Recent grammatical change in English: data, description, theory

Geoffrey Leech

Lancaster University

Abstract

This chapter begins by considering the contrast between the data-driven paradigm characteristic of corpus linguistics and the theory-oriented paradigm characteristic of some other schools of linguistics, particularly those espousing a generative framework. To illustrate the corpus linguistics paradigm in detail, I present a case study of grammatical differences observed in the LOB and FLOB corpora and also other corpora of the early 1960s and the early 1990s. By abductive or inductive inference from the observed data, (fallible) descriptive generalizations can be made, and tentative conclusions of theoretical interest can be drawn. In conclusion, I argue that corpus linguistics is not purely observational or descriptive in its goals, but also has theoretical implications. However, like a theory-driven inquiry in the classic formulation of Popper's hypothetico-deductive method (1972: 297), a corpus linguistic investigation can only lay claim to provisional truths, and therefore requires confirmation or refutation by further research findings.

Table 1. Summary of the contents of this article

A.	Metatheoretical preamble
B.	Case study:
	Recent grammatical changes in (mainly) written (mainly) British English – *viz.* frequency changes between 1961 and 1991-2:
	(a) modal auxiliaries and semi-modals
	(b) other grammatical categories relating to colloquialization
C.	Conclusions

1. Introduction

In the 1960s, one of the widely-accepted fundamentals of linguistics was to be found in Chomsky's hierarchy of three levels of adequacy (1964: 62-3):

(1) *Explanatory adequacy* is achieved when the associated linguistic theory provides a general basis for selecting a grammar that achieves [descriptive] adequacy over others that do not.

> *Descriptive adequacy* is achieved when the grammar gives a correct account of the linguistic intuition of the native speaker, and presents the observed data (in particular) in terms of significant generalisations that express the underlying regularities of the language.
>
> *Observational adequacy* is achieved if the grammar presents the observed data correctly.

One of the implications of this formulation was a downgrading of the importance of empirical observation: as Chomsky himself pointed out, observation adequacy could be achieved by a mere listing of the data. Another implication, as I saw it, was a confusion between two notions of 'intuition': Chomsky's concept of descriptive adequacy confused the knowledge of the language of a native speaker with the analytic knowledge or expertise of the linguistic scientist, able to make significant generalizations about the language. In Leech (1968) I argued this case, and suggested a different hierarchy of three levels, which would be a more realistic account of the main strata of investigation in linguistics:

(2) *Theory*: formal [and functional] characterization or explanation of language as a phenomenon of the human mind and of society.

> *Description*: formal [and functional] characterization of a given language, in terms of the theory.
>
> *Data collection*: collection of observations which a description, and ultimately a theory, has to account for [e.g. corpora]

Since that time, the more empiricist and more rationalist trends in linguistics have diverged so far as to be almost irreconcilable. However, I still find the formulation in (2) useful, although I would now prefer to insert the words in square brackets '[and functional]', showing my preference for a combination of formal and functional explanation which corpus linguistics is characteristically attracted to. The other words in brackets – '[e.g. corpora]' – are of course a reminder that corpus linguistics finds its *raison d'être* at the observational or data-collection stratum of these three, the one that Chomsky found to be of such little importance. However, my overarching goal in the present chapter is to explore the relation between these three interrelated levels, and to argue against the common assumption that corpus linguistics is concerned with 'mere data collection' or 'mere description'.

2. A case-study: recent changes in English grammar

Alongside this, I also have a more practical goal, which is to exhibit as a case study a particular area of linguistic description: recent quantitative change in English grammar, as observed through the comparison of the LOB and FLOB corpora. Although the main study has been focused on the LOB and FLOB corpora, and therefore on written British English, it has been supplemented where practicable by work on other corpora permitting a similar comparison between English in the early 1960s and in the early 1990s. I will use this case study as a means of illustrating the relation between the three levels of theory, description and data collection – or, to put them in the order which would more naturally occur to a corpus linguist – data collection, description and theory.

2.1 Data collection: using the LOB, FLOB, and other corpora

To begin with the level of observation: we began with a study of the two matching corpora LOB and FLOB, which had already been part-of-speech tagged, through the combined processing of two taggers: CLAWS4 and Template Tagger (see Smith 1997 on the tagging techniques).[1] By using the powerful annotation-aware search and retrieval tool Xkwic (Christ 1994), we found it possible to extract occurrences of a whole range of grammatical categories that have been suspected, with varying degrees of empirical backing, to have become more frequent or less frequent in the recent past. The main areas of grammar we focus on in this chapter are (a) the modal auxiliaries, together with the mixed array of verbal constructions conveniently termed 'semi-modals', and (b) a range of grammatical phenomena associated with a suspected trend of 'colloquialization'.[2]

Although we began with the LOB and FLOB corpora, we extended our study to a selective use of some other comparable corpora spanning approximately the same period of 30 years, as shown in Table 2.

The family of four matching corpora Brown, LOB, Frown and FLOB (henceforward termed 'the Brown family') is well placed to provide evidence of frequency changes in British and American English over the period between 1961 and 1991-2. Unfortunately no comparable corpora for spoken English exist, but we were reluctant to confine our attention to written (printed) language, especially considering that much grammatical innovation is likely to originate in the spoken language. With the permission and help of Bas Aarts and Gerry Nelson at University College London, we were able to identify small comparable spoken subsets from two other million-word corpora developed at UCL with data from around the early 1960s and the early 1990s.[3] These were the corpus of the Survey of English Usage (SEU), of which a large spoken part was computerized and distributed as the London-Lund Corpus, and the International Corpus of English (the British variant known as ICE-GB). Because of difficulties of matching samples, the spoken 'mini-corpora' from SEU and ICE-GB were even smaller, indeed much smaller, and were moreover less closely matched than the

Table 2. The corpora of English used in the study

Name of corpus	***American or British English***	***Date of data collected***	***Spoken or written***	***Corpus size and design***
LOB Corpus	BrE	1961	Written	Each corpus contains approx. a million words, in 500 text samples from 15 different genres. The four corpora are built according to the same design and sampling method.
Brown Corpus	AmE	1961	Written	
FLOB Corpus	BrE	1991	Written	
Frown Corpus	AmE	1992	Written	
SEU-mini-sp	BrE	1959-1965	Spoken	Each (sub)corpus contains approx. 80,000 words from a comparable and balanced range of spoken genres.
ICE-GB-mini-sp	BrE	1990-1992	Spoken	

Brown family of corpora. One difficulty was that, although the SEU corpus had been collected over a period of about 30 years, comparability with LOB and Brown dictated that we rejected any material not contemporaneous with the written corpora, a constraint we interpreted rather liberally to exclude any material outside the time frame 1959-1965. Another problem was that the SEU corpus was subdivided into texts of 5000 words each, whereas the ICE-GB texts were of 2000 words each. Hence a one-by-one matching of texts between the two spoken mini-corpora was not feasible, and partial and overlapping matchings had to be allowed.

Because of these drawbacks, particularly the restriction of the mini-corpora of speech to a mere 80,000 words each, our findings from the spoken corpora could only be seen as highly tentative indicators of what was happening to spoken English over this period. Nevertheless, we felt that such a study, however inadequate and provisional, would be preferable to a survey of recent grammatical change which took no account of the spoken language. In fact, differences observed between the mini-corpora in the frequency of modals and semi-modals were tantalizingly even greater than those observed between LOB and FLOB. A summary of the contents of the two spoken mini-corpora is given in Table 3.

The sophisticated ICECUP software available for searching the ICE-GB could not be used with SEU-mini-sp, and so to ensure comparability we decided to use the WordSmith retrieval package and XKwic for both mini-corpora.

Table 3. Mini-corpora for studying language change in recent British spoken English

Name of corpus:	Survey of English Usage spoken 'Mini-corpus'	International Corpus of English (Great Britain) spoken 'Mini-corpus'
Abbreviation:	SEU-mini-sp	ICE-GB-mini-sp
Period of texts:	1959-1965	1990-1992
Size:	80,000 words each	
Texts from these categories:	(in each corpus:) conversation, broadcast discussions, sports commentaries, other commentaries, broadcast news, broadcast talks	

This section of the chapter has been called 'Data collection', and under this heading we can bring together the basic evidence-providing tools of the corpus linguist's stock in trade. Obviously, these include the corpora used for this particular study, and the software used to extract the relevant grammatical phenomena – in this case the search and retrieval tools XKwic and WordSmith. Basic retrieval products such as concordances and frequency lists, especially when they incorporate the results of simple grammatical analysis such as POS tagging, might be considered to take us beyond mere data collection, and to bring us to the threshold of the descriptive level of analysis. However, the scale of abstraction represented by the three levels of data collection, description, and theory is best assumed to consist of many small steps, rather than three giant strides. I return to the matter of data collection versus description in 2.2 below.

Although so far my presentation of the three levels has worked from the bottom up, this is of course by no means inevitable in the methodology of corpus linguists. Some studies are problem-driven – where the need to investigate a particular theoretical or descriptive hypothesis may determine the collection or selection of a suitable corpus, and the selection of particular corpus data to be studied. But in the present case, the 'bottom-up' methodology prevailed. We did not start with a particular theoretical claim (say about the process of historical change) or a particular descriptive hypothesis (say about the English modals), although our study led to these. It was the existence of the LOB and FLOB corpora, and the particular equivalence relation between them (found also between Brown and Frown) which enticed us to follow the example already set by Hundt, Mair and others, and to use these corpora to investigate recent changes in grammar.[4]

2.2 Description: the modals and semi-modals

The descriptive level of linguistic investigation attempts to determine what can be truly said about some aspect or level of the language, in this case English grammar. On the face of it, an example of linguistic description is provided by Table 4, showing changes in the frequency of modal auxiliaries over the 30-year period as reflected by the paired corpora.[5] However, at this stage, statements are

being made about a particular set of corpora, rather than about the language that they exemplify. We could call this level of statement 'data description': an intermediate step between data collection and linguistic description.

Table 4. Frequencies of modals in the four written corpora (including negative forms)

	British English		Log likhd	Diff %		American English		Log likhd	Diff %
	LOB	FLOB				Brown	Frown		
would	3028	2694	20.4	-11.0	*would*	3053	2868	5.6	-6.1
will	2798	2723	1.2	-2.7	*will*	2702	2402	17.3	-11.1
can	1997	2041	0.4	+2.2	*can*	2193	2160	0.2	-1.5
could	1740	1782	2.4	+2.4	*could*	1776	1655	4.1	-6.8
may	1333	1101	22.8	-17.4	*may*	1298	878	81.1	-32.4
should	1301	1147	10.1	-11.8	*should*	910	787	8.8	-13.5
must	1147	814	57.7	-29.0	*must*	1018	668	72.8	-34.4
might	777	660	9.9	-15.1	*might*	635	635	0.7	-4.5
shall	355	200	44.3	-43.7	*shall*	267	150	33.1	-43.8
ought	104	58	13.4	-44.2	*ought*	70	49	3.7	-30.0
need	78	44	9.8	-43.6	*need*	40	35	0.3	-12.5
Total	14667	13272	73.6	-9.5	Total	13962	12287	68.0	-12.2

In this chapter we will be almost entirely concerned with description in terms of relative frequency, or relative likelihood, of occurrence.[6] Table 4 records the frequency of each modal auxiliary of the 'canonical' set of modals in each of the Brown family of corpora. In the absence of other explanations (such as the corpora being importantly different in other ways than in the dates of their composition) we can tentatively conclude that these differences reflect different states of the language: that between 1961 and 1991, the modals declined very significantly in frequency in written English in both American and British usage. (The overall percentage losses are –9.5% in BrE and –12.2% in AmE). The fourth and ninth columns in Table 4 tell us how much the frequencies of the modals have declined, as a percentage of the 1961 figures. The fifth and tenth columns provide a second measure of the degree of decline, this time using the *log likelihood ratio* (G^2) as a measure of significance (Dunning 1993). In these columns, any score of 3.8% or over is calculated to be significant at the chi-square level of $p < 0.05$, and any score of 6.6% or over is significant at the level of $p < 0.01$. The larger the log likelihood ratio, the greater the significance.

The individual modals show a decline varying between *can* (which actually increases its frequency in FLOB, and declines only 1.5% in Frown) and *shall* (which declines over 40% in both FLOB and Frown). In Table 4, the modals are

listed in order of frequency in LOB, and exactly the same order of frequency, with the exception of *should* and *must*, applies to Brown. It will also be seen that a roughly similar pattern of falling frequency is observed in both BrE and AmE corpora. Broadly, the most frequent modals decline least, and the least frequent modals decline most in percentage terms, the rare modals *shall, ought* (*to*) and *need* (+ bare infinitive) having become much rarer. Some middle-order modals (especially *must* and *may*) also show very significant falls in frequency.

The most interesting observation from Table 4, however, is that the overall frequency of modals is highest in LOB and lowest in Frown, with FLOB and Brown in intermediate positions. Alongside the decline between 1961 and 1991-2, there is an equally important difference between AmE and BrE, which invites interpretation as a time lag. It is as if BrE is following rather reluctantly in the wake of a change in AmE, with something like a generation gap. This is shown graphically (though not strictly to scale) in Figure 1.

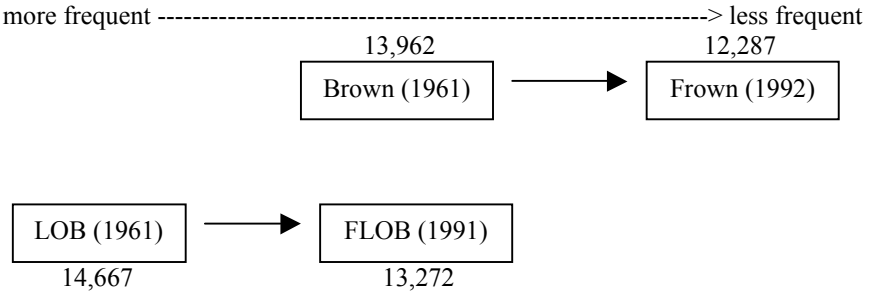

Figure 1: British English following an apparent American English trend

It might be proposed that the apparent decline in modal usage is due to the rise, in recent centuries, of the so-called semi-modals, such as *be going to* and *have to*, which are presumed to be still increasingly used. Perhaps these are gradually encroaching on the territory of the canonical modals. Such a hypothesis can be tested, up to a point, by noting the differences of frequency of semi-modals in the four corpora, as shown in Table 5. Although the class of semi-modals is not a well-defined set, those in Table 5 may be taken as fairly representative.

Ostensibly, there is no strong connection between the patterns shown by the modals and the semi-modals.[7] Altogether, the semi-modals are very much less frequent (in written English) than the modals, and their changes in frequency show a mixed picture. Some of them seem to have increased their usage massively in the period 1961-1991/2, but others have declined. One of the differences at first glance lending credence to the encroachment hypothesis is that AmE shows a greater increase in the semi-modals (+18.6%) in comparison with BrE (+10.0%) – a mirror image of what is happening with the modals. Unexpectedly, however, the overall frequency of semi-modals is found to be greater in the BrE than in the AmE corpora in both periods!

Table 5. Frequencies of some semi-modals in the four written corpora

BrE	LOB	FLOB	Log likhd	Diff (%)	AmE	Brown	Frown	Log likhd	Diff (%)
*BE going to**	248	245	0.0	-1.2	*BE going to**	219	332	23.5	+51.6
BE to	454	376	7.6	-17.2	*BE to*	349	209	35.3	-40.1
(*had*) *better*	50	37	2.0	-26.0	(*had*) *better*	41	34	0.7	-17.1
(*HAVE*) *got to**	41	27	2.9	-34.1	(*HAVE*) *got to**	45	52	0.5	+15.6
HAVE to	757	825	2.7	+9.0	*HAVE to*	627	643	0.1	+1.1
NEED to	54	198	83.0	+249.1	*NEED to*	69	154	33.3	+123.2
BE supposed to	22	47	9.2	+113.6	*BE supposed to*	48	51	0.1	+6.3
used to	86	97	0.6	+12.8	*used to*	51	74	4.3	+45.1
*WANT to**	357	423	5.4	+18.5	*WANT to**	323	552	60.9	+5.2
TOTAL	2069	2275	9.2	+10.0	**TOTAL**	1772	2101	28.4	+18.6

*Forms spelt *gonna, gotta* and *wanna* are counted under *be going to, have got to,* and *want to* respectively

Table 6. Comparison of SEU-mini-sp and ICE-GB-mini-sp: modals in spoken BrE (provisional figures)

	SEU-mini-sp	**ICE-GB-mini-sp**	**Log likhd**	**Difference (%)**
would	415 (5188)	271 (3388)	30.5	-34.7
will	248 (3100)	307 (3838)	6.3	+23.8
can	252 (3150)	295 (3688)	3.4	+17.1
could	145 (1813)	83 (1038)	17.1	-42.8
may	86 (1075)	36 (450)	17.5	-54.1
should	100 (1250)	84 (1050)	1.6	-17.3
must	87 (1088)	35 (438)	24.3	-60.7
might	56 (700)	50 (625)	0.3	-10.7
shall	26 (325)	17 (213)	1.9	-34.6
ought	20 (250)	9 (113)	4.3	-55.0
need	0 (0)	0 (0)	0.0	0.0
Total	1435 (17938)	1187 (14838)	23.5	-17.3

Note: The figures in parenthesis show frequency per million words, and are therefore comparable to the figures for the written corpora given in Table 4.

At this point, it is an attractive idea to look at the patterns of change observable in the spoken mini-corpora, small as they are. Surely the innovatively increasing use of semi-modals, and perhaps the corresponding fall in modals, are likely to show up far more in the spoken language than in the written. The differences in frequency (in spoken BrE only) between SEU-mini and ICE-GB-mini are shown in Table 6.

In general, the patterns of frequency shown in Table 6 suggest that trends in spoken English are similar to those in written English, but somewhat more exaggerated. The modals are more frequent in the written than in the spoken corpora, for both periods, but the decline in frequency is also greater – a loss of 17.3%. *May* and *must* are particular heavy losers, whereas *will* and *can*, in contrast, show a surprising increase from the 1961 to the 1991 corpus. This picture may be contrasted with the apparent considerable increase in semi-modals in spoken English between the early sixties and the early nineties, as observed in the spoken corpora, and as shown in Table 7.

Table 7. Comparison of SEU-mini and ICE-GB-mini: some 'semi-modals' in spoken BrE

	SEU-mini-sp	**ICE-GB-mini-sp**	**Log likhd**	**Difference (%)**
(BE) going to	88	120	4.9	+36.4
BE to	5	10	1.7	+100.0
(HAVE) got to	35	26	1.3	-25.7
HAVE to	79	104	3.4	+31.6
NEED to	2	15	11.3	+650.0
BE supposed to	8	12	0.8	+50.0
Total	217	287	9.8	+32.3

These numbers, of course, are ridiculously small – only three can be counted as significant in log likelihood terms. However, overall they suggest, as many would suspect, that the general increase of semi-modals is even greater in spoken than in written English.

2.3 Descriptive conclusions and further discussion on modals and semi-modals

The following overall findings can be presented by way of summary of the preceding section on modals and semi-modals. *On the basis of the evidence from the corpora*:

(i) In general terms, there is clearly an appreciable decline of frequency in the use of modal auxiliaries between 1961 and 1991-2.

(ii) During this period, individual modals have been declining at different rates, but there is a tendency for very common modals to hold their own (e.g. *will, can*), and for infrequent modals (e.g. *shall, ought to, need*) to decline sharply and to appear almost moribund. Some middle-ranking modals (e.g. *may* and *must*) have also declined sharply.

(iii) Alongside the decline of modals, there is no clear overall picture regarding semi-modals: although in general, semi-modal usage is increasing, some semi-modals are declining, and semi-modals as a whole are much less frequent than 'true' modals.

If we ignore the italicised phrase above ('*On the basis of the evidence from the corpora*') these statements are descriptive: they claim to tell us something that is true about the language, English. But rather than accept them uncritically, we have to bear in mind some hazardous assumptions which can be made in moving from data description to language description:

Hazardous Assumptions: from Data Description to Language Description

1. That the corpora are large enough and varied/balanced enough to allow us to extrapolate from corpus findings to what is happening in (relevant varieties of) the language in general.
2. That the corpora are sufficiently comparable in terms of samples of the varieties represented, and in using the same sampling methods.
3. That statistically significant results can be attributed to real linguistic differences, rather than to extraneous factors such as cultural shifts or faulty sampling.
4. That the grammatical categories are defined and used in a way that other grammarians or linguists find reasonable.
5. That the extraction of data from the corpora has been acceptably (if not totally) free from error.

The first of these assumptions – the well-known issue of representativeness – is perhaps the biggest hazard. In the lack of any practical, general measure of representativeness,[9] the statements (i)-(iii) must be regarded as hypotheses, well evidenced, it is true, but needing to be supported by further corpus studies as and when opportunities arise. The second assumption underlies the whole enterprise of comparing the Brown family of corpora. The third raises the thorny question of how to relate statistical significance to certain causative factors. For example, we might attempt to explain changes in the direction of spoken style as part of a general socially-driven trend of colloquialization (see sections 2.4 and 3) when it is possible that these changes can be more directly explained by an increase in the amount of quoted speech included in the 1991-2 corpora (see below). The fourth hazardous assumption reminds us that linguistic categories – even 'consensual' ones like modal auxiliary verb, are not God's truth but capable of being challenged. The fifth has already been discussed in Note 4.

In my view, none of these hazards justifies a response of extreme scepticism which says if one cannot prove the truth of these descriptions, one should not make them at all. Rather, they lead to the recognition that such results should be regarded as provisional and that there is a need to seek further corroborating evidence as well as means of increasing accuracy and reliability. This 'striving for perfection' can be a slow, gradual and time-consuming process, which might include further manual checking or even collecting and analysing fresh corpora.

It is, though, reassuring to bear in mind that even if an objection is raised to a hazardous assumption, this often fails to undermine the results in more than a minor way. For example, the discovery of occasional errors or differences of categorization in identifying modals is unlikely to cause more than a minor change in the frequency counts in Table 6, and hence in the statistical significance of the results. Thus if someone insists that *ought to* is not a modal but a semi-modal, this would change the overall findings only marginally. Or to take another example, on checking the examples of *may* in Frown, I found two examples of non-modal *may* lurking in the database, thus reducing the count of the modal *may* from 878 to 876: this makes almost no difference to the significance of the decline, and in fact increases it.

Returning to the colloquialization trend mentioned above (and to be taken up again in 3 below), the claim that this phenomenon is an illusion because of an increase in quoted speech in the later corpora can be checked by actually undertaking a measurement of quoted material in LOB and FLOB. This has been done by Nick Smith for LOB and FLOB, and shows that there is an increase of c. 9.5% in the incidence of quoted material in FLOB as compared with LOB.[10] However, this could account only in part for most of the changes that might be attributed to colloquialization (see Table 8 below), so there still remains something linguistically interesting to be explained here.

To counterbalance the 'hazardous assumptions', observations such as the following can have a compensating effect in increasing the plausibility if not authority of the results, and suggesting that they are not just a matter of chance or accident:

(a) Many results are highly significant as measured by log likelihood ratio.
(b) Trends are consistent across different items – e.g. the general frequency decline of the modals is replicated in almost every single modal auxiliary.
(c) Trends are often consistent across different subcorpora – e.g. if we subdivide each of the Brown family into genre categories Press (A-C), General Prose (D-H), Learned (J), and Fiction (K-R), often similar trends are observed in all these four subcorpora. An instance of this is the decline of the passive from LOB to FLOB (see Table 8). The passive is less frequent in FLOB as a whole by –12.4%, a trend repeated in a similar way for each subcorpus: Press –12.5%; Gen Prose –12.4%; Learned –16.6%; Fiction –3.6%).

I find it useful to use an analogy of scaffolding in the confirmation and extension of descriptive findings. If we think of the corpus-based methodology as the constructing of a building by erection of scaffolding, the superstructure of description of a language can be supported in three ways:

(i) Data observation: from below, struts or buttresses can be used to strengthen the grounding of data description (e.g. seeking confirmation from new data).
(ii) Description: at the same descriptive level, findings can be extended and deepened. For example, we can probe into the crude frequency changes of modals in Table 4 by analysing subcorpora as already noted in (c) above, or by undertaking a semantic analysis of examples. This we did for *may, must* and *should*, and noted a trend in *may* and *should* towards monosemy – viz. the dominant senses of *may* (epistemic) and *should* (deontic) increased their dominance in spite of loss of frequency (see Leech, 2003). *Must*, on the other hand, showed a decline of both its major senses, the epistemic and deontic meanings. Such further descriptive investigations help to pinpoint what is happening more precisely, in terms of how and where the modals are becoming less used.
(iii) Theory: pointing 'up' to the theoretical level, further descriptive investigations, for example by taking contextual factors into account, can help to identify appropriate theoretical explanations as to *why* the modals are declining. This is where broad explanatory concepts such as colloquialization come into play, and help to direct investigation into particular channels.

2.4 Continuing the case study: grammatical changes relating to colloquialization

Taking further the descriptive study of the LOB and FLOB corpora, we now turn to a wider-ranging set of grammatical categories, mostly belonging either to the verb phrase or to the noun phrase. What brings all these categories together is that they can all be associated with a trend towards colloquialization, that is *a tendency for the written language gradually to acquire norms and characteristics associated with the spoken conversational language.* Quantitatively, colloquialization can be shown in two ways: (a) by an increasing frequency of phenomena associated with spoken language, and (b) by a decreasing frequency of phenomena associated with the written language. Type (a) changes predominate in Table 8 below, but Type (b) changes are also seen, in the decreasing frequency of the passive, of the *of*-construction, and of the relative pied-piping construction.

Table 8. Changes apparently indicative of colloquialization (tokens per million words)

	LOB	FLOB	Log lkhd	Difference (%)
Categories within the verb phrase				
a. Present progressive (active)	980	1263	36.0	+28.9
b. Progressive passive	198	260	8.4	+31.3
c. Verb contractions (e.g. *it's*)	3126	3867	79.1	+23.7
d. Negative contractions (*-n't*)	1940	2462	62.6	+26.9
e. Passive forms (all)	13260	11614	109.8	-12.4
Miscellaneous 'colloquialization features' outside the verb phrase				
f. Questions (all)	2572	2816	11.1	+9.5
g. Verbless questions	310	424	17.7	+36.6
h. Tag questions	63	65	0.1	+4.5
j. Genitives	4935	6122	128.5	+24.1
k. *Of*-phrases	33715	32139	37.9	-4.7
l. *Of*-phrases competing with the genitive (2% sample only)	124	95	3.9	-23.6
Relative clauses				
m. *Wh*-relative pronouns	6971	6376	26.7	-8.5
n. Zero relative with stranding (sample)	18	73	36.4	+310.0
p. Pied-piping relatives	1394	1158	21.9	-16.9

Of the categories within the verb phrase, the first four (a.-d.) all show very convincing increases between LOB and FLOB. Previous corpus studies (e.g. Biber et al. 1999: 461-463) have shown the progressive to be more common in conversation than in written genres, and this is a justification for treating colloquialization as a possible explanation for a. and b. (However, the growing use of the progressive aspect can also be linked with grammaticalization, going back over 500 years.) The passive (e.), on the other hand, is strongly associated with the written medium (see for example Biber et al. 1999: 476-477), and so its decline in frequency can count as a negative manifestation of colloquialization.

The next set of categories in Table 8 (f.-h.) is more mixed. In fact f. and h. (questions) should arguably be excluded from the list of colloquialization phenomena, as the increase of quoted speech in FLOB compared with LOB (see Note 9) provides a readier explanation for the increasing occurrence of questions (+9.5%) and tag questions (+4.5%).

We have begun to investigate two further colloquialization themes in the noun phrase (see j.-p. in Table 8): the *s*-genitive vs. the *of*-phrase; and zero or *that*-relative clauses vs. *wh*- relative clauses. Results so far point in the direction of (a) a rise in the genitive with a corresponding decline in *of*-phrases; and (b) a rise in zero relative clauses ending with a stranded preposition and a corresponding decline in *wh*- relative clauses. The rise in stranding accords with an unsurprising and significant fall in the use of pied-piping constructions in which the *wh*-relative pronoun is preceded by a preposition (*in which, of whom*, etc.).

Summary of descriptive conclusions relating to colloquialization

(a) The use of the present progressive construction has increased overall by c. 30% between LOB and FLOB. This seems part and parcel of the spread of the progressive aspect usage over the past 500 years.

(b) In practice, this increase has been chiefly in the present progressive – the past progressive has actually shown a slight decline.

(c) As part of a general 'colloquialization' trend, the use of negative and verb contractions has increased by approximately a quarter (25%). Part of this, though, can be attributed to the increase in the proportion of quoted speech in the written corpora.

(d) Conversely there has been an appreciable decline in the use of the passive – a verbal category strongly associated with formal written language.

(e) The written corpora show an increase in 9.5% in the use of questions.

(f) This actually increases to approximately 36.6% if we confine our attention to questions which lack a finite verb – this 'fragmentary interrogative' type is particularly strongly associated with conversational English (see Biber et al. 1999: 211-212). Tag questions, on the other hand, have not increased much. Perhaps this is because they are essentially dialogic in a way that other questions are not. (In Biber et al. *ibid*, tag questions are shown to be of particularly low frequency in the written language.)

(g) In the noun phrase, historically, the competition between *'s* genitives and *of*-constructions has been interpreted as a competition between more and less oral styles of expression.[11] Genitives have increased by about 25% from LOB to FLOB, whereas *of*-phrases have declined by about 5%. However, if we confine our attention to *of*-phrases which could be replaced semantically by genitives, the decline of the *of*-construction (based on a 2% sample) goes up to 24%. This intriguing provisional result, which almost exactly balances the gain in the genitive, needs further corroboration with a larger sample.

(h) There is a general tendency for *wh*-relative clauses to decline. This applies not only to *whom* but also to *who, whose,* and *which*. The decline is not unexpectedly magnified if we confine our attention to pied-piping relatives (beginning with a preposition – e.g. *of which, to whom*).

(j) Conversely, there appears to have been an increase in the use of zero relatives, i.e. relative clauses with a zero relativizer (*the book I read*) especially when combined with a stranded final preposition *(someone I spoke to*). This is a provisional finding based on a small sample, and again needs further research.

As a conclusion to the descriptive sections of this chapter, I reiterate two caveats already mentioned. First, the results presented are provisional (particularly those based on a small sample, such as (j) above) since the research presented here is

still work in progress. (In fact I have gone so far as to suggest that it is in the nature of corpus research to be provisional.) Second, the 'hazardous assumptions' listed in section 2.3 have to be kept in mind throughout, and opportunities found to probe them further. I have yielded above to the temptation to talk in terms of the language change between LOB and FLOB: a kind of dynamic metaphor used to explain what are actually sets of synchronic observations about a 1961 corpus and a 1991 corpus. But the claims that these observations represent changes in the (use of the) language ultimately remain hypotheses, in need of further probing and confirmation.

3. Back to theory: conclusions

There is a great deal more to be done in terms of short term diachronic investigation of the Brown family of corpora. Once the gross frequency changes have been plotted, the next step is to investigate factors internal to the corpora that might help to explain these changes (e.g. differential results in different subsections of the corpus). Much more research also needs to be done – and some is being done – on the changing frequency of semantic categories such as epistemic modals and 'pragmatic' uses of the progressive. We are also making further comparisons between the British corpora and their American counterparts Brown and Frown. And of course, there is room for much more work on spoken language – the spoken mini-corpora used for this study are likely to reflect more fascinating indications of language change, but are obviously of inadequate size.

Explaining the changes in a deeper sense means finding historical reasons – investigating both language-internal and language-external (especially socially motivated) explanations of why these changes of frequency are taking place. In part the changes noted – e.g. in the increase of semi-modal use – may be related to well-known grammaticallization processes:

Grammaticalization – 'the process whereby lexical items and constructions come in certain linguistic contexts to serve grammatical functions, and, once grammaticalized, continue to develop new grammatical functions.' (Hopper and Traugott 1993: xv)

This is a linguistically-oriented explanation, invoking a whole theory of language change, applicable particularly to the growth of the semi-modals and the progressive aspect. But frequency studies such as the present one are less concerned with linguistic innovation than with diffusion and attenuation of aspects of language use, and invite social explanations in terms of such trends as:

Colloquialization – a tendency for features of the conversational spoken language to infiltrate and spread in the written language.

Democratization – speakers' and writers' tendency to avoid unequal and face-threatening modes of interaction (this may account in part for the decline of

deontic *must* and the rise of deontic *should, have to* and *need to*). For this kind of explanation in the realm of modality, see Myhill (1995).

Americanization – the influence of north American habits of expression and behaviour on the UK (and other nations). This shows up apparently in the loss of frequency of the modals, as depicted in Figure 1.[12]

However, these 'izations' manifest themselves patchily. For example, in contrast to the Americanization effect noted with the decline of modals, the growth of the present progressive shows very little difference between AmE (in the Brown and Frown corpora) and BrE, as demonstrated in Table 9.

Table 9. Comparison of increase of present progressive in LOB-FLOB and Brown-Frown (active only)

	1961 corpora	1991-2 corpora	Log likelihood	Difference
British	980 (LOB)	1263 (FLOB)	36.0	+28.9%
American	996 (Brown)	1316 (Frown)	43.6	+31.8%

So Americanization can be only tentatively invoked here, although it might be applied to other changes touched on earlier, such as the decline of the relative pronoun *which*. Another example of patchiness is the virtual stasis of the *get*-passive in LOB and FLOB (101 instances in LOB; 104 in FLOB): this obviously colloquial construction does not seem to follow the pattern observed elsewhere.

One explanation for the selectivity of these 'ization' trends is that the trends can be in conflict with one another. What happens, for example, to a formal (uncolloquial) construction characteristic of AmE? Does it increase in BrE because of American influence, or does it decline in BrE because of its negative association with colloquialization? An apparent example of this kind of conflict is the mandative subjunctive as in:

the Secretary of Labor requires that he <u>be</u> willing to risk his reputation
(Example from the Brown Corpus)

– a construction which (in a study by Serpollet 2001) increases from 14 in LOB to 33 in FLOB, while in AmE it is far more frequent, though declining: 91 in Brown and 78 in Frown. What seems to happen here is that in BrE, the Americanism of the construction outweighs its non-colloquialism. But different kinds of explanations might be applicable to other cases.

As we move from the level of description to that of explanation, it is appropriate to ask what kind or kinds of theory would be best able to explain the descriptive findings of corpus linguistics. Terms like *colloquialization* do represent some rather general attempt to explain change, but they do not amount to well-developed theories. As for grammaticalization, Croft (2000), like Krug (2000) is one of those who see grammaticalization taking place within a usage-based, communication-based, utterance-oriented theory of language change. Croft emphasises the important diachronic collaboration between innovation or

actuation – the creation of novel forms of language – and propagation or diffusion – the way the use of these forms expands into more general language use. The converse mechanisms of change – contraction and loss – also need to be given fuller consideration: we need a theory to explain the decline of the modals as well as the growth of the semi-modals.

In diachronic corpus comparisons we can observe the results of propagation and contraction. (It is unlikely that we will find true grammatical innovation – or that we would recognize it as such in a corpus even if we came across it.) This means that we need explanations which take full account of socio-cultural factors inducing language change. Croft argues (2000: 166) that the basic mechanism for propagation is the speaker's self-identification with a social group, and he cites in this connection a maxim put forward by Keller (1990/1994), 'Talk like others talk'. Here, the social-psychological theory of *accommodation* as a linguistic process comes into play.

This seems to place propagation of change firmly in the sphere of sociolinguistics, but it might be pointed out that the Brown family of corpora are not sociolinguistically sensitive in the normal sense: by definition, they contain published, i.e. public, language. So where does this leave the explanation of increase and decrease of frequency in the LOB and FLOB corpora? It is reasonable to suggest that the spread or shrinkage of linguistic usage in recent modern society has been influenced considerably by language use in the public media. So it can be helpful to complement the sociolinguistic perspective by perspectives oriented towards mass communication.

Table 10. Some principles of usage-based models of language (after Barlow and Kemmer 2000)

1.	The intimate relation between linguistic structures and instances of the use of language.
2.	The importance of frequency.
3.	Comprehension and production are integral, rather than peripheral to the language system.
4.	Focus on the role of learning and experience in language acquisition.
5.	Importance of usage data in theory construction and description.
6.	The intimate relation between usage, synchronic variation, and diachronic change.
7.	The interconnectedness of the linguistic system with non-linguistic cognitive systems.
8.	The crucial role of context in the operation of the linguistic system.

For example, with reference to colloquialization, Fairclough (1992) discusses 'the apparent democratization of discourse' in present-day English-speaking society. 'Conversational discourse,' he goes on, 'has been, and is being, projected from its primary domain into the public sphere (p.98). Social theories focusing on public discourse, like Fairclough's, here provide a valuable

supplement to the more established frameworks of historical linguistics and sociolinguistics. But there would be much benefit in investing in the support such theories may gain from the empirical findings of corpus research.

To conclude, I return to the opening theme of metatheory. Although I have not gone far towards suggesting theoretical solutions, I have worked my way around to suggesting the kind of theoretical approach that is better suited to corpus linguistics than is the Chomskyan paradigm. Corpus linguistics finds a good ally in the usage-based frameworks championed by Barlow and Kemmer (2000: viii-xxii), who, among other principles of this approach, list those in Table 10.

The usage-based conception of linguistics is not a monolithic theory, or a single school of thought, but is more like a confederation of linguists with similar goals, priorities and methods. Their tenets are the opposite of the generative paradigm in nearly every respect. Corpus linguistics finds a natural place in this body of linguists who believe that there is not a gulf, but on the contrary a natural bridge, between the study of naturally-occurring data and the cognitive and social workings of language.

Notes

1. In this chapter, *we* refers to Nicholas Smith and myself. I am grateful to Nick for much of the corpus processing, quantitative and analytic work that resulted in the findings reported here, as well as for discussion of broader issues and specific comments on this chapter. The project on Recent Grammatical Change in British English was supported by a research grant from the Arts and Humanities Research Board (UK) and a British Academy Larger Research Grant. In this research, we have benefited from collaboration with Christian Mair and Marianne Hundt at Freiburg University, to whom we owe support and inspiration, as well as the more practical benefit of the post-editing of most of the automatically-tagged FLOB corpus.
2. The colloquialization tendency for written style to drift towards more oral styles over time for some genres between and 17th and the 20th centuries is demonstrated statistically by Biber and Finegan (1989).
3. We are very grateful to Bas Aarts and Gerry Nelson for their help both in allowing use of these corpora, and extracting the data for the mini-corpora.
4. There has been a growing range of publications on the comparison of the LOB and FLOB corpora. Particularly relevant to the present study are Hundt (1997) and Mair (1997).
5. The findings on the modals in this chapter are presented and discussed more extensively in Leech (forthcoming 2003) and Smith (forthcoming 2003). Some of the counts in the tables are slightly different from those in these cited papers, owing to further research and further accuracy checks (see Note 6).

6. A caveat about frequency: most of the frequency figures in this study are *very close approximations* rather than guaranteed 100% accurate. Both manual procedures and automatic procedures can give rise to error, although the incidence of error is likely to be totally insignificant. The one exception to this is the margin of error arising from POS tagging (about 2% in the present context). Although we were able to use the results of manual correction for the LOB Corpus and most of the FLOB corpus, for the fictional genres (K-R) of FLOB and for the Frown Corpus we had to rely on automatic tagging only. A method of approximation was devised on the basis of comparing automatic tagging and manual tagging outcomes in cases where they were both available, and hence calculating an error coefficient for each tag. The procedure is described in the Appendix to Mair et al. (2003).
7. However, the decline of *must* may have some connection with the increase in use of *have to* and *need to* – see Smith (2003). In general, the varied behaviour of the semi-modals in this corpus confirm the impression that they comprise a miscellaneous category. In Quirk et al. (1985:136-148), where it is argued that they form a gradient between auxiliary and full verbs, four intermediate categories are distinguished: marginal modals, modal idioms, semi-auxiliaries, and catenative verbs.
8. However, the decline of *must* may have some connection with the increase in use of *have to* and *need to* – see Smith (forthcoming 2003). In general, the varied behaviour of the semi-modals in this corpus confirm the impression that they comprise a miscellaneous category. In Quirk et al. (1985:136-148), where it is argued that they form a gradient between auxiliary and full verbs, four intermediate categories are distinguished: marginal modals, modal idioms, semi-auxiliaries, and catenative verbs.
9. On representativeness, Biber (1993) is the classic reference; but Biber's position has also been criticised (e.g. by Váradi 2001). There is no test that could be used to ensure that statements about the LOB and FLOB corpora are representative of the varieties of English of which they are samples, except to collect independent samples of data of the same text types – in effect, to replicate the LOB and FLOB corpora but with different text samples.
10. Nick Smith has undertaken a count of quoted material in the LOB and FLOB corpora, helped by a program written by Izumi Tanaka. He found that the number of words within quotation marks in FLOB was c.127,000, compared with c.116,000 words in LOB – an increase of c. 9.5%. This figure of +9.5% is a reasonably close approximation, but needs to be followed up by further checks and edits.
11. Actually genitives are not so frequent in conversation as in some varieties of written English, especially news writing (see Biber et al. 1999: 302). This can be largely explained by the fact that nouns are notably infrequent in the spoken language: a construction which is 'rich in nouns' (a description that applies both to the genitive construction and the *of*-construction) is therefore comparatively rare. However, if we consider the likelihood of choosing a

genitive as contrasted with a semantically equivalent *of*-phrase, the odds in favour of the genitive are higher in spoken English than in a range of written varieties (see Leech et al. 1997).

12. Colloquialization and Americanization are discussed, with reference to the LOB and FLOB corpora, by Mair (1997, 1998). See also Hundt (1997).

References

Biber, D. (1993), 'Representativeness in corpus design', *Literary and Linguistic Computing* 8: 243-257.

Biber, D. and E. Finegan (1989), 'Drift and the evolution of English style: a history of three genres', *Language* 65.3: 487-517.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan (1999), *Longman grammar of spoken and written English*. London: Longman.

Barlow, M. and S. Kemmer (eds) (2000), *Usage-based models of language*. Stanford: CLSI.

Christ, O. (1994), 'A modular and flexible architecture for an integrated corpus query system', *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research (Budapest, July 7-10, 1994)*. Budapest, 23-32.

Croft, W. (2000), *Explaining language change: an evolutionary approach*. London: Longman.

Chomsky, N. (1964), 'Current issues in linguistic theory', in: J.A. Fodor and J.J.Katz, *The structure of language*. Englewood Cliffs, New Jersey, 50-118.

Dunning, T. (1993), 'Accurate methods for the statistics of surprise and coincidence', *Computational Linguistics* 19.1: 61-74.

Fairclough, N. (1992), *Discourse and social change*. Cambridge: Polity Press.

Hopper, P. and E. Traugott (1993), *Grammaticalization*. Cambridge: Cambridge University Press.

Hundt, M. (1997), 'Has BrE been catching up with AmE over the past 30 years?', in: M. Ljung (ed.), *Corpus-based studies in English: Papers from the 17th International Conference on English Language Research on Computerized Corpora (ICAME 17)*. Amsterdam, Rodopi, 135-151.

Keller, R. (1990/1994), *On language change: the invisible hand in language*. London: Routledge. (Translation and expansion of *Sprachwandel: von der unsichtbaren Hand in der Sprache*. Tübingen: Francke.)

Krug, M. (2000), *Emerging English modals: A corpus-based study of grammaticalization*. Berlin & New York: Mouton de Gruyter.

Leech, G. (1968), 'Some assumptions in the metatheory of linguistics', *Linguistics* 39: 87-102.

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

Leech, G., B. Francis and X. Xu (1997), 'The odds in favour of the genitive: a study of gradience in English', in: K. Yamanaka and T. Ohori, *The locus of meaning: Papers in honor of Yoshihiko Ikegami*. Tokyo: Kuroshio, 187-208.

Mair, C., M. Hundt, G. Leech and N. Smith (2002), 'Short term diachronic shifts in part-of-speech frequencies: a comparison of the tagged LOB and FLOB corpora', *International Journal of Corpus Linguistics*, 245-264.

Mair, C. (1997), 'Parallel corpora: a real-time approach to language change in progress', in: M. Ljung (ed.), *Corpus-based studies in English: Papers from the 17th International Conference on English Language Research on Computerized Corpora (ICAME 17)*. Amsterdam: Rodopi, 195-209.

Mair, C. (1998), 'Corpora and the study of the major varieties of English: issues and results,' in: H. Lindqvist et al. (eds), *The major varieties of English.* Växjö: Växjö University Press, 139-157.

Myhill, J. (1995), 'Change and continuity in the functions of the American English modals', *Linguistics* 33: 157-211.

Popper, K. (1972), *Objective knowledge* (revised edition). Oxford: Oxford University Press.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A comprehensive grammar of the English language.* London: Longman.

Rayson, P., A. Wilson, T. McEnery, A. Hardie and S. Khoja (eds) (2001), *Proceedings of the Corpus Linguistics 2001 Conference*. Lancaster University: UCREL Technical Papers 13.

Serpollet, N. (2001), 'The mandative subjunctive in British English seems to be alive and kicking… Is this due to the influence of American English?', in: Rayson et al. (2001), 531-542.

Smith, N. (1997), 'Improving a tagger', in: R. Garside, G. Leech and A. McEnery (eds), *Corpus annotation: Linguistic information from text corpora*. London: Longman, 137-150.

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

Váradi, T. (2001), 'The linguistic relevance of corpus linguistics', in: Rayson et al. (2001), 587-593.