4 Inferring Variation and Change from Public Corpora

LAURIE BAUER

When Lass (1987: 21) presents three versions of the same biblical passage to illustrate the differences between Old English, Middle English, and Early Modern English, he is using a technique which is well-established in the tradition of introducing historical variation to beginners. These three texts, he notes, are all called English, yet they are visibly different and in lectures they might also be shown to be audibly different. We infer change to the language in the periods between the times when these three translations appeared by considering an easily available sample of comparable material produced at three periods. These samples provide a small corpus of publicly available (and thus confirmable) data on the basis of which we infer language change. In this sense the whole philological tradition is based on the study of public corpora, and the study of language change that derives from this tradition is a corpus-based study.

Traditional dialectology takes this notion one step further by creating corpora rather than using pre-existing corpora. When Edmond Edmont cycled around France noting answers to the questions on his questionnaire, he was collecting a corpus, which became public with the publication of the *Atlas linguistique de la France* (Gilliéron, 1902–10). At that point, the corpus could be exploited by Gilliéron and other researchers to illustrate linguistic variation within the "Gallo-Roman" area, and to consider mechanisms of linguistic change.

Today, with the rapid improvements in computerized type-setting and scanning technology, with the decreasing cost and increasing efficiency of computer memory, we tend to interpret the word "corpus" rather more narrowly to mean "electronically searchable text database." This rather new usage means that we need to be explicit in defining our notion of a corpus. We also need to consider what exactly it means for a corpus to be "public."

1 The Notion of Public Corpus

1.1 What is a public corpus?

Kennedy (1998: 1) defines a corpus as "a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description." Even this wide definition may seem in some respects too narrow. For example, it excludes a body of sound recordings which can serve as a basis for linguistic analysis and description, although sound recordings (not necessarily transcribed in any sense whatsoever) might be the best basis for providing a linguistic analysis of the phonology of a particular variety. By focusing on "text," Kennedy's definition might be taken to exclude word-lists and the like, which might nevertheless be an appropriate basis for the description of things like word formation patterns in a particular variety. Accordingly, I would prefer to modify Kennedy's definition to read "a body of language data which can serve as a basis for linguistic analysis and description." This definition is (intentionally) inclusive, possibly excessively so, since it would even allow a set of sentences invented by so-called "arm-chair" linguists to prove a particular grammatical point as a corpus. We might wish to modify the definition further to guarantee that the language data are naturally occurring language data (although that might be hard to define strictly), or at least language whose original purpose was not to explain or justify some linguistic analysis.

"Public" is defined (in the relevant sense) by *The Macquarie Dictionary* (Delbridge 1997) as "open to the view or knowledge of all." In principle this looks clear-cut: Edmond Edmont's transcriptions noting the answers he received to the questions on his questionnaire were his own personal research notes and were not "open . . . to all," while the *Atlas linguistique de la France*, by virtue of being published, is open to all because it is in the public domain. In practice there are many intermediate stages. Let us consider just two.

In 1946 the New Zealand Broadcasting Service established a Mobile Disc Recording Unit which traveled round various parts of provincial New Zealand recording pioneer reminiscences for posterity. Unfortunately not all areas of New Zealand were covered, since a change of government curtailed finances for the project. Mainly elderly people were interviewed, although some middle-aged speakers were included. Their dates of birth range from approximately 1850 to 1900. This material was made available to the Origins of New Zealand English (ONZE) project under the direction of Elizabeth Gordon and Lyle Campbell at the University of Canterbury, New Zealand in 1989 (Lewis 1996). Some of this material has been transcribed orthographically and some has been subjected to detailed phonetic analysis, both auditory and acoustic by members of the ONZE team. Various results from the project have been published (see, e.g., Gordon 1994, 1998, Trudgill et al. 1998). This

material was collected by a public company, was intended for broadcast, and can be regarded as a kind of archive to which later broadcasters might have access. This seems "public" in the relevant sense. The corpus is available in two research libraries in New Zealand, and this too seems "public" in the relevant sense. Yet it is public in a very restricted way because it is hard for most researchers to get to the libraries, and the associated speaker data collected later by the ONZE team is not part of the library record. A release of material on CD is foreseen, but currently the corpus is only marginally "public," and the original intention behind the data collection is a poor guide to the current status of the corpus.

In order to create a corpus for tracking the progress of grammatical change in twentieth-century English, I used leading articles from *The Times* newspaper of London (Bauer 1994: 50). The first ten leading articles, excluding Sunday leaders in later years, were chosen from the March editions of the newspaper (the month was chosen at random) for the years 1900, 1905, 1910, 1915 . . . 1985. Given that information, it should be possible for other researchers to recreate precisely the database that I was working with without making any direct application to me. *The Times* can be consulted on microfilm in many good libraries all round the world, and is clearly public domain material. The collected material remains in my own research notes, and is not public material, but anyone can recreate the same files that I had to look at. This seems to me to be a slightly marginal case, but basically to be a public corpus. It is marginal because any other linguist wishing to check my findings would have to go to considerable effort to replicate the material I used; it cannot simply be purchased/downloaded as a body of material. Now that newspapers are producing back-copies in annual form on CD-ROM, patterns of what is possible may be changing here (Minugh 1997). Had the material been spoken instead of written, the case might have been different. This is the situation with Van de Velde's (1996) study of Dutch pronunciation on the basis of tapes from Dutch and Belgian radio. The original broadcasts were public, but it is extremely difficult for other linguists to reconstruct Van de Velde's data for themselves, since they would have to apply to all the original radio stations, or to Van de Velde himself. This corpus thus seems not to be a public one, and the implication is that a public corpus of spoken data for phonetic analysis will be extremely difficult to find, although CD-technology means that it is becoming more possible. The recently-published CD of Survey of English Dialects material (SED 1999) is an as yet rare example.

We have thus defined a public corpus as a body of data which can serve as the basis for linguistic analysis and description and which is available to linguists in general either as an identifiable whole or from easily accessible materials. We shall now go on to consider the types of such corpora which are available.

1.2 Types of corpus

We can distinguish several types of public corpus, in a multidimensional matrix. First, we must distinguish between corpora which are individual linguists' ad hoc collections of available materials (such as the one described for Bauer 1994 above) and those which are created as deliberately structured corpora, and which are made available as such. There are benefits to each type. The corpora which are created as such are more likely to have texts carefully chosen to be representative in some way – for instance the texts for the Brown corpus of written American English from 1961[1] were specifically "selected by a method that makes it reasonably representative of current printed American English" (Kučera and Francis 1967: xvii). On the other hand, such corpora may not be suitable for answering all questions of linguistic description. However representative the Brown corpus is of written American English in 1961, it could not hope to answer the questions of diachronic development for which I developed my own ad hoc corpus of material from *The Times*, already mentioned above, because that was not what it was constructed to do.

Next we must distinguish between corpora collected in paper form and those which are collected in electronic form, and thus allow electronic searches with any one of the large number of software packages now available for the purpose, e.g. The Oxford Concordance Program, TACT, WordCruncher or WordSmith to name but a few (for a review of these and others, see Hofland 1991, Kennedy 1998: 259–67, McEnery and Wilson 1996: 189–91). On the whole, this is a question of the age of the corpus and whether the corpus is being created for a particular project or being seen as something that can be exploited for multiple purposes. It was some time after the appearance of the Brown corpus in 1964 that its full value could be exploited by linguists all round the world, because in 1964 computing was still seen as something for the sciences rather than the humanities and as being very esoteric. Even today, the collection of a corpus is not a trivial matter and for many purposes it remains as easy to search a corpus by eye (or by hand) as to search it electronically (consider, for example, corpora of data from young children or second-language learners, where the same lexical item may appear in more forms than it could have in the standard language, and where the computer may not be able to predict the full range of relevant forms). On the whole, paper-based corpora do not get the same kind of wide distribution that electronic corpora get, but it must be recalled that corpora such as the Bible, the complete works of Shakespeare or *The Oxford English Dictionary* have been in use for many years, and some of the London-Lund Corpus of Spoken British English is available as a book as well as in electronic form (Svartvik and Quirk 1980).

Next we need to distinguish between corpora of written and spoken language. Corpora of spoken language may simply provide an orthographic transcription (e.g. The Wellington Corpus of Spoken New Zealand English (WCSNZE)),

or they may provide a more or less detailed phonetic transcription (e.g. The London-Lund corpus which is marked up for prosodic categories such as stress and intonation). In principle, even the orthographic transcription is not necessary if the sound recordings are available.

Next we need to distinguish between simple and comparative corpora. The labels are mine, but the idea is to distinguish been corpora which represent the language/variety to be described at one synchronic moment in time versus those which present diachronic data; to distinguish between those which represent a single variety as opposed to those which represent two or more varieties, and so on. A corpus such as Brown is a simple corpus, in these terms, describing written American English in 1961, but may nevertheless be used comparatively alongside the Lancaster-Oslo-Bergen (LOB) corpus which is closely modeled on Brown, but presents written British English in 1961. Although this division may seem a clear-cut one, it is not. Following the publication of the Brown corpus, a number of 1-million word corpora were developed which were modeled on it and which, like Brown, contain a number of thematic sections, such as press reportage, government documents, and scientific writing. It is a repeated research finding (and not at all a surprising one) that these different sections differ from each other linguistically, perhaps especially in terms of style (Biber 1988, Sigley 1997a, 1997b, 1997c). Thus it might be claimed that corpora which contain stylistically differentiated sub-components are implicitly comparative rather than simple. Taken to extremes this would mean that no corpus was ever simple, since even a single work can have passages which are stylistically distinct within it. This dichotomy has thus to be applied in a fairly uncritical manner.

Finally we need to distinguish between corpora which are made up of textual material and those which are made up of word-lists. Among the former are all the major electronic corpora that have been discussed so far, among the latter are dictionaries, thesauruses, vocabularies, and the like. Although most electronic corpora are made up of texts, these word-lists deserve the title of "corpora" (1) functionally, in that they allow comparisons of language types along several different dimensions and (2) formally since they are bodies of data created for one purpose which may nevertheless be exploited for other purposes "for linguistic analysis and description." Table 4.1 provides a classification of several public corpora according to these distinctions.

An appendix in Aijmer and Altenberg (1991) provides a short description of the electronic corpora discussed in that volume, along with some bibliographical references. McEnery and Wilson (1996: 181–7) provide brief characterizations of the electronic corpora they discuss in their text, with addresses (electronic and snail-mail) for further information. Kennedy (1998: 23–57) provides a detailed survey of most of the major electronic corpora available for English. Rissanen (2000) provides a useful overview of historical corpora of English in electronic form, including e-mail addresses or URLs for getting details on availability, etc.

Table 4.1 Classification of some public corpora

Corpora	Classification and comments
Brown, LOB, Kolhapur, ACE, WCWNZE, Frown, FLOB	Structured, electronic, written language, simple, textual. Since these corpora are all constructed on the same basic model, any two or more of them may be used comparatively.
ICE	Structured, electronic, both written and spoken language, comparative (in that written and spoken are both extensively covered, but also in that the various national sections of ICE can be compared), textual.
Jones, *English Pronouncing Dictionary* (Jones, 1917–)	Structured (in its attempt at exhaustivity), paper, spoken language, simple, word-list. The various editions together can be used as a comparative (diachronic variation) corpus, as in Bauer (1994).
The Bible	Ad hoc, paper (now also available electronically), written language, simple (although various editions can provide comparative diachronic data), textual.
Helsinki	Structured, electronic, written language, comparative, textual.
The Bank of English	Ad hoc (its representativeness arises through its sheer size rather than through the careful selection of texts), electronic, written, simple, textual.

2 Benefits and Problems Provided by Public Corpora

The main benefit accruing from the use of a public corpus is replicability. Sigley (1997c: 218) reports that *whose* and *of which* are more or less evenly divided in nonrestrictive relative clauses with inanimate antecedents – i.e. sentences such as *These menus present . . . alternate choices, whose selection/selection of which leads to further menus . . .* – in New Zealand English (55.6 percent and 44.4 percent respectively). If subsequent researchers find this distribution unexpected, they can check the results for themselves in the relevant corpus. While there may be some slight variability caused by experimental method (e.g. in the case in point, the definition of nonrestrictive or inanimate), we would not expect any gross deviations without contrasting theoretical presuppositions (e.g. a presuppositions that some of the tokens of *of which* belong to some completely separate grammatical structure). Replicability of this type is a sign of good science.

The other major benefit of corpora, the possibility of treating such phenomena numerically, accrues to all corpus studies, not just to studies based on public corpora.

There are also a number of problems which corpora give rise to. Again, these problems are not specific to public corpora, but to all corpus studies. Some of these are discussed, in no particular order, below.

First, by allowing numerical treatment, corpus studies allow an appearance of precision which may be totally spurious. The researcher needs to ask how far the corpus reflects anything but the collection of texts/words which make up the corpus. Does the 55.6 percent figure for *whose* in nonrestrictive relative clauses with inanimate antecedents really tell us about New Zealand English, or just about the Wellington Corpus of Written New Zealand English (WCWNZE, reflecting language in 1986)? Even if the Brown corpus was deliberately constructed to reflect printed American English in 1961, is there any guarantee that a different sample of actual texts built on the same framework for the same variety of English and in the same year would give the same results for any (let alone all) of the linguistic variables? We do not know what the margin of error is in linguistic corpora of these types.

Correspondingly, there are problems when comparing two or more corpora. The Freiburg versions of Brown and LOB (Frown and FLOB), with texts from 1991 and 1992 rather than 1961, are specifically designed to be comparable with their earlier congeners. Yet where a difference is discovered between the language of the two, it is not necessarily clear that it can be entirely attributed to the 30-year gap between the two corpora. Mair (1998: 148) reports that an increase in progressives between Brown and Frown (and also between LOB and FLOB) is due to the more frequent choice of an informal option rather than a non-progressive formal option in places where either is possible. While this might be a language change, it might equally be viewed as a societal change in perception of formality or as no change at all, just a different exploitation of precisely the same system. Moreover, that exploitational difference might reflect text-selection rather than any change in norms. Similarly, WCWNZE was based firmly on LOB, and is intended to be comparable with LOB. Yet at least one major difference between the two is drawn attention to in the manual which accompanies WCWNZE: the fact that because of different publishing traditions, mass-market fiction is not often published in New Zealand, and what is published in New Zealand is "more consciously literary" (Bauer 1993: 2). The fiction categories of the two corpora are thus different, and so the differences between the two corpora as wholes might reflect differences in style as much as differences in geographical origin or diachronic differences. Problems of this type are soluble only in terms of consistency of findings. Where research based on different corpora shows comparability of results in terms of varietal or temporal differences, we can be relatively sure that these represent genuine differences of variety. The more such results we have, the more we can trust results from corpora which have in other ways shown themselves to be representative.

Next, corpus size provides a problem. It is not necessarily clear in advance of testing whether a corpus is large enough to provide an answer or not. Yet too large a corpus means that the experimenter is left analyzing unnecessary data. In Bauer (1994: 50–1) I referred to this as Murphy's Law applied to corpora. The corpus I used from *The Times*, and another from the *New York Times*, were too small to provide particularly clear results on the subject of whether there is a change in comparative and superlative marking in English from a synthetic form like *remoter* to an analytic form like *more remote*. Work using the larger British National Corpus (BNC) (Kytö and Romaine 1997; Leech and Culpepper 1997) provides rather more definite results (although even those results are not entirely clear). Corpora used to investigate lexical matters generally have to be extremely large (hence the involvement of so many dictionary-publishers in the construction of the BNC); some grammatical phenomena are also so rare in texts as to require very large corpora if reasonable amounts of data are to be found: the use of *whose* and *whom* in modern English are phenomena in point.

It was implicit in what was said above that corpora are not consistent in their style level across all the texts they include. Some corpora, such as WCSNZE, the Helsinki corpus of historical English, and the various corpora involved in the International Corpus of English (ICE) project, with parallel 1-million word corpora planned from some 20 countries, include biographical details (including age, gender, ethnicity, educational achievement) of all the speakers/writers whose output is represented. This makes it possible to look for linguistic differences between the various social groupings for which the corpus is marked up. In most written corpora, this information is at best indirectly derivable and at worst unavailable: corpora of press materials, especially reportage, rarely mark authors' gender, for example, for the simple reason that it is usually unknown, in that most such items are not signed/by-lined; indeed, a single news item may have been edited so many times that it no longer has "an author" of identifiable gender. This does not mean that differences correlated with gender (or, a fortiori, the other social categories) are not present in the text; it just means that they cannot be isolated. Accordingly, the entire corpus will be undetectably biased for any relevant factors by the language of the social group which provided the majority of the texts. In practical terms, this should probably be interpreted as implying that marginally significant differences in linguistic behavior measured in corpora cannot be trusted, and that statistical significance has to be treated with great care.

We can summarize the general point here by saying that different corpora assume different degrees of idealization about the speech community they attempt to represent: Brown assumes homogeneity across the community, the WCSNZE assumes homogeneity only within the categories it identifies (gender groups with certain educational backgrounds, etc.). In principle, these differences lead to incommensurability between corpora; in practice, as long as the statistics are dealt with carefully, they need not prevent similarities and differences from being discovered.

3 Using Corpora to Infer Variation and Change

Although there may be difficulties in interpreting results from corpus-based comparisons, as we have just seen, discovering variation or change from public corpora would seem to be relatively straightforward. Find a corpus or corpora which allow comparison on the required dimension (e.g. corpora from different historical periods, corpora of different national varieties, corpora containing written and spoken material, corpora containing utterances from both males and females who are identified as such, etc.). Comparability is a problem, but has been dealt with above. Assuming a two-way comparison, with two varieties we can call A and B, measure the linguistic behavior in variety A in the corpus, and measure the linguistic behavior of variety B in exactly the same way with respect to some potential linguistic variable V. If V is indistinguishable in A and B, assume no variation/lack of change. If V is measurably and significantly different in the two cases, postulate variation/change.

Consider the following example from Hundt (1998: 32). Hundt counts the number of regular and irregular past tense forms of various verbs in the WCWNZE and in FLOB, with the results shown in table 4.2 (taken from Hundt's 1998: 32 Table 3.4). The difference between FLOB and WCWNZE, she says, "proved significant at the 1 percent level" using a chi-square test, and we have a case for variation between these two varieties. She also comments that in the Brown corpus 96.7 percent of the relevant verb-forms are regular, so that there is a difference between American English, British English and New Zealand English on this measure, with New Zealand English being more like British than like American English, but different from both.

This example is of interest not only because it illustrates a canonical instance of the argument showing regional variation in English, but also because this result feeds in to Hundt's final conclusion, namely that "synchronic 'snapshots' focusing on regional differences can be interpreted as stages in the (regional) diffusion of change" (Hundt 1998: 134) and that what we have seen operating in table 4.2 is regional variation, but regional variation originating from different speeds of diffusion of grammatical change. Such a conclusion would not be possible on the basis of a simple two-way comparison, such as is presented in table 4.2, but emerges because of the range of material Hundt is able to sample. In other words, we have to be careful in interpreting the results of any such experiments: there may be more (or less!) to them than superficially appears.

The problems of interpretation become greater when less canonical forms of the argument are used. A cautionary tale is provided by my own study of relativization strategies in English (Bauer 1994: 66–83). Although I had created my own diachronic corpus, it was not always sufficient, and I also tried to make comparisons with other linguists' corpora. In particular, I compared the percentage of various relativization strategies (the use of a *wh*-word, the use of *that*, or the lack of any complementizer) across various corpora. Among a

Table 4.2 Irregular and regular past tense forms of various verbs

	WCWNZE	FLOB
burned	13	16
burnt	28	11
dreamed	9	5
dreamt	4	5
leaned	26	25
leant	4	13
leaped	0	3
leapt	6	7
learned	69	81
learnt	37	22
smelled	7	6
smelt	5	4
spelled	0	4
spelt	3	2
spilled	3	5
spilt	2	5
spoiled	0	4
spoilt	9	2
-ed	127 (56.4%)	149 (68.7%)
-t	98 (43.6%)	68 (31.3%)
Total	225 (100%)	217 (100%)

Source: Hundt (1998: 32)

number of linguists who reported the use of *that* as a relativizer in 15–20 percent of restrictive relative clauses, one stood out as reporting *that* use in 54 percent of cases (Biesenbach-Lucas 1987). Because the proportion was so different, I concluded that "Biesenbach-Lucas (1987) does not seem to be counting the same thing as I am, when we talk about restrictive relatives" (Bauer 1994: 80). This seemed the only way to explain the large discrepancy, particularly given that terms like "restrictive" are well-known to be hazardous. I thus made what seemed like a conservative decision that, despite appearances to the contrary, our corpora were not comparable because we were not measuring the same thing. This is a genuine problem in interpreting others' results, but in this particular case, I was wrong. In a considerably more sophisticated analysis,

Sigley (1997a: 467–73) shows that there *is* a difference between American English and other varieties on this point. Sigley's analysis shows that the difference between American English and other varieties is significant only in press usage (Biesenbach-Lucas studies the language of *The Washington Post*). He attributes the difference to overt prescription in American English for *that*-usage in restrictive relative clauses, e.g. in *The Chicago Manual of Style* (anon. 1993: §5.42). My caution in interpreting corpus results made me miss one of the most important markers of regional variation in this part of the grammar.

One point to note in all of this is that using electronic corpora limits the types of phenomena that can practicably be studied, while making it easier to study those which can be retrieved. While it is relatively simple to consider the differences in use between *which* and *that* using electronic corpora, it is extremely hard to study the use of the zero relative (as in *the man Ø I met yesterday*) because there is no form to search for. Although Sigley (1997a) managed to do a lot of searching by finding relevant environments, in the end a manual pass through the text is required. In principle, tagged corpora solve this problem, because they mark relative clauses as such, and all relative clauses can then be pulled out automatically. In practice, unless the corpus has been manually tagged, reliance on tagging is dangerous, since it is likely to miss particular types of data in a systematic manner. Since manual tagging is so time-consuming, most tagged corpora have the tags determined by computer program. While the best of these claim an accuracy of approximately 95 percent, the errors tend to congregate in particular categories, with the tag for singular noun, in particular, being over-assigned (DeRose 1991: 11; Kennedy 1998: 220). In any case, many tagged corpora do not get beyond part-of-speech tagging, and it is not possible to find relative clauses using such tagging. A notable exception is the ICE-GB corpus, which allows syntactic trees to be matched and retrieved from the corpus. It is not yet possible to assign such trees automatically, so that this extremely useful outcome is the result of a huge amount of preparation of the corpus.

4 Results

Kennedy (1998: 180–203) provides an excellent summary of the major findings concerning English that are discussed in the literature. A brief recapitulation of some of these is provided below.

4.1 Dialectal variation

On the whole corpora have been built for national varieties of English rather than for regional dialects within one country. Thus we do not have public electronic corpora that would allow us to investigate differences in the syntax of

Newfoundland and Vancouver Englishes, or of Cornish and Tyneside dialects. The presupposition is that such comparisons would either be meaningless (in that it is not clear that comparing the syntax or vocabulary of Cornish and Tyneside would be any more meaningful than comparing the syntax and vocabulary of two distinct languages), or vacuous (in that no distinctions are expected – e.g. Newfoundland vs. Vancouver). Certainly, a number of studies based on the Brown and LOB corpora found very few significant differences between American and British English grammar except for the wider use of *that* in restrictive relative clauses in American English, the greater use of regular forms of certain verbs already mentioned above, and the use of the minor construction *this prevented me from leaving* in British but not American English (Mair 1998). Studies of Australian and New Zealand Englishes have also found differences in the weighting of different strategies, but very few absolute grammatical distinctions such that one variety uses a particular construction and another invariably uses a contrasting construction (or has no equivalent construction) in the same environments. Those that have been suggested (the transitive use of *farewell*, the mediopassive use of *screen* in New Zealand English – Hundt 1998) are very much at the lexical end of grammar. However, the extent to which particular grammatical structures are used by speakers/writers may be an entirely different matter, and here Biber (1987) shows that American English uses more nominalizations, more passives and more *it*-clefts than British English, while British English uses more place and time adverbials and more subordinator deletion than American English.

That there are lexical differences between the major international varieties of English is not something about which there is any doubt, and here the best corpora are perhaps the dictionaries (although dictionaries do not give information about extent of usage, which may give a rather different picture of what is really going on; see Kennedy and Yamazaki 1999, Kennedy 2001). There are probably also differences in interactional styles. Tottie (1991) finds that American speakers provide three times as many backchannel agreement markers as their British counterparts.

4.2 Written and spoken language

The most important work in considering the differences between written and spoken English is clearly Biber (1988). Biber argues that the differences between different text types within written or spoken English are sometimes greater than the differences between written and spoken. This observation, of course, is based on treating both written and spoken language as text, and ignores the phonetic nature of spoken language completely. Others have commented on differences of vocabulary in the two media. Kennedy (1998: 184) points out that *pretty* is mainly used as a descriptive adjective in writing (the LOB corpus), and mainly used as an intensifier in speech (the London-Lund corpus).

4.3 Style

Questions of grammatical determinants of style have been considered particularly by Biber (1988) and Sigley (1997b). In particular Biber points out that syntactic complexity and lexical complexity may not (as is often assumed) go together.

4.4 Language change

Studies of language change have always, in some sense, been corpus-based (see above). The interesting discussions on the basis of modern corpora are thus not those which simply report or date a change (although some of those can provide surprising results, e.g. Peitsara 1993), but those which show a pattern to the change, which re-examine change in the light of modern sociolinguistic theory. An example is Nevalainen (2000), where, on the basis of the Corpus of Early English Correspondence developed at the University of Helsinki, women are shown to lead the introduction of verbal *-(e)s* (replacing *-th*) in Early Modern English and the use of *you* in subject position, but men are shown to lead in the move to single negation instead of double negation. The development of Frown and FLOB at Freiburg also means that we can expect to see more studies emerging of the progress of change in twentieth-century English.

4.5 Sociolinguistic variation

All variationist studies are corpus-based, but most of the corpora have not been public ones, and the results are thus not strictly germane to this survey. What we can say, perhaps, is that the use of public corpora has not led to radically different conclusions in areas such as the correlations between ethnicity, gender, or geographical origin and linguistic usage from those provided by the private corpora which preceded them. This is just what we would expect. The classic variationist studies also dealt exclusively with phonetic/phonological variation, while the rise in the use of public corpora is increasing the range of phenomena that can be studied within this framework. Morphological and syntactic phenomena, it turns out, can often be considered in precisely the same ways when sufficient searchable text is available. Nevalainen's study mentioned above is a simple example. Sigley (1997a) considers variation in relative clause construction from speakers/writers of different genders, and educational levels, and finds some significant (but numerically not very important) differences. Holmes (1998a, 1998b) illustrates the fact that corpus studies can be used to go beyond the study of forms into functions and interactional choices.

4.6 Variation in word-list corpora

The discussions above have been largely based on results obtained from public electronic corpora. Since the kind of use to which word-list corpora can be put is perhaps less obvious, it is worth devoting some space considering the kind of results that can be obtained from them.

Bauer (1994) illustrates the use of pronouncing dictionaries to show variation in particular phonological phenomena synchronically, but also, by using a number of editions of the same works, to show change in these phenomena. Specifically, stress in polysyllabic words and /j/-dropping following coronal consonants are studied, with the latter being considered in different national varieties. Bauer (1994) also uses the *Supplement to the Oxford English Dictionary* (Burchfield 1972–86) to illustrate the change in sources of borrowed words in English over a 100-year period. Dictionaries are now being widely used by morphologists as evidence for changing patterns of productivity in particular patterns of word-formation (Anshen and Aronoff 1997, Bolozky 1999, Bauer 2001).

5 Descriptive Use

A quick glance at many of the collections of papers on corpus linguistics may make it look as though the main interests which corpora present to the linguist are the problems of constructing them in the first place and the problem of parsing them once they are constructed. Both are matters of extreme complexity, and I do not wish to underestimate their importance or difficulty. In the present context, however, what is more important is the descriptive use that can be made of corpora. At one level, corpora can be used to make sure our descriptive facts are correct, and to improve the quality of grammatical descriptions and lexicological descriptions (consider, for example, the extensive use of corpora made by some of the major dictionaries aimed at non-native learners of English, such as COBUILD – Sinclair 1987). As long as care is taken, this descriptive basis can be extended fairly readily to a consideration of variation and change. While the study of variation and change is made easier by the existence of electronic corpora which are deliberately created for this purpose (the ICE corpus, the various parallels to the Brown corpus, the Helsinki corpus), variation (which may indicate change in progress) can also be discovered in simple corpora, because the set of texts used in a corpus can never be stylistically completely homogeneous.

Whereas in the past, corpora tended to be collected by individuals for their own use, the availability of corpora has increased enormously in the last 30 years or so as computers have become more readily accessible, computer memory has become cheaper, and scanning techniques have improved, and, correspondingly, the amount of work on language variation and change that

uses them has also increased. It is now easier for individual scholars to make their databases generally available than ever before. While there are often legal and ethical problems involved in doing so, it is to be hoped that this will continue to happen in the foreseeable future, because the greater the amount of genuine data that is available in this way, the better the descriptions that will be possible and the surer linguists will be of the replicability of their findings. The main point about public corpora as opposed to private ones is that their representativeness can be openly considered, and that they provide a large and readily-available body of agreed-upon data against which hypotheses can be tested. Corpora, even public corpora, are not new; the widespread use of them derives from the fact that they have become so valuable and so available. We are now almost reaching the stage where corpus studies based on public corpora are the default way of providing robust descriptions.

Appendix: List of Electronic Corpora Cited

Corpus	Abbreviation	Variety of English	Written/ spoken	Sample date	Corpus family	Size in words
Australian Corpus of English	ACE	Australian	written	1986	Brown	1m
Bank of English		mainly British	mainly written	1960–		300m+
British National Corpus	BNC	British	written and spoken	1960–		100m
Brown		American	written	1961	Brown	1m
Corpus of Early English Correspondence	CEEC		written			2.7m
Freiburg Brown	Frown	American	written	1991	Brown	1m
Freiburg LOB	FLOB	British	written	1992	Brown	1m
Helsinki		Historical	written	b. 850–1720		1.6m
International Corpus of English	ICE	Various	written and spoken	c. 1990		1m for each country
Kolhapur		Indian	written	1978	Brown	1m
Lancaster–Oslo–Bergen	LOB	British	written	1961	Brown	1m
London–Lund		British	spoken	1975–81		
Wellington Corpus of Spoken New Zealand English	WCSNZE	New Zealand	spoken	1986		1m
Wellington Corpus of Written New Zealand English	WCWNZE	New Zealand	written	1986	Brown	1m

ACKNOWLEDGMENTS

* I should like to thank Janet Holmes and Tom Lavelle for their perceptive comments on an earlier draft, which have led to many improvements. They are of course not to blame for what I have included.

NOTE

1 See Appendix for a list of the electronic corpora mentioned in the text.

REFERENCES

Aijmer, Karin and Bengt Altenberg (eds.) (1991). *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London and New York: Longman.

anon. (1993). *The Chicago Manual of Style*, 14th edn. Chicago and London: Chicago University Press.

Anshen, Frank and Mark Aronoff (1997). Morphology in real time. In Geert Booij and Jaap van Marle (eds.), *Yearbook of Morphology, 1996*. Dordrecht: Kluwer. 9–12.

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

Bauer, Laurie (1994). *Watching English Change*. London and New York: Longman.

Bauer, Laurie (2001). *Morphological Productivity*. Cambridge: Cambridge University Press.

Biber, Douglas (1987). A textual comparison of British and American writing. *American Speech* 62: 99–119.

Biber, Douglas (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biesenbach-Lucas, Sigrun (1987). The use of relative markers in modern American English. In Keith M. Denning, Sharon Inkelas, Faye C. McNair-Knox and John R. Rickford (eds.), *Variation in Language: NWAV-XV at Stanford*. Stanford, CA: Department of Linguistics, Stanford University. 13–21.

Bolozky, Shmuel (1999). *Measuring Productivity in Word Formation*. Leiden: Brill.

Burchfield, R. W. (ed.) (1972–86). *A Supplement to the Oxford English Dictionary*. Oxford: Clarendon Press.

Delbridge, A. (ed.) (1997). *The Macquarie Dictionary*. Sydney: Macquarie Library.

DeRose, Steven J. (1991). An analysis of probabilistic grammatical tagging. In Stig Johansson and Anna-Brita Stenström (eds.), *English Computer Corpora*. Berlin and New York: Mouton de Gruyter. 9–13.

Gilliéron, Jules (1902–10). *Atlas linguistique de la France*, 13 vols. [Paris]: Champion.

Gordon, Elizabeth (1994). Reconstructing the past: written and spoken evidence of early New Zealand

speech. *New Zealand English Newsletter* 8: 5–10.
Gordon, Elizabeth (1998). The origins of New Zealand speech: the limits of recovering historical information from written records. *English World-Wide* 19: 61–85.
Hofland, Knut (1991). Concordance programs for personal computers. In Stig Johansson and Anna-Brita Stenström (eds.), *English Computer Corpora*. Berlin and New York: Mouton de Gruyter. 283–306.
Holmes, Janet (1998a). Generic pronouns in the Wellington corpus of spoken New Zealand English. *Kotare* 1: 32–40.
Holmes, Janet (1998b). Narrative structure: some contrasts between Maori and Pakeha story-telling. *Multilingua* 17: 25–57.
Hundt, Marianne (1998). *New Zealand English Grammar, Fact or Fiction? A Corpus-Based Study in Morphosyntactic Variation*. Amsterdam and Philadelphia: Benjamins.
Jones, Daniel (1917–). *English Pronouncing Dictionary*. London and Melbourne: J.M. Dent. 2nd edn, 1924; 3rd edn, 1926; 4th edn, 1937; 5th edn, 1940; 6th edn, 1944; 7th edn, 1945; 8th edn, 1947; 9th edn, 1948; 10th edn, 1949; 11th edn, 1956; 12th edn, 1963; 13th edn, 1967; 14th edn, 1977.
Kennedy, Graeme (1998). *An Introduction to Corpus Linguistics*. London and New York: Longman.
Kennedy, Graeme (2001). Lexical borrowing from Maori in New Zealand English. In Bruce Moore (ed.), *Who's Centric Now? The Present State of Post-Colonial Englishes*. Melbourne: Oxford University Press. 59–81.
Kennedy, Graeme and S. Yamazaki (1999). The influence of Maori in the New Zealand English lexicon. In J. Kirk (ed.), *Corpora Galore: Analyses and Techniques in Describing English*. Amsterdam: Rodopi. 33–44.
Kučera, Henry and W. Nelson Francis (1967). *Computational Analysis of Present-day American English.* Providence, RI: Brown University Press.
Kytö, Merja and Suzanne Romaine (1997). Competing forms of adjective comparison in Modern English: what could be *more quicker* and *easier* and *more effective*? In T. Nevalainen and L. Kahlas-Tarkka (eds.), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Société Néophilologique. 329–52.
Lass, Roger (1987). *The Shape of English: Structure and History*. London and Melbourne: J.M. Dent.
Leech, Geoffrey and Jonathan Culpepper (1997). The comparison of adjectives in recent British English. In T. Nevalainen and L. Kahlas-Tarkka (eds.), *To Explain the Present: Studies in the Changing English Language in honour of Matti Rissanen*. Helsinki: Société Néophilologique. 353–73.
Lewis, Gillian (1996). The origins of New Zealand English: a report on work in progress. *New Zealand English Journal* 10: 25–30.
Mair, Christian (1998). Corpora and the study of the major varieties of English: issues and results. In Hans Lindquist, Staffan Klintborg, Magnus Levin and Maria Estling (eds.), *The Major Varieties of English*. Acta Vexionensia Humaniora 1. Växjö: Växjö University. 139–57.
McEnery, Tony and Andrew Wilson (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
Minugh, David (1997). All the language that's fit to print: using British and American newspaper CD-ROMs as corpora. In Anne Wichmann, Steven Fligelstone, Tony McEnery and

Gerry Knowles (eds.), *Teaching and Language Corpora*. London: Longman.
Nevalainen, Terttu (2000). Gender differences in the evolution of standard English. *Journal of English Linguistics* 28: 38–59.
Peitsara, Kirsti (1993). On the development of the *by*-agent in English. In Matti Rissanen, Merja Kytö and Minna Palander-Collin (eds.), *Early English in the Computer Age*. Berlin and New York: Mouton de Gruyter. 219–33.
Rissanen, Matti (2000). The world of English historical corpora. From Cædmon to the computer age. *Journal of English Linguistics* 28: 7–20.
SED (1999). *The Survey of English Dialects on CD-ROM*. London and New York: Routledge.
Sigley, Robert J. (1997a). Choosing your relatives: relative clauses in New Zealand English. Unpublished Ph.D. thesis, Victoria University of Wellington.
Sigley, Robert (1997b). Text categories and where you can stick them: a crude formality index. *International Journal of Corpus Linguistics* 2: 199–237.
Sigley, Robert (1997c). The influence of formality and channel on relative pronoun choice in New Zealand English. *English Language and Linguistics* 1: 207–32.
Sinclair, John (ed.) (1987). *Collins COBUILD English Language Dictionary*. London and Glasgow: Collins.
Svartvik, Jan and Randolph Quirk (eds.) (1980). *A Corpus of English Conversation*. Lund: Lund University Press.
Tottie, Gunnel (1991). Conversational style in British and American English: the case of backchannels. In K. Aijmer and B. Altenberg (eds.), *English Corpus Linguistics: studies in honour of Jan Svartvik*. London: Longman. 254–71.
Trudgill, Peter, Elizabeth, Gordon and Gillian Lewis (1998). New dialect formation and southern hemisphere English: the New Zealand short front vowels. *Journal of Sociolinguistics* 2: 35–51.
Velde, Hans van de (1996). *Variatie en verandering in het gesproken Standaard-Nederlands (1935–1993)*. Nijmegen: Katholieke Universiteit Nijmegen.