Extending the possibilities of corpus-based research on English in the twentieth century: a prequel to LOB and FLOB

Geoffrey Leech
Nicholas Smith
Lancaster University

Abstract

This paper explains the rationale for a new corpus being assembled at Lancaster University to complement the existing Brown 'family' of corpora; that is, English language corpora modelled on the original Brown University corpus, such as LOB, Frown, FLOB, Wellington, etc. The purpose of the new corpus, called Lancaster1931, is to extend the chronological span of these corpora into the first half of the twentieth century, and so to afford researchers a stronger empirical basis for examining recent grammatical change in English. We discuss some methodological issues encountered in extending the Brown model to earlier historical periods. We also outline some developments under way to permit more rigorous computer-assisted analyses within and across these corpora, namely (i) encoding of all the corpora with XML, (ii) adoption of a common grammatical tagset, known as 'C8', and (iii) implementation of a semantic annotation scheme.

1 Introduction

In 1964, when the pioneering one-million-word Brown University corpus of written American English was completed and published, W. Nelson Francis, its chief architect, announced it as "a standard corpus of edited present-day American English" (Francis, 1965). His use of the term 'standard' is glossed, without any prescriptive overtones, in terms of utility for a research agenda for the future:

> It should … be of help to have a common body of material on which studies of various sorts can be based, among which comparisons can be made. It is in this sense that the corpus is hopefully called 'standard'. It can certainly be matched by parallel corpora of British English or of English of other

> periods such as the eighteenth or seventeenth century… But I am quite willing to let someone else prepare the next million words! (ibid. 273).

The anticipation of new, 'parallel'[1] corpora subsequently being added is reflected in the code 'E1' – i.e. first English corpus – which Francis inserted into the Brown corpus coding scheme. To date, Francis's invitation has been taken up for several varieties of English, all of them so far within the late twentieth century. The first round of development was essentially synchronic. The LOB corpus (Johansson et al. 1978) of written British English matched the Brown corpus with respect to its year of sampling, 1961, and (almost exactly) its sampling frame and representation of different text types. During the 1980s work on the addition of other 'standard' varieties to the Brown family led to improved corpus representation of Englishes worldwide; new corpora included Indian English (Shastri, 1988), Australian English (Collins and Peters, 1988), and New Zealand English (Bauer 1993).[2]

It was not until the 1990s that a clearly diachronic element was introduced to the collection, along the lines envisaged by W. Nelson Francis. The FLOB and Frown corpora compiled at Freiburg University represented, respectively, written British English in 1991 and American English in 1992. Because of the thirty-year 'generation' gap between Brown/LOB and Frown/FLOB, and their closely matching design, the four corpora offered an unprecedented opportunity for linguists to investigate and compare real-time changes within two major varieties of written language. Moreover, because of the recentness of FLOB and Frown, such changes could be justifiably claimed to represent changes 'in progress'in written usage (see Mair 1998 and Hundt and Mair 1998). Copyright clearance of permissions meant that researchers in other sites could share in the benefits.

This article presents the first research we are aware of undertaken to extend the Brown model *backwards* in time. For want of a better term, we refer to the new Lancaster1931 corpus as a 'prequel' to LOB and FLOB, in the sense that it predates these corpora with respect to sampling period, although it follows them in its date of compilation. The next planned extension to our project will be a further prequel of this kind, that is, a matching corpus of British English texts published in 1901, provisionally named Lancaster1901. On completion, the four corpora, FLOB, LOB, Lancaster1931, Lancaster1901, will hopefully provide linguists with an extensive and strictly comparative basis on which to track change in written British English in the twentieth century. We aim to clear copyright permissions for release of Lancaster1931 by 2006.

2 Existing corpora of English spanning the twentieth century

As yet, there is a dearth of corpora of English spanning the whole of the twentieth century, or more particularly spanning the early part of it.[3] Those that exist, to our knowledge, tend to be restricted in that they either are not generally available to the research community, or sample a very restricted range of genres. Westin's (2002) Corpus of English Newspaper Editorials (CENE), for instance, is a collection of institutional editorials sampled at ten-year intervals across the twentieth century. It is based on three 'broadsheet' British newspapers (*The Times*, the *Guardian* and the *Daily Telegraph*). Due to copyright restrictions, the CENE is not yet available to the public. This is a pity, because Westin makes a number of important observations and claims on stylistic change in editorials across the twentieth century, including, notably, that there has been an increasing informality of style, and increasing density of lexical information (Westin, 2002).[4] This raises the question, do other genres covering the same period reveal the same or different trends?

Similar remarks apply to Bauer's corpus of *The Times* (Bauer, 1994). It too consists of editorials, sampled at decade intervals, and is not publicly distributed. Bauer points out, however, that the newspapers sampled for the creation of the corpus are all accessible in public libraries around the world, although the laborious process of recreating his corpus would make it difficult for anyone to replicate the analyses and findings.

The early twentieth century is partially covered by David Denison's Corpus of Late Modern English Prose (Denison, 1994), consisting of informal private letters written between 1861 and 1919, by British writers. This corpus sits more firmly in the nineteenth century than the twentieth century, and is again of a single genre; but it is available to the research community.

In contrast to the above, the ARCHER corpus (Biber et al., 1994) covers a wide chronological span (1650 to the present) and a diverse range of genres (e.g. drama, medical, historical and news reportage texts). On the other hand, only two period samples are taken per century, bracketed into fifty-year blocks. ARCHER provides a rich array of diachronic data, but is as yet not available to the research community.

3 Characteristics of the Brown family of corpora

Following the design blueprint of the other corpora in the Brown family, Lancaster1931 consists of one million words of printed English spread across 500 texts. Each text sample consists of approximately 2000 running words, selected at a random point in the original source. The sampling range covers 15 text categories, including, for example, newspaper reportage, popular lore, learned writing, and romance fiction.[5]

These categories correspond loosely to the traditional notion of ‘genre’, ‘register’ or ‘text type’. It should be pointed out, however, that within many of the categories there is a great deal of internal heterogeneity: news reportage, for instance, comprises a miscellany of news types, including political, cultural, sports, and ‘spot’ news. Category ‘E’ comprises hobbies (books and magazines) and professional skills and trade journals. Within each text category, there are subcategories such as *national press* vs. *provincial press*; *books* vs. *periodicals*; *natural sciences* vs. *social/behavioural sciences*. At a higher level of generality, six of the categories can be grouped under the heading of ‘fiction’ (or ‘imaginative’), and nine under the heading of ‘non-fiction’ (or ‘informative’).[6]

Although the corpora represent a rich variety of genres, they do not record any biographical or demographic information about the writers: e.g. their geographical origin, sex, age, education. In many cases such information was simply not known. (See discussion of problems of this kind in Bauer 2002).

Some of the corpora are part-of-speech (POS) tagged, and in some cases the tags are post-edited and where necessary corrected. However, up to now there have been differences of mark-up and tagging conventions preventing easy comparisons of distribution of grammatical features across the four corpora. In collaboration with Christian Mair and his team at Freiburg, we have been progressively remedying this situation by applying a common tagset and a common mark-up scheme to the four existing corpora Brown, LOB, FLOB and Frown, as well as to the new and still incomplete corpus Lancaster1931. These improvements are described in section 5 below.

Apart from these limitations, it is well-known that the Brown model is representative of the language only in a very limited sense. It is restricted to written language, and further to mainstream standard varieties of public, printed text. It does not include more peripheral or ‘exotic’ registers such as poetic and dramatic texts; advertising; ephemera; private correspondence.

4 Issues in sampling

4.1 Target sampling interval (periodization)

We chose 1931 (± three years – see 5 below) as the target sampling year in order to maintain the thirty-year gap already established between the existing corpora of British English (LOB and FLOB) and of American English (Brown and Frown). The planned Lancaster1901 corpus will conform to this pattern: our intention is to include British English texts published in 1901 (± 3 years), thus maintaining the thirty-year gap. The equidistant positions of the corpora in chronology will provide evidence as to whether a particular change is speeding up, slowing down, or following an even trend. Studies based on LOB/Brown and FLOB/Frown have shown that even thirty years are long enough to reveal significant changes in the distribution profiles of numerous grammatical categories. For example, from the evidence in these corpora, modal auxiliaries have undergone a dramatic, almost wholesale, decline (Leech 2003, Smith 2003) and the present progressive construction has increased significantly in frequency (Smith 2002). Among relativization options, the *wh*-relative pronouns have declined, and the use of *that* and zero relativization has increased. Further sharp differences in grammatical frequency between the 1961 and 1991 corpora are discussed in Leech (2004).

One departure from the practice established for the existing corpora,[7] however, is that the sampling procedure for Lancaster1931 permits a leeway of three years on either side of the target year (i.e. sampling is from 1928 to 1934 inclusive). This change of practice was unavoidable because constraints on the budget and the duration of the project made it impractical to confine the sampling to a single year of publication. Since sampling for most diachronic corpora of earlier centuries of English has taken place across much wider date ranges, up to half a century in some cases, it seems unlikely that the seven-year span of Lancaster1931 texts will undermine the utility of the corpus for the study of diachronic change in the twentieth century.

4.2 Genre evolution: a problem for comparability of corpora

Text genres tend to be more synchronically heterogeneous, and more fluid over time, than is generally recognized (see Wright 1994). Some recent empirical studies over the course of the Modern English period have shown that many established genres have evolved considerably; see Atkinson (1999), Biber and Finegan (1989) and McIntosh (1998). Further corpus-based studies suggest that in the twentieth century alone, major stylistic changes have taken place in one or more genres: notably informality of style has increased (colloquialization), as has density of lexical information (Biber and Clark 2002, Westin, 2002; Mair, 1998; Hundt and Mair 1999). This observation applies to both AmE and BrE, and probably to other regional varieties such as Australian English.

4.2.1 *An issue for the linguist to bear in mind*

At its extreme, the stylistic evolution of genres is liable to obscure the question of whether a particular change has taken place in the grammatical usage of the language. This is an important issue for linguists to bear in mind, since it is largely for the purpose of detecting grammatical change – the diachronic emergence or decline of particular grammatical features – that many look to matching corpora such as LOB and FLOB.

One way for the linguist to take account of, though not to solve, this problem is to carry out a *general* profile of stylistic change across the historical corpora, surveying a considerable array of features. This then serves as a background against which to compare change in the linguistic feature of immediate interest. The stylistic study can be either quantitative – e.g. based on the ‘multidimensional’ model of Biber (1988) – or qualitative, taking a more socio-cultural perspective, as for example in McIntosh (1998). Smitterberg’s (2002) analysis of the progressive in nineteenth century English is exemplary in combining broad stylistic profiling with a close study of a particular grammatical structure. Only by studying a range of stylistically relevant features can we tell where a change in frequency – e.g. a decline in the use of the passive – is to be seen as part of a more general trend, or as something peculiar to the feature in question.

4.2.2 *An issue for the corpus compiler*

This problem of genre evolution is perhaps particularly acute in the twentieth century, a period of extremely rapid social change, accompanied by increased professional and academic specialization, and hence genre specialization. How does this affect sampling? We will consider a case affecting the sampling of texts for the Lancaster1931 corpus.

In the late 1920s and early 1930s, some written genres were already well established, with clearly recognizable features. Detective fiction, for example, was between the wars in its 'golden age' (Knight 2003), becoming a central genre in the publishing market. Matching texts from this period to those of end of the twentieth century is unproblematic.

Meanwhile other genres during the 1920s and 1930s were still in their infancy, or at least far removed from their present form, e.g. psychology, sociology, science fiction, and romantic fiction. The 1929 edition of *Sociological Review* (vol.21, no.1), for instance, contains two articles that resemble the form and style of a modern sociology paper, and three that seem quite different – more in the tradition of philosophic and humanistic scholarship.

This problem of genre evolution is likely to become even become acute as we extend the diachronic corpus back in time, e.g. to the eighteenth or seventeenth century, as Francis suggests above. Some genres will be absent entirely. It is doubtful whether we could maintain this design for such periods, without diluting the specification of the genres (e.g. replacing psychology and sociology texts with other learned writing).

Two important desiderata of equivalent corpora we have already mentioned are representativeness and comparability ('matchingness'). It is widely assumed that if a corpus is to be reliably used as a basis for statements about (a variety of) a language, it must be in some sense representative. Francis himself (1982: 7) defined a 'corpus' with representativeness in mind: "a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis". The wide range of text sampling of the Brown model ensures that each corpus is broadly representative of published English for the relevant period and regional variety. However, representativeness and comparability can be at odds with one another. This can happen when the importance of matching of text samples in corpora like LOB and FLOB is allowed to overrule the importance of achieving a balanced representation of the language in synchronic terms, as we discuss in 4.3. below.

4.3 Diachronic matching: random sampling versus publication matching

There is a slight but possibly significant difference in sampling frames used for LOB and FLOB. In the case of LOB the sampling frame – that is, the total population of publications considered – included *all* books and periodicals that met the twin criteria of: (i) falling within the fifteen genre categories listed above, and (ii) being published in 1961. The titles were extracted from exhaustive listings: for books, the British National Bibliography Cumulative Subject Index, 1960-1964, and for periodicals, the Willings Press Guide (1961).

In the case of FLOB the sampling frame for periodicals was largely predetermined by the selected titles in LOB. An effort was made to select articles from newspapers, magazines and journals that were in print both in 1961 and 1991-92 (Sand and Siemund, 1992: 120). Only if a publication was discontinued was another substituted. The effect of this strategy is that the later corpus is not nearly so *randomized* a sample of available publications as the earlier corpus. This policy permits a closer control over the diachronic comparison of language use by matching individual text samples as closely as possible: it gives greater *comparability*. However, this will be at the expense of *synchronic representativeness* if some of the publications in question have become peripheral to the reading public, and other titles, much more widely read, have been excluded as a result. We are not aware that any problems of this kind arose in the case of LOB and FLOB. However, the comparability between LOB and Lancaster1931 is in principle subject to the same kind of distortion, except that the time factor is reversed. That is, the situation we want to avoid is one where the matching of individual texts leads to an under-representation of some genre or sub-genre which was more frequent 30 years earlier, or conversely, the over-representation of some genre or sub-genre which was less frequent 30 years earlier. The cases of sociology and science fiction text types mentioned above are minor instances of this kind. However, they do not distort the picture to the extent that we have felt it necessary to abandon the practice of matching text by text, rather than relying on random sampling. We have felt it wiser to replicate as far as possible the practice employed at Freiburg in sampling texts for the FLOB corpus, so that the threefold comparison Lancaster1931 – LOB – FLOB can proceed on the basis of close text-by-text equivalence.

5 Corpus processing enhancements to the Brown family corpora

This section describes some enhancements to corpus encoding and annotation that are being applied to the new corpus, as well as to other corpora in the Brown family, beginning with the British English data (Lancaster1931, FLOB, LOB).

The purpose of these enhancements is to improve the potential use of the corpora for linguistic analysis, by ensuring an optimal degree of consistency of practice and hence comparability across these samples of twentieth century written English language.

5.1 Encoding with XML

Each of the corpora in the Brown family contains *mark-up*, that is, a set of codes representing different types of structuring and formatting information applied to the original excerpted texts. The compilers of each corpus have gone to great lengths to retain this metatextual information so as to enable ‘scoped queries’, i.e. queries operating only within the scope of a structural or formatting feature, such as direct speech quotations, or text highlighted in bold or italics. Our experience, however, is that few users of these corpora manage to exploit the potential of the mark-up, and typically filter out all mark-up codes before running a query. The problem partly stems from differences of mark-up conventions from one corpus to another. (LOB has a different scheme to Brown, and both of these differ from the mark-up of FLOB and Frown.) In addition, there has been a shortage of generally available software to allow easy exploitation of these mark-up codes.

To remedy this situation, we propose to convert existing mark-up in the corpora to XML, and to use XML in the Lancaster1931 corpus. XML is emerging as a standard, non-proprietary encoding scheme, for which a growing range of analysis software is available. Using XML and programs such as XAIRA (Burnard 2004), corpus users will find it easier to undertake context-sensitive searches, e.g. searches for all instances of infinitives in headlines, or for all occurrences of *get* not in quoted material.

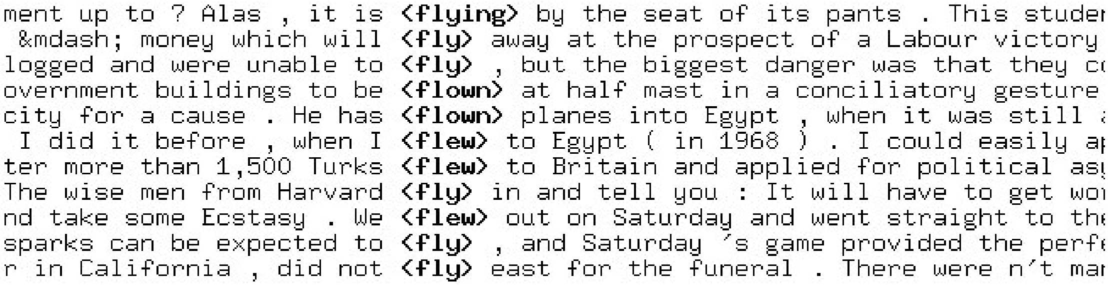
5.2 Adoption of a common grammatical tagset (Claws ‘C8’)

Morphosyntactic wordclass annotation (POS-tagging) of a corpus is an invaluable aid for the retrieval of many types of grammatical structure, e.g. passives, progressives, prepositions, relative clauses, and conjunctions. The existing corpora Brown, Frown, LOB and FLOB have already been automatically POS-tagged using the C8 tagset. This enhancement is continuing with the manual post-editing of the corpora: a procedure that is still in progress. In LOB and approximately 50 per cent of FLOB, the tags have already been post-edited and corrected. Our plan is also to tag Lancaster1931 as soon as it is complete. The original CLAWS tagset devised in 1980 for the tagging of the LOB Corpus was itself an elaboration and modification of the earlier tagset devised for the Brown Corpus, and since then has gone through many further refinements. Our latest version, C8, includes minor improvements on the C7 tagset, the richer of two versions of the tagset used for annotating the British National Corpus (BNC). Unlike C7, C8 distinguishes auxiliary from lexical forms of the primary verbs *BE, HAVE and DO.* Linguists used to using earlier tagsets in the same series will have little difficulty adapting to C8.

The tagging of Lancaster1931, LOB and FLOB with the same tagset C8, applying the same tagging system, will permit the comparative analysis of all three corpora in terms of grammatical categories, so that it will be possible to ascertain whether the trends already observed in the comparison of LOB and FLOB can be traced back to 1931 using the earlier corpus.

5.3 Lemmatization

POS-tagging is also a pre-requisite to accurate lemmatization. With the aid of lemmatization, users of the corpora will be able to make lexico-grammatical searches more efficient (because the variants do not need to be found separately), and more discriminatory (because lemmas are part-of-speech based). Figure 1 illustrates this with an extract from a KWIC concordance of the verb lemma FLY, sampled from the BNC.



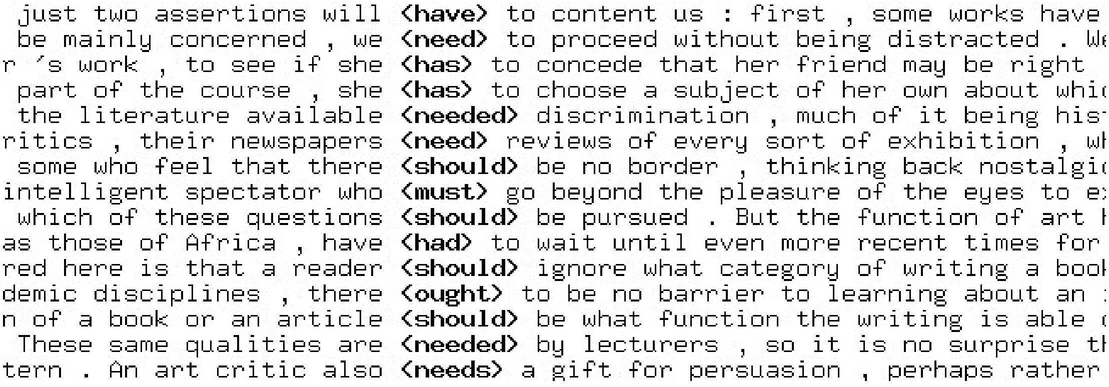
```
ment up to ? Alas , it is <flying> by the seat of its pants . This studen
 &mdash; money which will <fly> away at the prospect of a Labour victory
logged and were unable to <fly> , but the biggest danger was that they c
overnment buildings to be <flown> at half mast in a conciliatory gesture
city for a cause . He has <flown> planes into Egypt , when it was still
 I did it before , when I <flew> to Egypt ( in 1968 ) . I could easily a
ter more than 1,500 Turks <flew> to Britain and applied for political as
The wise men from Harvard <fly> in and tell you : It will have to get wo
nd take some Ecstasy . We <flew> out on Saturday and went straight to th
sparks can be expected to <fly> , and Saturday 's game provided the perf
r in California , did not <fly> east for the funeral . There were n't ma
```

Figure 1: Concordance of the verb lemma FLY, sampled from the BNC.

The lemmatization procedure being used is an adaptation of Beale (1987), and has been applied exhaustively to other corpora including the BNC (Leech, Rayson, Wilson 2001).

5.4 Semantic annotation

Annotation with a state-of-the-art semantic tagger (Rayson et. al 2004) will allow researchers to add an unprecedented level of sophistication to their linguistic querying of the corpora. They will be able to systematically take into account the contribution of semantic factors in the distributional behaviour of grammatical categories, and compare these patterns over time. As an example, Figure 2 shows a KWIC concordance view of verbal (including auxiliary) expressions of obligation and necessity based on a combination of a semantic tag and a grammatical tag.



```
 just two assertions will <have> to content us : first , some works have
 be mainly concerned , we <need> to proceed without being distracted . W
r 's work , to see if she <has> to concede that her friend may be right
 part of the course , she <has> to choose a subject of her own about whi
 the literature available <needed> discrimination , much of it being his
ritics , their newspapers <need> reviews of every sort of exhibition , w
 some who feel that there <should> be no border , thinking back nostalgi
intelligent spectator who <must> go beyond the pleasure of the eyes to e
 which of these questions <should> be pursued . But the function of art
as those of Africa , have <had> to wait until even more recent times for
red here is that a reader <should> ignore what category of writing a boo
demic disciplines , there <ought> to be no barrier to learning about an
n of a book or an article <should> be what function the writing is able
 These same qualities are <needed> by lecturers , so it is no surprise t
tern . An art critic also <needs> a gift for persuasion , perhaps rather
```

Figure 2: Concordance of verbal (including auxiliary) expressions of obligation/necessity. (Based on a sample of semantically tagged material in the BNC.)

6 Conclusion

The new Lancaster1931 corpus represents a new direction of diachronic expansion of the Brown family of corpora. It will add valuable evidence on grammatical and other changes over the sixty-year period 1931-1991. Whereas a significant difference of frequency between LOB and FLOB may suggest an ongoing trend, the evidence of the third, 1931 corpus will enable the analyst to confirm that trend by plotting three points on a graph, instead of two. In addition, it will be fascinating to discover what proportion of ongoing grammatical and stylistic changes, like those observed between LOB and FLOB, are recent developments, and what proportion are simply continuations of trends that were already established in the 1930s.

With the addition of the Lancaster1931 corpus we also plan to introduce some processing enhancements to the Brown family corpora, to facilitate comparisons of within and across the corpora, and so improve their usefulness for further linguistic research.

References

Atkinson, Dwight (1999) *Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London, 1675-1975*. Mahwah, N.J.: Lawrence Erlbaum

Bauer, Laurie. 1993. *Manual of Information to accompany the Wellington Corpus of Written New Zealand English*. Department of Linguistics, Victoria University of Wellington.

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

Bauer, Laurie. 2002. Inferring Variation and Change from Public Corpora. In Chambers, J., Peter Trudgill, and Natalie Schilling-Estes (eds.) *The handbook of language variation and change*. Malden, Mass.: Blackwell, 97-111.

Beale, Andrew. 1987. Towards a Distributional Lexicon. In: R. Garside, G. Leech and G. Sampson (eds). 1987. *The Computational Analysis of English: A Corpus-based Approach*. London: Longman, pp. 149 - 162.

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

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8.243-257.

Biber, Douglas and Edward Finegan. 1989. Drift and evolution of English style: a history of three genres. *Language* 65.3: 487-517.

Biber, Douglas, Edward Finegan and Dwight Atkinson. 1994. ARCHER and its challenges: Compiling and exploring A Representative Corpus of Historical English Registers. In *Creating and using English language corpora*, ed. by U. Fries, G. Tottie and P. Schneider, 1-14. Amsterdam: Rodopi.

Biber, Douglas and Victoria Clark. 2002. Historical shifts in modification patterns with complex noun phrase structures: How long can you go without a verb?. In

Fanego, Teresa, Javier Pérez-Guerra and María José López-Couso (eds.) *English Historical Syntax and Morphology*, 43–66.

Burnard, Lou. 2004. Xaira: an XML-aware indexing and retrieval architecture. Talk presented at *Digital Resources for the Humanities*, 5-8 September 2004, University of Newscastle upon Tyne.

Collins, Peter and Pam Peters. 1988. The Australian corpus project. In *Corpus linguistics, hard and soft,* ed. Merja Kytö et al. Amsterdam: Rodopi, 103-120.

Denison David. 1994. A Corpus of Late Modern English Prose. In M. Kytö et al. (eds.), *Corpora Across the Centuries*. Amsterdam: Rodopi, pp. 7-16.

Francis, W. Nelson. 1965. A standard corpus of edited present-day American English. *College English* 26: 267-273.

Francis, W. Nelson. 1982. Problems of assembling and computerizing large corpora. In Johansson, Stig (ed.) Computer corpora in English language research. Bergen: Norwegian Computing Centre for the Humanities. pp.7-24.

Greenbaum, Sidney (ed.). 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press

Hundt, Marianne and Christian Mair. 1998. 'Agile'and 'uptight' genres. The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4(2): 221-242.

Johansson, Stig, Geoffrey Leech and Helen Goodluck. 1978. *Manual of Information to Accompany the Lancester-Oslo/Bergen corpus of British English.* Department of English, University of Oslo, Oslo.

Knight, Stephen. 2003. The golden age. *The Cambridge Companion to Crime Fiction*. Cambridge: Cambridge University Press, pp. 77-94.

Leech, Geoffrey. 2004. Recent grammatical change in English: data, description, theory, in K. Aijmer and B. Altenberg (eds), *Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora* (*ICAME 23) Göteborg 22-26 May 2002*, Amsterdam: Rodopi, pp. 61-81.

Leech, Geoffrey, Paul Rayson and Andrew Wilson. 2001. *Word Frequencies in Written and Spoken English: based on the British National Corpus*. London: Longman.

Mair, Christian. 1998. Corpora and the study of the major varieties of English. Issues and results, in: Lindquist, H. et al.: *The major varieties of English: Papers from MAVEN 97*, Växjö, S. 139-157.

McIntosh, Carey. 1998. The evolution of English prose 1700-1800: Style, politeness and print culture. Cambridge: Cambridge University Press.

Rayson, Paul, Dawn Archer, Scott Piao and Tony McEnery. 2004. The UCREL semantic analysis system. In *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks* in association with 4th International Conference on Language Resources and Evaluation (LREC), May 2004, Lisbon, Portugal, pp. 7-12.

Sand, Andrea and Rainer Siemund 1992. LOB-30 years on ... ICAME Journal 16:119-122.

Shastri, S. V. 1988. The Kolhapur Corpus of Indian English and work done on its basis so far. *ICAME Journal*, 12, 15-26.

Smitterberg, Eric. 2002. The progressive in nineteenth century English: a process of integration. Uppsala: unpublished doctoral thesis.

Westin, Ingrid and Christer Geisler. 2001. A multi-dimensional study of diachronic variation in British newspaper editorials.

Westin, Ingrid. 2002. *Language Change in English Newspaper Editorials*. Amsterdam: Rodopi.

Wright, Susan. 1994. The Place of genre in the corpus. In Kyto, Merja, Matti Rissanen and Susan Wright (eds.) *Corpora across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora* Amsterdam: Rodopi: 101-7.

[1] The term ‘parallel’ is placed in quotation marks because it is commonly used in another sense, to refer to corpora consisting of aligned texts, typically source and translation texts, of two or more languages. Here we prefer the term ‘matching’ corpora or ‘equivalent’ corpora.

[2] Although they are not quite synchronous with Brown and LOB, the common design these corpora share with each other affords a wider survey of the linguistic characteristics of global Englishes in recent times. A similar enterprise, following a different design model, is represented by the corpora in the *International Corpus of English* project (Greenbaum, 1996).

[3] As in the present project, there were compiled principally for purposes of grammatical research.

[4] Westin’s findings are based on a multidimensional analysis, using the methodology of Biber (1988). They are broadly consonant with findings in Biber and Finegan (1989) and Biber and Clark (2002).

[5] The corpus texts were compiled into electronic form through OCR scanning and manually keying in the source texts.

[6] The exact composition of the categories varies slightly between the American and British corpora, due to differences of publishing environment in the respective countries. This difference (which is maintained in Lancaster1931) does not seriously detract from inter-corpus comparability, or the use of the term ‘Brown model’ for corpus design.

[7] The Wellington Corpus (of New Zealand English) covers a two-year sampling period, 1986-87.