in circumstances where the writer is in a position to choose.

4.3 Experimenting with *shall/will* alternants in DCPSE

Our experimental data should ideally be restricted to include only cases in contexts where *will* and *shall* are interchangeable. In what follows we outline a number of ‘knock-out’ contexts, attempting to focus on those cases where *will* and *shall* are true alternants and can therefore be said to represent a choice. In addition to declarative cases, *shall* and *will* can appear in interrogative and negative constructions.

(1) a. Interrogatives: *Shall we go to the park?* vs *Will we go to the park?*
b. Negatives: *I won’t/will not go to the park* vs *I shan’t/shall not go to the park.*

However, the semantics of the interrogative cases are distinct from the declarative cases, different usage constraints may apply, or use may be sensitive to genre. Another concern is that the negative cases include the increasingly archaic and informal *shan’t*. We therefore chose to concentrate on the base form in positive declarative utterances, and exclude these ‘knock-out’ contexts.

In Section 2 we discussed the fact that every text unit in DCPSE is given a tree analysis and we can use Fuzzy Tree Fragments (FTFs) to identify cases conforming to a particular structure. To extract declarative cases, we limit cases to where *shall* and *will* are classified as auxiliaries following a subject NP. This will retrieve from the corpus all cases of *shall* and *will* preceded by a pronoun or a noun phrase subject and exclude instances of subject–auxiliary inversion. Figure 4 illustrates the FTF for finding declarative cases of *shall*, results for *will* are obtained by simply substituting the word. At this stage, the lexical slot for the subject NP is unspecified (‘¤’), but we will revisit this later.

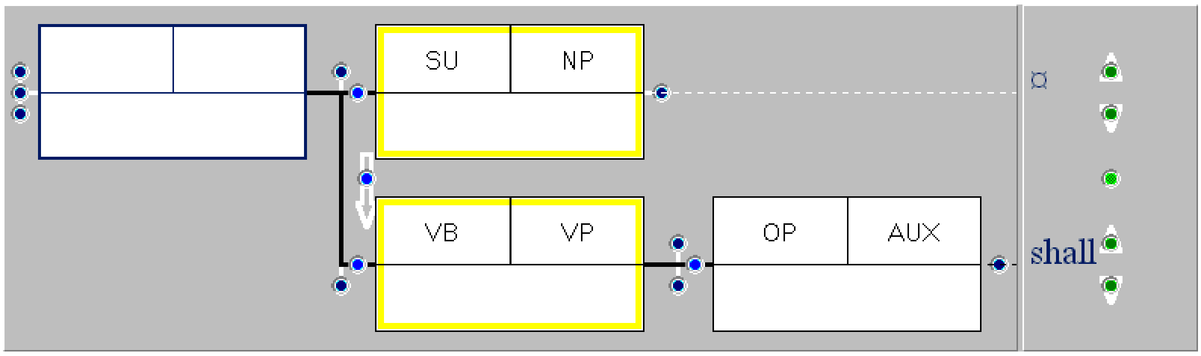

Figure 4: An FTF used to search for *shall* after any subject NP.

A second, similar, FTF was used to retrieve instances of *shall/will not* and these cases were then subtracted from the results. We exclude all negative cases, including *shall not/shan't/will not/won't*. For now, we also exclude the contracted form *'ll*. Results are summarised in Table 5a.

We evaluate the alternation with a 2 × 2 χ^2 test. The χ^2 figures in the bottom row are equivalent to goodness of fit χ^2 tests against the total.[15] The final column contains three figures: the percentage swing $d^{\%}$ from LLC to ICE-GB for *shall* out of the total (see Section 3), the 2 × 2 ϕ effect size measure (see Appendix 2) and the 2 × 2 χ^2 result. The results show a significant change between the two subcorpora, and that most of the variation over time appears, perhaps unsurprisingly, to be attributable to the decrease in the frequency of *shall* (note the high values in the χ^2(*shall*) column).

(spoken)	*shall*	*will*	**Total**	χ^2(*shall*)	χ^2(*will*)	**summary**
LLC (1960s)	124	501	625	**15.28**	2.49	$d^{\%}$ = -60.70% ±19.67%
ICE-GB (1990s)	46	544	590	**16.18**	2.63	ϕ = 0.17
TOTAL	170	1,045	1,215	**31.46**	**5.12**	χ^2 = **36.58**

Table 5a: 2 × 2 χ^2 for *shall* and *will* between ICE-GB and LLC (spoken, positive and declarative; bold is significant for p<0.05).[16] The contracted form *'ll* is excluded.

If we analyse the figures for *shall* and *will* for British English presented by Mair and Leech (see Table 4) using the same method we obtain the results in Table 5b.

(written)	*shall+*	*will+'ll*	**Total**	χ^2(*shall+*)	χ^2(*will+'ll*)	**summary**
LOB (1960s)	355	2,798	3,153	**15.58**	1.57	$d^{\%}$ = -39.23% ±12.88%
FLOB (1990s)	200	2,723	2,923	**16.81**	1.69	ϕ = 0.08
TOTAL	555	5,521	6,076	**32.40**	3.26	χ^2 = **35.65**

Table 5b: 2 × 2 χ^2 for *shall* (+*shan't*) and *will* (+*'ll*, *won't*) between LOB and FLOB (written), data from Mair and Leech.

These results are significant, but the effect size measures ($d^{\%}$ and ϕ) are lower than in our spoken data in Table 5a.[17] The question we might ask therefore, is, are the results *significantly* different?

To answer this question we used a further test. Wallis (2010) defines a 'statistical separability' test to compare the results of two 2 × 2 tests.[18] This finds that the results are significantly separable at

[15] Values in bold are significant at p<0.05 (if they exceed 3.841). The figures on the bottom row indicate whether a particular value (*shall*, *will* etc.) significantly changes over time. The 2 × 2 result simply tells us that 'a change is taking place', but does not tell us where this is happening. High individual χ^2 values indicate cells which have unexpected values.

[16] The contracted form *'ll* and negative cases are excluded. Note that $d^{\%}$ represents the percentage swing of *shall*. Cramér's ϕ is a similar measure, but is calculated across both *shall* and *will* – it measures the size of the *shall/will* alternation (0 = no change over time and 1= complete change). It is particularly useful for comparing results.

[17] In other words, the change is smaller, but still sufficiently large to be judged 'significant' given the data available.

a 0.05 error level, so we are justified in claiming that our experiment obtains a significantly *stronger* result than that obtained using Mair and Leech's method for written data. However, it is not clear whether this fact derives from the exclusion of 'knock-out' contexts, a focus on spoken rather than written material, or simply the different ways in which the corpora were sampled. To investigate this, we modify the experimental design in a series of steps and repeat the separability analysis.

First, we apply Mair and Leech's data collection method to DCPSE. It turns out that the results obtained from our spoken corpus are very similar to their FLOB/LOB results. Changing the corpus does not change the result. The issue therefore seems to concern 'knock-out' contexts.

Staying with our lexical queries, we now eliminate cases of *'ll*. We find that these results *are* significantly distinct from Mair and Leech's, but are *not* significantly different from those obtained with FTFs (Table 5a).

(spoken)	*shall*	*will*	**Total**	χ^2(*shall*)	χ^2(*will*)	**Summary**
LLC	193	812	1,005	**13.87**	2.39	$d^{\%}$ = -48.88% ±16.46%
ICE-GB	91	836	927	**15.04**	2.59	$\phi = 0.13$
TOTAL	284	1,648	1,932	**28.91**	4.98	χ^2 = **33.89**

Table 5c: 2 × 2 χ^2 for the simple lexical auxiliary verb queries for *shall* and *will* between ICE-GB and LLC, all cases, i.e. excluding the contracted form *'ll*.

Results obtained from our spoken data are consistent with those obtained from the written corpora FLOB and LOB. However, if the contracted forms are removed the *shall/will* alternation increases in strength. The use of FTF queries focusing on declarative and positive cases is more restrictive still, but does not obtain a stronger result than this.

We did not test for the impact of eliminating interrogative constructions such as *shall we…?*, or negative constructions such as *you shall not.*[19] Note that this process of testing for statistical separability does not eliminate the need to refine the experimental design: it tells us which changes in the design give significantly distinct results with the data in our possession. As usual in discussions of this kind, with more data a smaller difference between experimental outcomes would be significant.

A further refinement would be to test alternation on a case-by-case basis. In discussing *shall* and *will* we emphasise the need to restrict our queries to *shall* and *will* where the speaker has a choice. Using ICECUP it is straightforward to review the set of cases found by a query line by line. (If the number of cases is large one can check a random subsample to estimate the proportion of problematic cases.) This type of 'health check' is extremely important in corpus linguistics.[20]

When can the alternation take place? Until now we have assumed that all cases of declarative *shall* and *will* can alternate. Here are two examples where the alternation is unproblematic.

(1) a. *…who shall remain nameless* [DI-B04 #208] → *…who will remain nameless*
b. *now Svevo I shall refer to him henceforth* [DL-J02 #240] → *I will refer to him henceforth*

However, some replacements sound awkward to our modern ears. A small number (up to 8) appear to be formulaic, and the alternation may be less likely simply because the word selection is determined

[18] This test uses a z test for two proportions taken from independent populations (Sheskin 1997: 229) to compare the exact swings ($p_1 - p_2$) obtained from each experiment.

[19] Note that in this case Cramér's ϕ is more indicative of a significant difference in strength than $d^{\%}$. The $d^{\%}$ measure is difficult to compare because the difference is divided by the initial value p_1, so similar $d^{\%}$ values can in fact represent distinct results (and vice versa). This is a further argument for favouring ϕ over $d^{\%}$ (see Appendix 2).

[20] As a rule one should always check cases found by a query. This is to minimise the proportion of 'false positives' (cases that should be excluded, possibly because they were incorrectly parsed) and to minimise the number of 'false negatives' (cases that were not found but should have been). Lexical searches can often help identify possible alternative parses and thus false negatives.

by quotation. Thus it is impossible to replace *shall* with *will* in the formulaic *ye shall be saved* (DL-J01 #49) without changing the purpose of the utterance.

A number of linguists have argued that *shall/will* alternation is likely to be more restricted than this. Coates (1983) reviews modal meaning in two 1960s corpora (the *Lancaster Corpus* and the LLC), and argues that second and third person subject *shall* is only found in cases of obligation – a rare meaning for *will*. Similarly, Collins (2009) investigates meaning in a 1990s corpus based on ICE-GB, ICE-AUS and US data. He finds few cases of second person *shall* and, in the third person, almost exclusively deontic *shall*. In expressions of futurity, he casts doubt on whether a traditional prescriptivist rule (*shall* to be used for first person, *will* for second and third) is being followed.

Mindful of these observations, we decided to limit our search to cases where the subject is the first person. We modify the FTF so that the subject consists of a single node and insert the set {*I*, *we*} in the word slot (Figure 5). The results are shown in Table 5d.

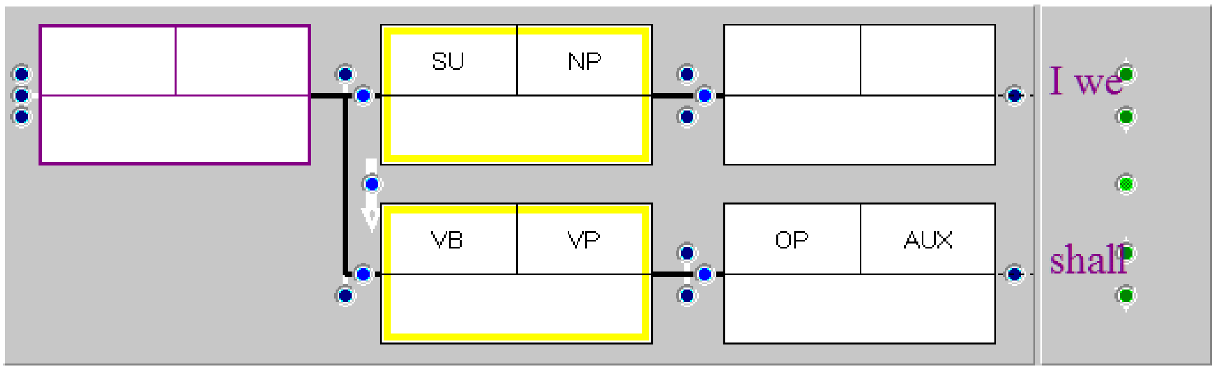

Figure 5: An FTF for a first person subject followed by *shall*. To search for *will* and *'ll* the lexical item *shall* is replaced.

(spoken, 1st ps subject)	*shall*	*will*	**Total**	χ^2(*shall*)	χ^2(*will*)	**Summary**
LLC	110	78	188	1.32	1.45	$d^{\%}$ = -30.24% ±20.84%
ICE-GB	40	58	98	2.53	2.79	$\phi = 0.17$
TOTAL	150	136	286	**3.85**	**4.24**	$\chi^2 = \mathbf{8.09}$

Table 5d: 2 × 2 χ^2 for *shall* and *will* between ICE-GB and LLC (spoken, first person subject, declarative), excluding the contracted form *'ll* and negative cases.

In our declarative data from DCPSE, second and third person *shall* is rare (below 7 percent of cases) whereas the majority of cases of *will* (around 86 percent) are in the second and third person. This tends to support the argument that with second and third person subjects *shall* is rarely an alternative to *will*, even if *will* substitutions are deemed to be acceptable. However, Table 5d shows that in first person cases, if *'ll* is excluded, far from being a residual usage, *shall* is in the majority across DCPSE.

We have already eliminated interrogative constructions, because they may behave differently from the declarative case (section 3.3). It is similarly legitimate to focus on first person declarative constructions and to distinguish between cases where *shall* alternates with *will* and with both *will* and *'ll* together.

4.4 Examining the contracted form *'ll*

So far we have seen that the decision to include or exclude *'ll* in a dataset is likely to lead to different results. The obvious question therefore concerns how we should treat *'ll*. First, we note that it is generally assumed that *'ll* is a contraction of *will*.[21]

We will first analyse all three types as if they mutually alternate at the same level, and then consider a two-level hierarchical analysis. An initial attempt at data retrieval, using the FTF in Figure 5 (*mutatis mutandis*), obtains the following.

	shall	*will*	*'ll*	Total	χ^2(*shall*)	χ^2(*will*)	χ^2(*'ll*)
LLC	110	78	379	567	**9.42**	0.16	2.39
ICE-GB	40	58	370	468	**11.42**	0.20	2.90
TOTAL	150	136	749	1,035	**20.84**	0.36	**5.29**

Table 6a: 3 × 2 χ^2 for *shall/will/'ll* between ICE-GB and LLC, declarative, positive, first person subject cases (ϕ = 0.13).

We have argued that it is important to focus on the investigation of choice between variants as far as possible. However, an obvious issue with the contracted form is its restricted distribution. If we contrast *shall, will* and *'ll* using the FTFs in Section 4.3 (which also exclude negative examples), we overlook the fact that *'ll* cannot plausibly replace *will* (or *shall*) in all syntactic environments. For example, *'ll* is not possible immediately preceding a syntactic gap such as an ellipsis site, or in sentence-final position. In the corpus examples below, therefore, *'ll* is not a possible alternative to *shall* and *will*:

(2) *So I won't be in on Monday but he will* [**he'll*]. [DI-B04 #258]
(3) *Of course we shall* [**we'll*]. [DL-D05 #0109]

We therefore further restrict instances of *shall* and *will* by excluding cases where the auxiliary is in the final position in the VP. The modified FTF is in Figure 6. The black line highlighted as '**Last child:** no' requires that in any matching case the auxiliary cannot occupy the final position in the VP. Results are given in Table 6b.

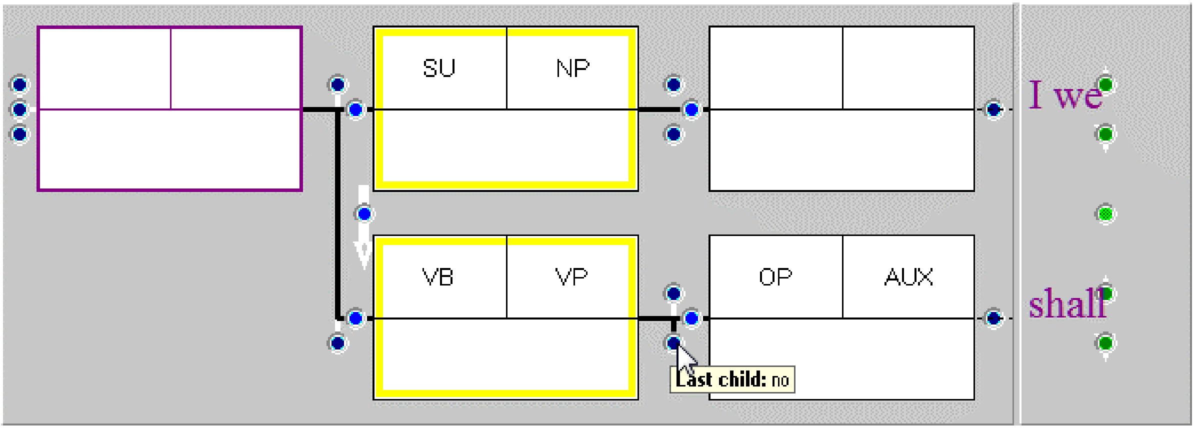

Figure 6: An FTF which specifies that *shall* may not be in VP final position.

[21] Leech (personal communication) takes the view "that *'ll* is a contraction of *will*, because (a) [w]-elision is common in the history of English (cf. *for'ard, Norwich, hussy, Harwich, Warwick, Dulwich, innards*) whereas [ʃ]-elision is not; and (b) whereas in PDE *shall* is largely restricted to first-person subjects, *'ll* occurs equally well with first-, second- and third-person subjects (*I'll, you'll, she'll,* etc.)."

	shall	*will*	*’ll*	Total	χ^2(*shall*)	χ^2(*will*)	χ^2(*’ll*)
LLC	104	69	371	544	**9.98**	0.13	2.33
ICE-GB	36	52	365	453	**11.98**	0.16	2.80
TOTAL	140	121	736	997	**21.96**	0.30	**5.13**

Table 6b: A comparison of the three alternants in declarative first person, VP non-final position (ϕ = 0.11).

In terms of raw frequencies, the number of cases of *shall* falls dramatically from LLC to ICE-GB (from 104 to 36), *will* falls at a slower rate, and *’ll* appears numerically stable.

However, this is potentially misleading. As a proportion of the three variants (i.e. examining their relative proportion), we find that *will* appears to be numerically stable over time at around 12 percent, *shall* falls from around 19 percent to 8 percent, whereas *’ll* increases its share from 68 percent to 80 percent. This pattern is reflected in the χ^2 values in the columns on the right. The 3 × 2 χ^2 test is significant, and the pattern of relative change is similar to before the VP-final exclusion was applied.

We can also carry out an analysis of this data by grouping the modals hierarchically {*shall*, {*will*, *’ll*}}. The idea is that speakers are making decisions at two levels – to employ *shall* or *will*, and whether or not to contract *will*. We therefore employ two 2 × 2 χ^2 tests, one at each level. We find that the *shall* vs. *will*+*’ll* alternation is significant and the proportion of *shall* cases significantly falls over time, but the contraction alternation does not obtain a significant result.[22]

4.5 Plotting trends over time

DCPSE date-stamps each spoken recording with the year that it was made. As our evidence suggests a decline in the use of *shall* over time, we can plot this trend on a year-on-year basis. We plot *shall* against two baselines: against the uncontracted *will* and against *will* plus the contracted form *’ll*. In so doing we revisit the concept of what we called the ‘true rate’ of alternation.

In carrying out a plot over time, we introduce an additional potential source of variation, because the number of texts per year and the sampling conditions under which they were obtained, are not evenly balanced in each annual subcorpus. However, the advantage of considering our data as a time series – compared to the contingency table approaches thus far – is that we can adjust for the differences in LLC and ICE-GB sampling periods. The LLC portion, while nominally described as ‘1960s’, was sampled over a period from 1958 to 1977, whereas ICE-GB was recorded between 1990 and 1992.

Table 7 shows data for first person *shall* vs *will* by year on the left hand side. For each year, *p*(*shall*) is the fraction of cases of *shall* out of the total *n*. On the right hand side we carry out the same procedure for *shall* vs *will*+*’ll*. Data retrieval involves the same method as we employed previously: employing the FTF pattern in Figure 6 and subtracting negative cases.[23]

[22] Pairwise comparisons with 2 × 2 tests find that *shall* vs. *will* and *shall* vs. *’ll* both yield significant results at p<0.05. It is possible to come up with a number of permutations of tests, but note that there are only two degrees of freedom in the table, which are described by the hierarchical decision tree proposed. (All other results are a corollary of those cited.)

[23] Probabilities cited here have one obvious problem: the total number of cases in a given year varies widely. We address this through the use of confidence intervals when we plot the data.

Year	*shall*	*will*	**Total** *n*	*p*(*shall*)
1958	1	0	1	1.0000
1959	1	0	1	1.0000
1960	5	1	6	0.8333
1961	7	8	15	0.4667
1963	0	1	1	0.0000
1964	6	0	6	1.0000
1965	3	4	7	0.4286
1966	7	6	13	0.5385
1967	3	0	3	1.0000
1969	2	2	4	0.5000
1970	3	1	4	0.7500
1971	12	6	18	0.6667
1972	2	2	4	0.5000
1973	3	0	3	1.0000
1974	12	8	20	0.6000
1975	26	23	49	0.5306
1976	11	7	18	0.6111
1970	0	0	0	?
1990	5	8	13	0.3846
1991	23	36	59	0.3898
1992	8	8	16	0.5000

Year	*shall*	*will*+*'ll*	**Total** *n*	*p*(*shall*)
1958	1	3	4	0.2500
1959	1	5	6	0.1667
1960	5	9	14	0.3571
1961	7	40	47	0.1489
1963	0	4	4	0.0000
1964	6	17	23	0.2609
1965	3	16	19	0.1579
1966	7	24	31	0.2258
1967	3	17	20	0.1500
1969	2	32	34	0.0588
1970	3	3	6	0.5000
1971	12	21	33	0.3636
1972	2	15	17	0.1176
1973	3	3	6	0.5000
1974	12	23	35	0.3429
1975	26	165	191	0.1361
1976	11	38	49	0.2245
1970	0	5	5	0.0000
1990	5	33	38	0.1316
1991	23	246	269	0.0855
1992	8	138	146	0.0548

Table 7: Frequency and probability data from DCPSE reflecting a declining use of *shall* over time as a proportion *p*(*shall*) of the set of alternants {*shall*, *will*} (left) and {*shall*, *will*, *'ll*} (right), following first person subjects (non VP-final).

First, we plot *shall* against a baseline set {*shall*, *will*} in Figure 7a. We employ a scatter-plot to record the probability (*p*) of *shall* rather than *will* being selected by a speaker, against the year the material was recorded. The dotted lines represent the upper and lower estimated trend lines and the crosses represent the mid-points of the LLC and ICE-GB data.

The vertical 'I'-shaped error bars express the Wilson confidence interval[24] for each data point. A large confidence interval means a greater level of uncertainty. Where samples are tiny (as here), confidence intervals will be extremely broad. The LLC data in particular is a 'cloud' from which no real trend can be inferred (hence two questionable trend lines).

[24] We calculate error bars using Wilson's score interval (see Appendix 1). In preference to the commonly-used Gaussian method, the Wilson interval compensates for skewed data (*p* can even be zero or 1, as Figure 7 reveals), and may be used with tiny samples (Wallis, 2009). Error bars are unequal and tend toward the centre of the probability range (i.e. 0.5).

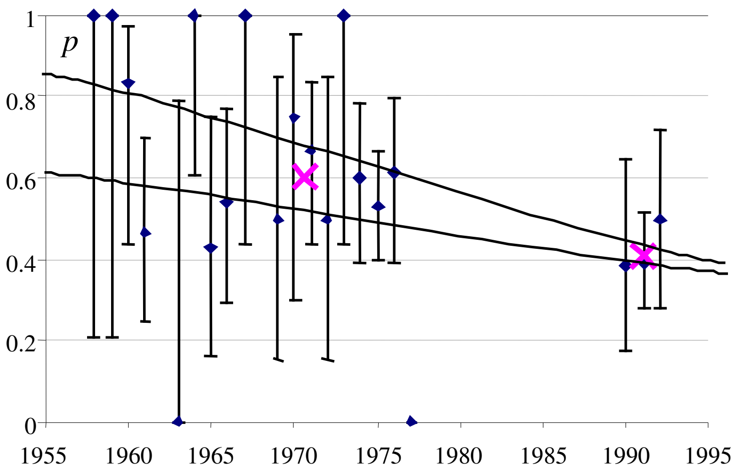

Figure 7a: Declining use of *shall* as a proportion *p* of the set {*shall*, *will*} with first person subjects, annual data, with Wilson intervals. The broad confidence intervals (I-shaped 'error bars') make it difficult to infer a single trend line (hence the upper and lower estimated trend lines indicated by the dotted lines). 'X' marks the centre-point of each sub-corpus.

The problem with this graph is the spread of data. Perhaps a better strategy with this dataset is to aggregate years together into five-year periods. Note that we are not really expecting to see an annual steady decrease in the use of *shall*, rather we are attempting to estimate the rate of change over the period. We can group data into half-decades indicated in Table 7, and plot the results in Figure 7b. The trend becomes clearer as a result.

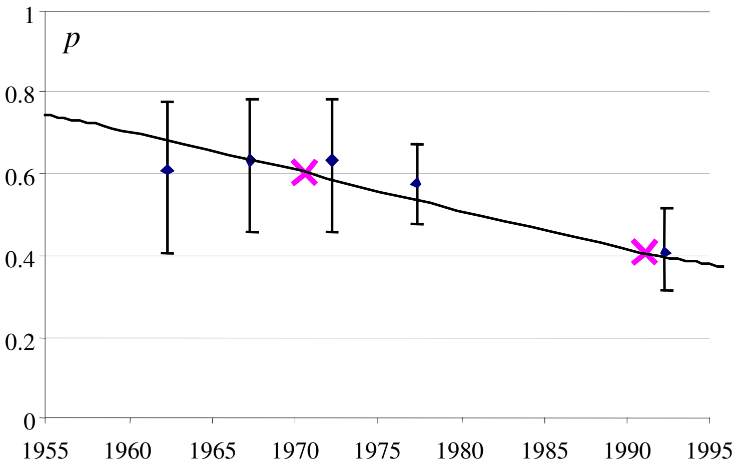

Figure 7b: Declining use of *shall* as a proportion *p* of the set {*shall, will*} with first person subjects, half-decade data ('1960' = 1958-62 inclusive, '1965' = 1963-67, etc.)

Putting Figure 7b into words: in declarative first person contexts, *shall* appears to be being replaced by *will*, with *shall* falling from around 60 percent of cases in or around 1970, to about 40 percent by the early 1990s. This suggests a switch from one dominant form (and therefore what speakers might consider to be the default choice of modal auxiliary verb) from *shall* to *will* over this period.

These results may also tie in with Collins' (2009) observation that the traditional prescriptive rule regarding preference for the first person usage of *shall* did not appear to apply to his 1990s data. If this is the case then it could be that the almost total dominance of *will* in second and third person usages is undermining this rule.

Finally we examine the effect of plotting *shall* as a proportion of the set including both forms of *will*, i.e. {*shall*, *will*, *'ll*}. The data is given in the right hand part of Table 7 and plotted in Figure 7c below. We can estimate a true rate for *shall* falling from around 20 percent in 1970 to below 10 percent in the early 1990s. Considered in this way, the data does not appear to represent a change in the dominant form.

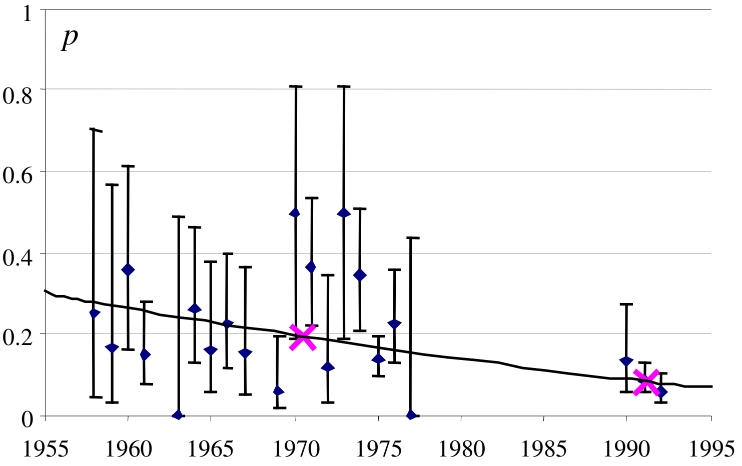

Figure 7c: Declining use of *shall* as a proportion *p* of {*shall, will, 'll*} with first person subjects, annual data. Note how the expansion in the baseline condition to include *'ll* makes *shall* a minority choice over the time period.

In conclusion, the choice of baseline that a researcher adopts is premised on the hypothesized set of alternants available to a speaker at any given point in time.

Both baselines are plausible and the results are meaningful: we can restrict our study of alternation to cases where the speaker chooses to use the uncontracted form, in which case it appears that we see a change in the dominance of the uncontracted modal; alternatively we can opt to include all cases of modal futurity and see that *shall* declines as part of a larger set.

We might argue that *shall* does not alternate *as freely* with *'ll* as with *will* – perhaps informal contexts cause speakers to employ contracted forms more frequently. In this case we would be justified in excluding *'ll* from a study, just as we excluded other 'knock-out' contexts.

4.6 Modal meaning

In our discussion of *shall* and *will* we have not addressed the issue of modal meaning. We have assumed that *shall* and *will* compete regardless of their meaning. However, work by Smith (2003), Leech (2003), Leech *et al.* (2009) and Close and Aarts (2010) suggests that this is unlikely. It is therefore necessary in our investigation of *shall* and *will* to investigate the level of competition according to semantic classification. This is also necessary if we are to reach any conclusions about reasons for change in the modal system.

All first person positive declarative instances of *shall* and *will* (but not *'ll*, which was omitted for reasons of time) were therefore manually coded according to whether the modal expressed Root or Epistemic meaning. We follow the classification system proposed in Coates (1983) whereby the Root meanings of *shall* include 'obligation', 'intention' and 'addressee's volition' (typically found in interrogatives, which were not included here), while Epistemic refers to 'prediction' (= 'futurity') (Coates 1983: 185). With respect to *will* the Root meaning includes 'willingness' and 'intention' (both of which can be subsumed under the heading 'volition') and Epistemic meanings include 'predictability' and 'prediction' (Coates 1983: 169-170). Illustrative examples from the DCPSE corpus are as follows:

(4) Root:
a. *I've got some at home so I shall take it home.* [DI-A18 #30]
b. *I will answer you in a minute.* [DI-B30 #293]

(5) Epistemic:

a. *So I shall have roughly from the twenty-ninth of June to the eighth of July on which I can spend the whole of that time on those two papers.* [DL-B01 #62]

b. *It's certainly my long term hope that I will have some kind of companion...* [DI-B53 #0257]

According to Coates (1983: 170), there are many cases of 'merger' found with *will* which makes coding difficult. In particular, in active clauses with an agentive subject and an active verb which is not progressive or perfective it is often difficult to decide whether *will* refers to a future event which is likely to take place (Epistemic meaning), or whether the subject is indicating an intention to carry out an action (Root meaning). The examples provided in (6) are ambiguous: in (6a) it is unclear whether the speaker intends to do half as much work or whether his statement is to be interpreted as 'it is inevitable that (in the future) I will have no choice but to do half as much work', and in (6b) *will* is ambiguous between intention and prediction (future).

(6) a. *So I said, "this just means I shall do half as much work", and he said, "very well".* [DL-B16 #224]

b. A: *Are you going to stay at that house then?* [DL-B30 #39]
B: *Well, I will be for the next couple of months.* [DL-B30 #40]

Obviously coding is a subjective exercise, and this raises problems when comparisons between results from different studies are compared. This is unavoidable.

Our results are summarised in Table 8a. Investigating the distribution of semantic types as a proportion of the total reveals a shift in the use of *shall* over time. The overall fall in *shall* appears to be due to a sharp decline in the number of cases of Epistemic *shall*, over 80 percent of which appear in the earlier subcorpus.

	Source corpus	**Root**	%	**Epistemic**	%		**Unclear**	%	**Total**
shall	**LLC**	33	30.84	72	67.29		2	1.87	107
	ICE-GB	22	59.46	14	37.84	← **sig**	1	2.70	37
will	**LLC**	44	55.70	28	35.44		7	8.86	79
	ICE-GB	37	66.07	14	25.00		5	8.93	56
Total		136		128	↑ **sig**		15		279

Table 8a: Distribution of semantic types of *shall* and *will* in first person positive declarative utterances in DCPSE.[25] Percentages are quoted of the total for *shall* and *will* in each row. Significant results of 2 × 2 χ^2 tests (at p<0.05 level) applied to the Root and Epistemic columns (Total row), and to the *shall* and *will* rows (column) are indicated by 'sig'.

Our results lend support to the argument that change in the modal system is related to the semantics of the modal auxiliaries (see Leech 2003, Smith 2003, Leech *et al*. 2009, Close and Aarts 2010). Specifically, we observe a sharp decline in Epistemic *shall*.

Table 8a contains three variables (source corpus, lexical item and modal meaning). In order to break down this three-way design we select two variables and subdivide the data by the third.

First, let us consider alternation over time for the Root and Epistemic subsets. Root and Epistemic *shall/will* alternation is analysed in Tables 8b and 8c, respectively.

Root *shall/will* is stable and the results are not significant. However, the alternation for Epistemic *shall/will* is statistically significant: indeed, out of the choice of *shall* and *will* in Epistemic contexts, *shall* declines in use as a proportion of the total by an estimated thirty percent (although note the large confidence interval). This analysis separates out Epistemic *shall* from the baseline

[25] Note that the frequencies are slightly lower than those presented in Table 5d because unfinished or unclear utterances were excluded in the coding process.

(Epistemic modals). The fall in *shall* is therefore not simply attributable to the sharp fall in Epistemic modals from 100 to 28: rather, we have evidence for a shift in use from Epistemic *shall* to *will*.

Root	*shall*	*will*	**Total**	χ^2(*shall*)	χ^2(*will*)	**Summary**
LLC	33	44	77	0.11	0.08	$d^{\%}$ = -12.99% ±38.83%
ICE-GB	22	37	59	0.15	0.10	$\phi = 0.06$
TOTAL	55	81	136	0.26	0.18	$\chi^2 = 0.32$ns

Table 8b: Analysis of change over time for first person declarative Root {*shall*, *will*}. The results are not significant and the overall change ϕ is small. Percentage swing $d^{\%}$ represents the change over time in the proportion of cases of *shall*. This is not significant (note that the confidence interval is bigger than the estimated change). 'ns' = non significant.

Epistemic	*shall*	*will*	**Total**	χ^2(*shall*)	χ^2(*will*)	**Summary**
LLC	72	28	100	0.34	0.71	$d^{\%}$ = -30.56% ±27.33%
ICE-GB	14	14	28	1.23	2.52	$\phi = 0.19$
TOTAL	86	42	128	1.57	3.23	χ^2 = **4.80s**

Table 8c: Analysis of the first person declarative Epistemic {*shall*, *will*} alternation set over time. *Shall* declines from being the majority Epistemic modal in the LLC '1960s' data, to being equal in frequency to *will* in the ICE-GB subcorpus. The results are significant ('s' = significant) and the overall change ϕ is substantial.

We may also examine whether there is any change in how *shall* and *will* are used. We carry out 2 × 2 χ^2 tests for the upper and lower rows in Table 8a (i.e., excluding the 'Unclear' column). We find that Epistemic *shall* is declining while Root *shall* increases its proportion ($\phi = 0.27$). The *will* data (lower rows) does not obtain a statistically significant difference.

Overall, Table 8a appears to indicate that Root *shall* had already declined to a 'rump' by the 1960s, and the numerical decline in Root *shall* in our data is not significant. Our analysis identifies a secondary decline in usage of Epistemic *shall*, taking place in spoken British English between the 1960s and 1990s. Returning to the comment made by Barber (1964: 134) that the "distinctions …between *shall* and *will* are being lost", we suggest that the decline in Epistemic *shall* is actually making *shall* and *will* more distinct (or, to put it another way, making *shall* more marked).

An examination of the percentages of *will* and *shall* synchronically shows that, in the 1960s data, two thirds of cases of first person *shall* were Epistemic, whereas around 55 percent of cases of *will* were Root. The decline of Epistemic *shall* means that around 60 percent of cases of *shall* during the 1990s were Root[26] – a similar proportion to *will*. If we also consider cases of *shall* and *will* in second and third person contexts, we find that the vast majority of cases of *will* (around 80 percent) in both time periods were Epistemic. A possible explanation for the decline of first person Epistemic *shall* signalling 'prediction', therefore, is simply that a dominant alternant, i.e. Epistemic *will*, is spreading from second and third person contexts to the first person.

4.7 *Be going to* versus the modals

A current change study of *shall, will* and *'ll* would not be complete without some discussion of the semi-auxiliary *be going to,* also known as the *'going-to* future' for its ability to replace *will* and/or *shall.* Our concern here is not the development of *be going to* (for this the reader is referred to Hopper and Traugott 2003, Mair 2006, and references therein), but the possible competition with the modals *shall* and *will.*

At this point, we wish to compare the distribution of *be going to* against that of *shall, will* and *'ll*, so we shall follow the principles laid out above. That is, we will retrieve from the corpus all

[26] According to Coates (1983), Root meaning of *shall* is most often found in interrogatives, which we excluded from this study. Including them would increase the overall proportion of Root uses of *shall*, further widening the divide between *will* and *shall.*

instances of *be going to* which may alternate with each of the other variants. We use the FTF in Figure 8, again exploiting the parsed corpus. The grammatical annotation of the corpus makes a distinction between the *be going to* future and the verb *go* followed by a preposition, e.g. *I'm going to London,* which makes data retrieval straightforward.

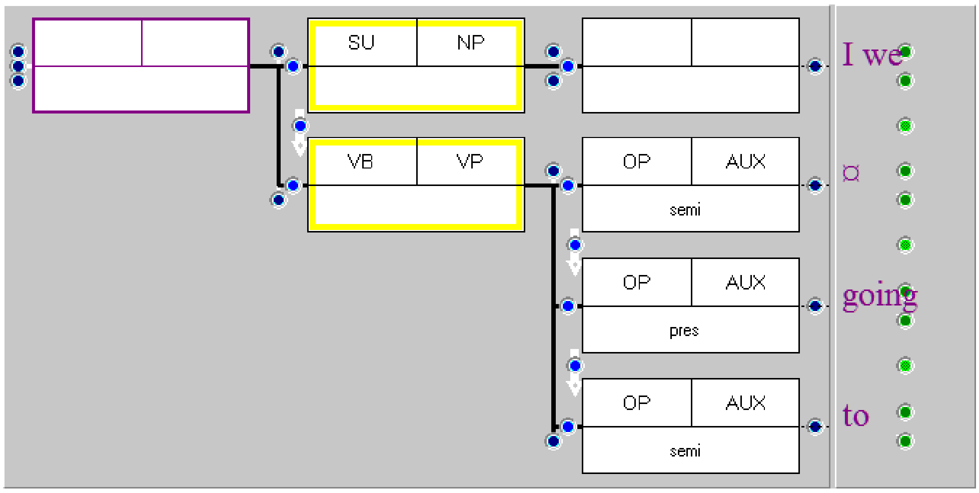

Figure 8: An FTF for *be going to* in the same syntactic context as before.

Cases of *be going to* can be retrieved from DCPSE with a single FTF which specifies the lexical items *going to* (identified in the corpus as 'auxiliary') and the feature 'present'. Without specifying the tense as 'present', data such as *I was going to say…* are retrieved. As these cannot alternate with *will* and *shall* we exclude them from our study.

Recall that we are only concerned here with positive, declarative first person contexts. The presence of an auxiliary node preceding *going* rules out the possibility of retrieving instances of subject–auxiliary inversion, and therefore excludes interrogative cases. To exclude negative cases an additional FTF was created which specified the presence of *not* or *never* between the first auxiliary node and *going*. These numbers were then subtracted from the total tokens retrieved using the FTF in Figure 8.

Finally, as with *'ll*, only cases of *be going to* that do not precede a syntactic gap or occur in sentence-final position were retrieved, as these cases can alternate with the contracted form.

	shall	*will*	*'ll*	*be going to*	**Total**	χ^2(*shall*)	χ^2(*will*)	χ^2(*'ll*)	χ^2(*be…*)
LLC	104	69	371	138	682	**10.46**	0.18	1.92	0.11
ICE-GB	36	52	365	124	577	**12.36**	0.22	2.27	0.13
TOTAL	140	121	736	262	1,259	**22.82**	0.40	**4.20**	0.24

Table 9: Results including the alternation for *be going to* in the first person.

This 4 × 2 χ^2 test is significant, but to identify where values are changing with time requires us to investigate further. A useful next step simply compares each type (*shall*, *will* etc.) against the remainder of the variant set with a 2 × 2 test. In effect, we ask 'does this type differ in its behaviour from the rest of the alternant set'? Out of the set of four types, *shall* significantly decreases its share of cases, whereas *'ll* significantly increases.

To conclude (and to neatly return to our discussion of baselines of change in Section 3), Figure 9 summarises the pattern of observed change over time in two ways. We plot percentage swing in 'per million word' ('absolute') and 'within set' ('relative') terms. The results are distinct. Measured simply against the numbers of words in the corpus (Figure 9, left), *shall* falls significantly over time. The overall set falls in number, but this change is not deemed significant.

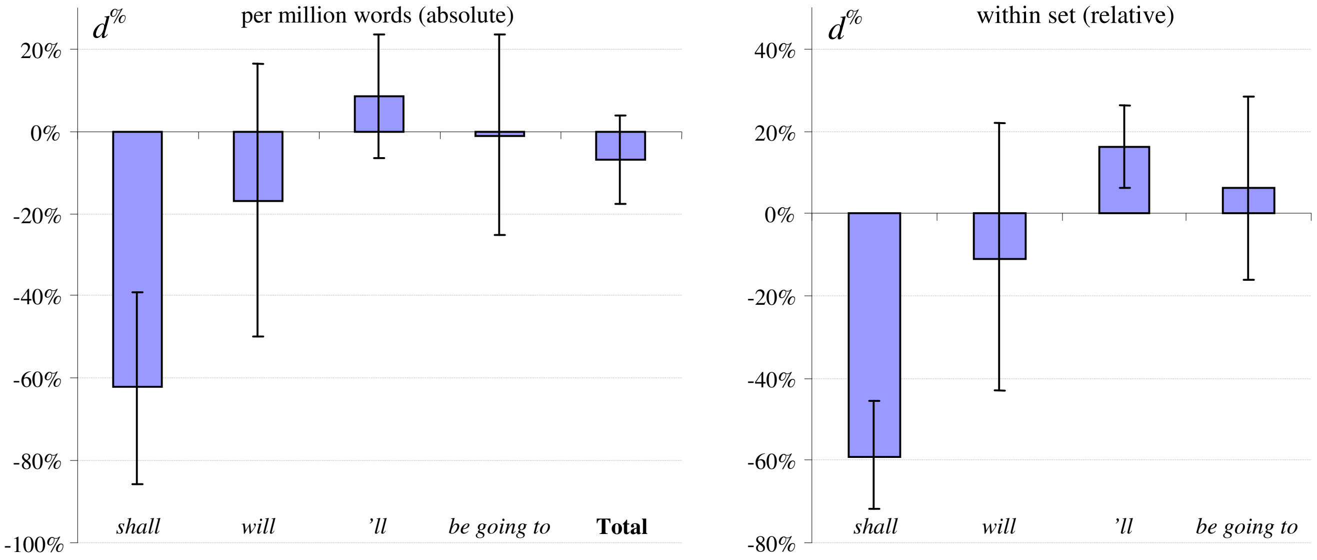

Figure 9: Summarising changes for *shall*, *will*, *'ll* and *be going to*, first person positive declarative (non VP-final) over time, plotting percentage swing ($d^{\%}$), on an absolute ('per million word') and relative ('within set') basis, with 95% confidence intervals (Newcombe-Wilson method, see Appendix 1).

If we examine relative change within the set of alternants (Figure 9, right), we 'factor out' any overall decline. Each error bar visualises the relevant 2 × 2 test mentioned above. The graph is an easy way to identify changes in the *share* of the set of alternants over time: where confidence intervals do not cross the axis they are significant. We can now also see that not only is *shall* falling as a proportion, *'ll* significantly rises. This pattern of change is obscured in the left graph.

One remaining question is whether alternation could genuinely occur in each case. We have restricted the context of our cases to those involving the first person because it seems probable that only these are likely to alternate with *shall*. However, *will* and *be going to* appear in second and third person contexts with future meaning, and these should be investigated separately.

Note that whilst we indicate that *'ll* and *be going to* may alternate with *shall/will*, the proposition that alternation is feasible in each case may need further investigation. For reasons of time we were unable to exhaustively evaluate every single case of *'ll* and *be going to* in our datasets to test them for replacement with either *shall* or *will*. Nonetheless, the overall result seems clear: *shall* is falling relative to its possible alternants, and *'ll* is increasing in use.

5 Conclusions

In this paper we have shown how the *Diachronic Corpus of Present-day Spoken English* can be used to track short-term changes in the use of the progressive construction and in the use of the modal auxiliaries *shall* and *will*. We summarised a number of methodological considerations for variants to be compared meaningfully. In particular we emphasise the importance of focusing on alternation where a choice may exist. The alternative is to include cases which do not alternate, and thereby introduce confounding variation into the experiment.

We hope that we have demonstrated that this type of focusing can make a very real difference to our understanding of change. The picture one obtains by examining change (e.g. over time) within a set of variants may be qualitatively different from that obtained by measuring change on a per million word basis (see Figure 9). Both pictures are true, but they need to be understood together. Only one of these pictures allows us to investigate whether a decline is due to a changing outcome of a choice.

Analysing corpus data is (inevitably) an *ex post facto* analysis of naturally occurring data. Unlike a lab experiment, we cannot constrain experimental conditions and ensure that the choice exists in advance. We are obliged to *infer* that a choice existed at the point the utterance was made, working backwards from our data, constraining case retrieval grammatically and examining cases. In our favour, our results are natural and uncued, and we cannot inadvertently introduce an 'experimenter bias' through the artificiality of data-collecting. It is necessary to carefully construct an experimental design, such as the one in this paper, to have confidence in the results.

In particular, we have

a) focused on variation between genuine alternating speaker choices as far as possible (and failing this, used a baseline as close to the choice as possible);[27]

b) plotted change over time series data;

c) examined change within subsets of the data, identifying differing behaviour of subsets classified by modal meaning; and

d) compared these results with those of other alternating types, extending a simple pair-wise alternation into a hierarchical set of binary choices.

Our initial results for *shall* vs *will* demonstrated a significantly greater change than that found in Mair and Leech's data. By carrying out a small number of intermediate experiments and comparing their results, we narrowed down the difference to the exclusion of 'knock-out' contexts of interrogative and negative cases, and finally, second and third person subjects. We also showed how it was possible to plot the fall in the use of *shall* over a time series, revealing an apparent shift in dominance from *shall* to *will* between 1960 and 1990.

By examining modal meanings we found that the fall in *shall* was attributable wholly to Epistemic *shall*, with Root cases remaining stable over time. Extending the alternation experiment to include *'ll* and *be going to*, both in non VP-final position, permitted us to identify that the fall in *shall* was robust and held up when cases of *'ll* were included with *will*. Moreover, when the set is expanded further to include *be going to*, the contracted modal *'ll* can be seen to rise as a proportion of the set. With the exception of modal semantics, where manual coding was necessary, our experiments exploited the parsed corpus to obtain results.

Appendix 1: Employing statistical tests and handling small, skewed samples

In this article we have concentrated on refinements to the experimental design, rather than the use of particular statistical tests. The development of parsed corpora such as DCPSE and ICE-GB permits more precise experiments to be elaborated using FTFs than are possible with POS-tagged corpora. Inferential statistics is secondary to sampling. One can only partially compensate for a defect in data collection by an improved test. However, in discussing methodology we must occasionally deal with statistical issues, and a greater understanding of statistics can improve experimental design.

The central test employed in the type of experiments described in this book is the chi-square (χ^2) **contingency test**, expressed as a 2 × 1 'goodness of fit' test or as a 2 × 2 test for independence or 'homogeneity' (see Wallis 2010). These tests tell us whether a deviation from an expected value, or pattern of values (a distribution), is sufficient to be significant at a given error level, typically 5%.

The principles that underpin χ^2 tests also support the calculation of **confidence intervals**. A confidence interval is the range of values around a single observation that the true value in the population might be in given the evidence. This allows us to observe, for example, with 95% confidence (or a 5% probability of error), an increase in progressives per million words in DCPSE of

[27] One of the characteristics of a 'real change' is that one may be able to observe it in many ways, including by less optimal methods than those recommended here. However, the conclusion should be clear: as far as is feasible one should focus on alternation where a choice exists, and distinguish between subcategories. It may not always possible to 'drill down' to the variants as we have done in this paper. In particular, difficulties may arise when using unparsed corpora, or when investigating a variant with no obvious variants (or too many variants). In cases such as these, the aim should be a 'best fit' of what we propose here. For example, the task of identifying all latinate verbs which could alternate with phrasal verbs is likely to be particularly arduous in a sizeable corpus. However, one may take a random subsample to estimate their true rate and thereby populate a contingency table.

+22.13% ±5.48% (Table 1). Confidence intervals and significance tests are related. Since 22.13 – 5.48 > 0, the change is *significantly different from zero*, i.e. 'significant'.

By far the most common method for calculating confidence intervals assume that repeated sampling at or around an observation obtains a symmetric, approximately Normally-distributed ('Gaussian') interval (Wallis 2009). The formula for the popular Gaussian single-sample interval is simply $p \pm z\sqrt{p(1-p)/n}$, where z is the critical value of the Normal distribution, and n the total number of observations.

However this rough approximation is rather inaccurate when an observation is very skewed (close to 0 or 1) or limited data is available. As p approaches either 0 or 1, the confidence interval must tend toward the centre, because not only the value, p, but also *the confidence interval around* p, must logically fall within the range of probability [0, 1]. Since low probability terms are not infrequent in linguistic data it is important to examine this question carefully![28]

There are, in fact, two problems with this formulation, distinguished by Wallis (2009), namely

i) that the Normal approximation to the Binomial is inaccurate for small samples, and should be corrected for continuity, and

ii) that the Binomial model on which contingency tests are based predicts a confidence interval on the population probability P rather than the observation p.

The first problem is widely known, and a number of alternative methods have been proposed. Comparing standard 2 × 1 χ^2, Yates' χ^2 and log-likelihood G^2 against exact Binomial calculations for all values of p and different sample sizes n, Wallis (2009) showed that Yates' formula had the lowest overall error, followed by χ^2. Log-likelihood, considered by some (e.g. Rayson 2003) to be an improvement on χ^2, was in fact rather less accurate.

The second problem was first recognised by Wilson (1927). He proposed a different formula for the interval on p, termed the *Wilson score interval*. Wilson's formula is at first sight more intimidating, but it is actually straightforward to construct in a spreadsheet.

(E1) *Wilson's score interval* $(w^-, w^+) \equiv \left(p + \frac{z^2}{2n} \pm z\sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}}\right) \Big/ \left(1 + \frac{z^2}{n}\right)$.

Newcombe (1998a) shows that this interval is superior to competing methods and argues for the Gaussian interval to be actively discontinued in its favour. It can also be corrected, Yates-like, for continuity. Wilson's interval on p neatly reflects the Gaussian interval on P, that is, if p is at the upper bound for P, P will be at the lower bound for p (Wallis 2009).

Crucially, Wilson's method allows us to robustly estimate confidence intervals on skewed values of p. In this article, we cite Gaussian intervals on $d^{\%}$ for ease of quotation, but we always compare results with those obtained using Wilson's score intervals. However, with very small data sets or skewed values there is no choice. The time series data in Table 6 can only be analysed using Wilson's method. The data set contains tiny annual samples and some of these are highly skewed. Figures 7a-c display Wilson intervals as error bars.

So far we have discussed the single sample ('goodness of fit') interval. Newcombe (1998b) considers alternatives to 2 × 2 χ^2 tests where a single interval representing the difference between two observed samples is calculated (Wallis 2009). He finds that a difference interval based on Wilson's formula is the most accurate, improving on log-likelihood and chi-square tests. For precision, Figure 9 uses this interval. We have checked our *will/shall* experiments against Newcombe's interval, and the formulae are available in an online spreadsheet.[29]

[28] Linguists are not the only ones with this problem. Medical scientists are often concerned with research into the effect of intervention in cases of low probability events, e.g. heart attacks. Medical statisticians have addressed this problem and we here draw linguists' attention to their findings.

[29] Method 10 in Newcombe (1998b). See also Wallis (2010). A spreadsheet www.ucl.ac.uk/english-usage/staff/sean/resources/2x2chisq.xls contains an implementation of this interval (and thus a more accurate 2 × 2 "chi-square test").

In summary, employing Wilson intervals and Yates' tests improve precision where conventional χ^2 tests break down: in small, highly skewed datasets. These methods are particularly valuable for corpus linguists, who frequently deal with data of this kind.

Appendix 2: Measures of change

A second methodological question concerns the measurement of *effect size*, i.e. estimating the size of the change in the rate of 'VP(prog)', the decline of *shall*, etc. Statistical significance tells us that the difference is unlikely to be zero (at a given level of confidence, see above). It does not tell us how large this difference actually is.

In the paper we quote the percentage increase (or decrease) $d^{\%}$ of a variant relative to the first subcorpus of DCPSE (the material from the LLC) with a baseline of 100 percent, and we calculate confidence intervals on $d^{\%}$. This approach is relatively intuitive, but it can be misleading – not least because an increase of 20% (say) followed by a decrease of 20% does not bring you back to the start ($p \times 1.2 \times 0.8 = 0.96p$, not p). It also has the rather unhelpful mathematical property of being unconstrained (it can have any value from minus to plus infinity).

In the statistics literature a number of measures of effect size are occasionally cited. These include the odds ratio, the contingency coefficient C and Yule's Q (Sheskin 1997: 244). A standard measure called Cramér's ϕ can be applied to any rectangular ($r \times c$) χ^2 contingency table. Like Pearson's r^2, ϕ measures the degree to which two discrete variables correlate, and ranges from 0 to 1.

The way this measure works is like this. A 2 × 2 table where all cells are equal, or equal in either rows or columns, obtains $\phi = 0$. We might say that the two variables do not interact at all, irrespective of the amount of data that might be available. On the other hand, a table where the cells form the identity matrix [[1, 0], [0, 1]] (or some multiple thereof), returns a value of $\phi = 1$. In this case the value of one variable *determines* the value of the other. Although ϕ is relatively unknown, it deserves to be more widely used, because it is a better indicator for comparing experiments than $d^{\%}$.

Cramér's formula may be written as follows.

(E3) $$\text{Cramér's } \phi = \sqrt{\frac{\chi^2}{(k-1)N}}$$

where N is the total number of cases in the table and, for an arbitrary $r \times c$ table (rows × columns), k is the smaller of r and c.

This formula[30] also neatly summarises the relationship between χ^2 and ϕ: ϕ measures the size of the effect of one variable on another, χ^2 tells us whether this effect size is large enough, given the quantity of supporting data, N, to be significant.

This measure may be used for any $r \times c$ test, such as a 2 × 2 test. It differs from percentage swing because it averages change over all four cells in the table. Percentage swing ($d^{\%}$) mirrors the 'goodness of fit' approach by concentrating on just one column (cf. progressive or *shall*). Wallis (2010) summarises a method for calculating an equivalent goodness of fit ϕ', which measures change for one column only (see also Bowie, Wallis and Aarts, this volume). Wallis' ϕ' can be interpreted as a measure of the degree to which a variable for one term in a set differs from the variable applied over the entire set.

References

Aarts, B., J. Close and S.A. Wallis. 2010. Recent changes in the use of the progressive construction in English. In: B. Cappelle and N. Wada (eds.) *Distinctions in English grammar, offered to Renaat Declerck*. Tokyo: Kaitakusha. 148-167.

[30] A different signed formula for 2 × 2 tables (Sheskin 1997: 244) records the opposing direction of change as a negative value, but obscures this relationship.

Aarts, B., Wallis, S. A. and Bowie, J. (forthcoming). Profiling the English verb phrase over time: Modal patterns. In: Juana I. Marín Arrese and Johan van der Auwera (eds.) *Current Issues on Evidentiality and Modality in English: Theoretical, Descriptive and Contrastive Studies*.

Barber, C. 1964. *Linguistic Change in Present-Day English*. Edinburgh and London: Oliver and Boyd.

Bauer, L. 1994. *Watching English Change: An Introduction to the Study of Linguistic Change in Standard Englishes in the Twentieth Century*. London: Longman.

Bowie, J., Wallis, S. A. and Aarts, B. (forthcoming). The Modals in Present-Day Spoken English: Changing Usage across Text Types. To appear in Johan van der Auwera and Juana I. Marin Arrese (eds.) Current issues on Evidentiality and Modality in English: Theoretical, Descriptive and Contrastive Studies.

Close, J. and B. Aarts. 2010. Current change in the modal system of English: a case study of *must*, *have to* and *have got to*. In: Ursula Lenker, Judith Huber, and Robert Mailhammer (eds), *The History of English Verbal and Nominal Constructions.* Volume 1 of *English Historical Linguistics 2008: Selected Papers from theFifteenth International Conference on English Historical Linguistics (ICEHL 15), Munich 24-30 August 2008.* Amsterdam: John Benjamins. 165-181.

Coates, J. 1983. *The Semantics of the Modal Auxiliaries.* London: Croom Helm.

Collins, P. 2009. *Modals and Quasi-Modals in English*. Language and Computers: Studies in Practical Linguistics 67. Amsterdam: Rodopi.

Denison, D. 1998. Syntax. In: S. Romaine (ed.). *The Cambridge History of the English Language*. IV: 1776-1997. Cambridge: Cambridge University Press. 92-329.

Hopper, P. J. and E. Closs Traugott. 2003. *Grammaticalization.* Second edition. Cambridge: Cambridge University Press.

Huber, M. 2007. The Old Bailey Proceedings, 1674-1834: evaluating and annotating a corpus of 18th- and 19th-century spoken English. In: A. Meurman-Solin and A. Nurmi (eds.) *Annotating Variation and Change* (Studies in Variation, Contacts and Change in English 1).

Kytö, M. and J. Culpeper (eds.). 2010. *Speech in Writing: Explorations in Early Modern English Dialogues*. Cambridge: Cambridge University Press.

Labov, W. 1969. Contraction, deletion and inherent variability of the English copula. *Language* 45.4. 715-762.

Leech, G. 2003. Modality on the move: the English modal auxiliaries 1961-1992. In: R. Facchinetti, M. Krug and F. Palmer (eds.) *Modality in Contemporary English*. Berlin and New York: Mouton de Gruyter. 223–240.

Leech, G., M. Hundt, C. Mair, and N. Smith. 2009. *Change in Contemporary English: a Grammatical Study*. Cambridge: Cambridge University Press.

Mair, C. 2006. *Twentieth-Century English: History, Variation and Standardization*. Studies in English Language. Cambridge: Cambridge University Press.

Mair, C. and G. Leech. 2006. Current changes in English syntax. In: B. Aarts and A. McMahon *The Handbook of English Linguistics*. Malden MA: Blackwell Publishers. 318-342.

Nelson, G., S. A. Wallis and B. Aarts. 2002. *Exploring Natural Language: working with the British Component of the International Corpus of English.*Varieties of English around the World series. Amsterdam: John Benjamins.

Newcombe, R. G. 1998a. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 17: 857-872.

Newcombe, R. G. 1998b. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* 17: 873-890.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985) *A Comprehensive Grammar of the English Language*. London: Longman.

Rayson, P. 2003. *Matrix: a statistical method and software tool for linguistic analysis through corpus comparison*. PhD thesis, Lancaster University.

Sheskin, D. J. 1997. *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, FL: CRC Press.

Smith, N. 2003. Changes in modals and semi-modals of strong obligation and epistemic necessity in recent British English. In: R. Facchinetti, M. Krug and F. Palmer (eds.) *Modality in Contemporary English*. Berlin and New York: Mouton de Gruyter. 241–266.

Smith, N. 2005. *A corpus-based investigation of recent change in the use of the progressive in British English*. PhD, Lancaster.

Smitterberg, E. 2005. *The Progressive in 19th-century English: a Process of Integration.* (Language and Computers: Studies in Practical Linguistics 54.) Amsterdam: Rodopi.

Wallis, S.A. 2003. Completing parsed corpora: from correction to evolution. In: A. Abeillé (ed.) *Treebanks: building and using Syntactically Annotated Corpora*. Boston: Kluwer. 61-71.

Wallis, S.A. 2009. *Binomial distributions, probability and Wilson's confidence interval*. London: Survey of English Usage, UCL. www.ucl.ac.uk/english-usage/statspapers/binomialpoisson.pdf

Wallis, S.A. 2010. z-*squared: the origin and use of* χ^2. London: Survey of English Usage, UCL. www.ucl.ac.uk/english-usage/statspapers/z-squared.pdf

Wilson, E. B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 22. 209-212.