Proceedings of the XXII International Conference of AEDEAN, 13-28. Eds. P. Gallardo i E. Llurda. Lleida: Universitat de Lleida, 2000

Corpus Linguistics and the Teaching and Learning of Languages[1]

Chris Butler

The aim of this paper is to claim that language teachers and their students have much to gain from the appropriately organised study of authentic language, as represented in computer-readable collections of texts. The advantages of corpus-related work for language teaching and learning are discussed, in terms of authenticity of materials, availability of software for corpus analysis, and conformity with the trend towards pedagogical grammars and towards "data-driven learning", involving the student as researcher. The problems associated with corpus work are then reviewed: the technical challenges involved, the danger of confusion if too much material is presented, the limitations of available corpora. Sources of corpus materials and available software packages are then outlined. A summary of the areas of language teaching which have profited from corpus work is followed by sections on bi- and multilingual corpora and on corpora of learner language.

1. Introduction

Theoretical linguists tend to divide into two basic camps: those whose interests lie in a postulated abstract "competence" underlying actual language behaviour, and accessed largely through native speaker intuitions, and those who are concerned not only with the language system, but also with the concrete acts of "languaging" which instantiate that system and also provide evidence for it. Those whose professional concerns lie in the applied language areas, including language teachers, are naturally inclined to the second approach. Their aim is to enable their students to acquire a practical mastery of those aspects of a language which are necessary for the purposes for which the language is being learned: put in another way, what they are trying to do is help learners to achieve as close a match as possible with what Tribble (1997),[2] following Bazerman (1994), has called "expert performances" in the appropriate domain(s), whether general (e.g. "informal conversational English", the production of written narratives) or more specific (e.g. presenting a seminar on a particular topic, the writing of a scientific article in English, or the construction of an instruction manual).

Language teachers and their students thus have a lot to gain from the appropriately organised study of authentic language. It is not surprising, then, that

1. This version of the plenary lecture given at the AEDEAN conference has been reduced, due to space limitations.
2. I am grateful to Chris Tribble and to participants in the AEDEAN conference for their useful comments on earlier forms of this paper.

one of the most exciting and important influences on language teaching and learning today comes from the area known as corpus linguistics, concerned with the construction of bodies of textual material and their manipulation and exploitation using computer-assisted techniques. The recent proliferation of introductory texts on corpus linguistics, with varying slants, bears witness to the high level of interest in this area (see McEnery & Wilson 1996, Stubbs 1996, Kennedy 1998, Biber et al 1998, Partington 1998).

In this paper, I will attempt to provide an overview of the usefulness and limitations of corpus-based work in the teaching and learning of languages (for a brief survey, see also Leech 1997). I will not cover other interesting areas such as the use of corpora in the teaching of linguistics, or the teaching of corpus linguistics as an academic subject. Furthermore, in view of the interests of AEDEAN, I will concentrate mainly on applications to the teaching of English as a foreign language.

It would be as well to begin by stating just what a corpus is. The following definition is taken from the standard text by McEnery and Wilson:

> ... a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration. (McEnery & Wilson 1996, 24)

McEnery and Wilson go on to point out, however, the possibility of deviations from this prototype, and we will see later that bodies of text not conforming to all the stated conditions may be very useful in language teaching and learning.

2. Advantages in the use of corpora in language teaching and learning

2.1. *Authenticity*

The most obvious advantage in the use of corpora is the one I mentioned in the introduction to this talk: they make available to the teacher and learner quantities of authentic language, and if the corpora have been properly documented, details of the circumstances in which the language was produced will also be available. Corpora, if truly representative of a particular variety of the language, are a window on to how people actually speak and write. The importance of this can hardly be overstressed, since native speaker reports on what they do linguistically are often, and notoriously, inaccurate.

Sinclair, in particular, has emphasised that many of the patterns which emerge from the detailed study of large corpora are simply not accessible to native speaker intuition. As he points out (Sinclair 1997, 32), our intuitions are valuable, since they give us instant information about the meanings of isolated words and the well-formedness of isolated sentences, as well as about language varieties. What they do not tell us reliably, however, is what happens when words or sentences combine in real communication. Throughout his writings, Sinclair provides many examples to support this case. Which of us could lay hand on heart and swear s/he had noticed that combat as a noun is overwhelmingly concerned with physical fighting, whereas combat as a verb is used mainly in the context of social struggle (Sinclair 1992, 14)? How many of us are truly conscious of the fact that lap as a part of the body is characteristically used in prepositional constructions and not as subject or object (Sinclair 1992, 14)? How many of you have noticed that the innocuous and much-criticised adjective nice occurs attributively with the indefinite article, but almost never with the definite article, and that when predicative it tends to attract degree modifiers such as very, pretty, extremely (Sinclair 1997, 33)? These are just a random selection from the many fascinating glimpses of real usage with which Sinclair´s work abounds. The implications for language teachers and learners surely need no further emphasis.

Authenticity, then, is a cardinal point in discussion of corpora in relation to language teaching. The matter is not, however, quite so simple as my discussion so far might suggest. As Widdowson observes:

> An authentic stimulus in the form of attested instances of language does not guarantee an authentic response in the form of appropriate language activity (Widdowson 1983, 30)

For Tribble (1997), this has important consequences for the use of corpora in language teaching: it is crucial not only that students be exposed to samples of authentic linguistic production, but also that these samples be taken from genres and topics with which the students themselves have some engagement; otherwise, the response is unlikely to be positive. I will take up this point again later, in the context of the availability of corpus materials for ELT.

2.2. *Availability of tools for corpus exploitation*

Corpora are of no use to the language teacher unless they are accompanied by user-friendly tools for their analysis and exploitation in the teaching and learning context. As we will see in more detail later, a number of computer programs, or varying degrees of sophistication and ease of use, are now readily available. Most of these will allow the rapid production of a number of useful forms of output: word lists, arranged alphabetically or numerically; concordances listing the occurrences of a given word form in one or more texts, with an amount of context which may be controllable by the user, and with the possibility of resorting the output by the items at various positions to the right or left of the word under study; statistics about word and sentence lengths, etc. Most will also allow at least some exploration of collocational patterning, the way in which words associate in running text. Some will also produce graphical plots of the distribution of a word or phrase in a text, and even an indication of which words are 'key' in a text, by comparison with a larger reference text or corpus.

2.3. *The trend towards pedagogical grammar*

One objection which some might raise to the use of corpora is that language learning is being approached through an analysis of the language, in clear contrast to the tenets of at least the more extreme forms of communicative language teaching (CLT). However, as Hadley (forthcoming) has pointed out, applied linguists have for some time been questioning the more extreme forms of CLT. Skehan (1996, 30), for example, observes that the tendency for CLT to stress fluency more than accuracy may well restrict learners to particular strategic solutions, inhibiting them from developing structurally and in terms of accuracy. Hadley detects a resurgence in interest in pedagogical grammars, predicated on the assumption that communicative and grammatical approaches are not necessarily mutually exclusive, but rather can complement each other in a productive manner. He sees corpus-based work as exemplifying one particularly powerful kind of pedagogical grammar which, while avoiding the excesses of grammar-translation or purely structural grammars, allows a closer integration to be achieved between knowledge about the language and ability to use it, and helps to promote a better balance between fluency and accuracy.

2.4. *Data-driven learning: the student as researcher*

One of the most important features of corpus-based work in language teaching and learning is that it follows the current trend towards a shift in the respective roles of teacher and learner. Increasingly, over the whole spectrum of academic areas, students are being expected to take more responsibility for their own learning, with the teacher acting as facilitator of learning ("the guide on the side"), rather than as all-knowing fount of knowledge ("the sage on the stage"). Attractive as this philosophy may seem initially to the teacher with dreams of sitting at the back of the classroom with arms folded, as the students beaver away at their tasks, it actually makes very great demands on the teacher, who must prepare the students very carefully for their tasks, and must always be willing to accept that s/he is, together with the students, facing the potentially unknown, rather than operating within a situation in which the teacher is in full control. The resulting feeling of insecurity is one of the main problems, for many teachers, with such an approach. Nevertheless, from the point of view of productive learning, this way of thinking has much to recommend it.

Firstly, students develop a sense of ownership of the knowledge gained, which is often not the case where they are expected simply to absorb knowledge

meted out by the teacher. This, in turn, will probably lead to greater retention of the knowledge. It is entirely possible for students to come up with novel findings about language, and this, for many students, provides a thrill and a sense of importance which are highly motivating.

Secondly, students can, if given appropriate corpora, work on texts, and on areas of language structure and function, which are of interest to them personally and are of relevance to their careers. Again, as I noted earlier, this is more likely to lead to a positive learning outcome than if the whole class is working on something which may be far removed from the interest and needs of particular students. Further, students can work at their own pace, concentrating on aspects which they find particularly interesting or difficult, and passing more lightly over easier or (to them) less fascinating features.

One of the most active proponents of corpus-assisted language teaching, Tim Johns of the University of Birmingham, UK, has characterised the student-as-researcher approach as one in which learning is driven by the data:

> The perception that "research is too serious to be left to the researchers": that the language learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data. (Johns 1991, 2)

As the term **data-driven learning** now has wide currency, I will use it, or the abbreviation DDL, in the remainder of this talk. I have no time to dwell further on this important concept, but would refer you to the following as a selection from the large range of discussion on this topic: Johns (1991, 1994), Tribble & Jones (1990/1997) Murison-Bowie (1993), Stevens (1995).

3. Problems and challenges

Teachers have long known that no method is a panacea for all ills, and that whatever the advantages of a particular approach, that same approach will also bring problems. So it is with data-driven learning of languages.

3.1. *The technical challenge*

Without doubt, one of the most serious difficulties in the implementation of the DDL approach is the apprehension which many teachers, and some of their students, experience when it is suggested that they might like to explore the use of computer-assisted corpus analysis in their work. It is still the case that many experienced EFL teachers, while they may feel quite comfortable with stand-alone CALL materials, have had little or no exposure to the kinds of exploratory activity involved in corpus analysis, or to the tools which enable such analysis to be carried out. Students, especially those of school age, are increasingly computer literate, but again few language students will be at all familiar with either the concepts or the technology involved in corpus-based study. Neither is this simply a question of fear of the computer as such: teachers and students may feel, as many also do about, for instance, linguistic and/or computational approaches to literary style, that such analysis is cold, mechanical and destructive, under-valuing the rich resource which we know as a language. Such attitudes are often hard to shake, and many will take some convincing that computer-based analysis is an appropriate path to take. In my experience, however, once bitten by the corpus bug, teachers and students alike (and in this area, we have seen that there may not be so much of a gap between them) tend to get hooked very easily.

What all this means, of course, is that there is a need to make language teachers and their students very much more aware of the possibilities afforded by corpus work, and to persuade them, preferably by example, that such work is not only fruitful but also not particularly complicated once you get used to it. Training is thus essential, and initiatives such as the free World Wide Web-based course in corpus linguistics offered by Lancaster University[3] and the

3. This course can be found at *http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm*

publication on the internet of Cathy Ball's tutorial notes on concordances and corpora[4] are very much to be welcomed. Even more useful would be a web-based course designed specifically for language teachers interested in DDL, though much can be learned from the examples which can be gathered from the literature, as well as from Tim Johns' web site.[5] Even if such resources become readily available, there is, of course, still the problem of the time in which to get up to speed with an unfamiliar approach. A useful account of the processes of corpus analysis and interpretation, and how to guide students through them, is given in Gavioli (1997).

3.2. *Not waving but drowning*

A frequent feeling among students exposed to a large range of corpus examples of a given linguistic phenomenon is that there is just too much data to handle. Hadley (forthcoming), for example, in an experimental course with Japanese learners of English, found that what seems at first like an advantage, the availability of large numbers of examples of authentic productions, in fact tended to seem overwhelming, although overall the students' reaction to the DDL approach was quite favourable. Clearly, a balance needs to be struck between providing enough data for valid generalisations to be made, but not so much that the students feel lost or overburdened. This is one argument, among several, for the use of smaller corpora than those standardly used in linguistic research as such, a point to which we will return later.

3.3. *Limitations of available corpora*

A further important problem is the degree of matching between the availability of corpora, and the needs of language teachers and learners. Most of the readily available corpora are of English, and while this is good news for EFL specialists, it does not help teachers and learners of other languages. This situation is, however, steadily improving, with the collection of corpora in several major world languages. A related problem is that even some of the corpora which currently exist are not generally available for academic use, usually for reasons of commercial sponsorship and ownership. And even when corpora are indeed available in principle, teachers may still, in many cases, lack access to funds to buy them or equipment on which to use them.

Even in the corpora which are genuinely available to teachers and researchers, there are often problems of balance and representativeness. It is obviously much cheaper to collect written material than to amass spoken text, which must be transcribed; for this reason, much more written than spoken language is available in corpus form. This is particularly unfortunate from the point of view of the many language teachers and learners whose primary interest is in spoken communication. Again, though, the situation is changing: as we will see later, recent corpora of English have sizeable spoken components: for instance, the British National Corpus (BNC), although it consists of 90% written material, does include 10 million words of spoken English, as does the 50 million word subset of the COBUILD Bank of English which is available online).[6] Questions of balance also extend to topic: some corpora (e.g. LOB, Brown, BNC), are carefully constructed to include particular proportions of material from specific varieties of language; others (e.g. the Birmingham "monitor" corpus) are deliberately more open-ended in their make-up. The issue of representativeness is a complex and thorny one, and I can do no more here than refer you to discussion of the topic in, for example, Biber (1993) as well as in the textbooks on corpus linguistics listed earlier.

A further issue relates to the types of information available in a corpus. Few corpora consist simply of raw text: most have at least some annotation

4. URL *http://www.georgetown.edu/cball/corpora/tutorial.html*
5. URL *http://web.bham.ac.uk/johnstf/timconc.htm*
6. This and other corpora mentioned here will be described briefly later.

labelling various parts of the corpus (e.g. identifying individual texts or text excerpts, speakers in a corpus of spoken material, pages or chapters in a novel, etc.). Increasingly, corpora are being annotated in more sophisticated ways, the most common being part of speech tagging (Lancaster-Oslo-Bergen [LOB], Brown, BNC, among others), though a few corpora with basic grammatical parsing are also now available (e.g. Lancaster Parsed Corpus, Lancaster-Leeds Treebank, Penn Treebank, SUSANNE Corpus). Such annotations are clearly of potential usefulness in language teaching an learning, though there are those who would object that a tagging system imposes a framework on the corpus from outside, rather than allowing categories to emerge from the corpus itself, and that dealing with a tagged corpus can lead to the overlooking of valuable subtleties in the interaction between grammar and vocabulary (see e.g. Tognini Bonelli (1996, 58-62). For a gentle but rather old introduction to corpus annotation, I refer you to Leech and Fligelstone (1992); a fuller discussion can be found in Garside, Leech & McEnery (1997).

There are, then, some important issues facing teachers who wish to make use of the large corpora standardly available. It has, however, been suggested that such corpora may not in any case be the most appropriate for exploitation in language teaching and learning. I mentioned earlier that Tribble (1997) has argued that corpus materials made available to students should be taken from genres and topics appropriate to their interests and needs. Some general purpose corpora may indeed provide suitable material: for instance, the written component of the British National Corpus contains material from various domains: imaginative, arts, belief and thought, commerce and finance, leisure, natural and pure science, applied science, social science, world affairs, unclassified (Aston & Burnard 1998, 29). Tribble's view, however, is that what learners really need is a modestly sized collection of 'expert performances' in the relevant genres, and that these can be put together by teachers from readily available sources such as multimedia encyclopaedias. Tribble provides convincing examples of how one such source can be exploited in the context of helping students who are beginning to write professionally oriented texts in formal English.

4. Sources of corpus materials

Details of standard corpora are given in the standard textbooks (see e.g. McEnery & Wilson 1996, 181-7; Biber et al 1998, 281-4), and lists of available corpora, with links to the appropriate sites, are also readily accessible on the internet.[7] I will therefore just mention a few of the most useful sources for English.

The 1 million word Brown Corpus of written American English, organised into 15 text categories, and the Lancaster-Oslo-Bergen Corpus, of parallel size and structure but for written British English, are readily available on the CD-ROM produced by the International Computer Archive of Modern English (ICAME) at the University of Bergen,[8] but are small by today's standards (though as we have seen, this may not be a serious disadvantage in the teaching and learning area) and, more importantly, represent the English of nearly 40 years ago. At the University of Freiburg corpora have now been produced which conform as exactly as possible to the model of Brown and LOB, but are taken from material published in the early 1990s.[9] The Freiburg-LOB and Freiburg-Brown corpora will be included on the new version of the ICAME CD-ROM to be published later in 1999.

A similar problem of ageing besets the London-Lund Corpus of spoken English, based on the Survey of English Usage and collected in the 1960s and early 70s. This corpus, also available on the ICAME CD-ROM, consists of half a million words, prosodically transcribed.

7. See, for example, *http://info.ox.ac.uk/bnc/corpora.html* at the University of Oxford, the Lancaster University information at *http://www.comp.lancs.ac.uk/computing/research/ucrel/corpora.html*, or the web site maintained by Michael Barlow at *http://www.ruf.rice.edu/~barlow/corpus.html*
8. URL *http://www.hs.uib.no/icame.html*
9. For further information, see *http://www.hd.uib.no/icame/flob/flobinfo.htm*

The British National Corpus consists of 100 million words (distributed, as we have seen, into 90% written and 10% spoken). The written component was collected according to domain (the 10 categories listed earlier), time (largely 1975 to 1993, but with some material from as far back as 1960), and medium (book, periodical, etc.). In the spoken material, there are roughly equal amounts of informal speech by a socially-stratified sample of speakers, and more formal language collected in meetings, from the radio, and so on.[10] The BNC comes with its own built-in software, SARA, for searching. Because of its size and the consequent demand on hardware, the BNC is at present largely housed on institutional servers rather than on individual stand-alone PCs. A CD-ROM containing a 2-million word sample of the corpus has been promised. A particularly useful facility, currently mounted on an experimental basis, is a search facility for words or phrases which gives 50 randomly selected examples, completely without charge, over the internet.[11]

The Bank of English, the corpus on which the COBUILD dictionaries and other materials are based, has now grown to enormous size (some 329 million words as of July 1998) and is a 'monitor' corpus, added to daily, and intended to reflect the mainstream of English today. A subset of it is available on subscription from COBUILD Online, and consists of 50 million words, tagged for part of speech, in 10 subcorpora, covering written English (newspapers, transcripts of broadcasts, etc) from the UK, the USA and Australia, and also 10-million words of spoken British English. The online service provides sophisticated concordancing facilities, and also the generation of collocations, the quantitative importance of which can be assessed by different statistical indicators. The *COBUILD on CD* CD-ROM contains examples from a 5 million word selection from the Bank of English, in addition to the Collins COBUILD English Language Dictionary, Collins COBUILD English Usage and Collins COBUILD English Grammar. COBUILD also produces a useful CD-ROM of collocations derived from the Bank of English.[12]

Mention should also be made of the International Corpus of English (ICE), which when complete will consists of one million words of English, spoken or written between 1990 and 1996, from each country or region in which English is a first or major language. The corpus will be part of speech tagged, and the UK subcorpus is now available.[13]

A 2-million word corpus of spoken professional American English, constructed from transcripts of academic meetings and White House press conferences, is available from Athelstan.[14]

Two useful corpora for teaching purposes are those originally marketed by Oxford University Press for use with their analyser MicroConcord (see later), each consisting of about one million words. One corpus is of articles in various topic domains, the other consists of material from the Independent and Independent on Sunday newspapers.[15]

For those interested in the history of English, the Helsinki Corpus, a collection of texts spanning the Old, Middle and Early Modern English periods and available on the ICAME CD-ROM, is invaluable.

Finally, it is worth emphasising that textual material in computer-readable form is now widely available in forms other than the organised collections we call corpora: CD-ROMs containing large quantities of text (e.g. literary works, newspapers[16]) are proliferating; there are a number of electronic text archives from which material can be obtained; large quantities of text are now also available through the internet.

10. For details see Aston and Burnard (1998:28-33), and for further information on purchasing the BNC visit the web site at *http://info.ox.ac.uk/bnc*
11. URL *http://thetis.bl.uk/lookup.html*
12. For further details of the Bank of English and other COBUILD products, visit *http://titania.collins.cobuild.co.uk/* and follow the appropriate links
13. For information, see *http://www.ucl.ac.uk/english-usage/*
14. URL *http://www.athel.com/*
15. These corpora are now available from Athelstan.
16. For an account of work in EFL using newspaper CD-ROMs as corpora, see Minugh (1997).

5. Software tools for corpus analysis

As we have just seen, some corpora, such as BNC, come with their own search software. In general, however, the user must select one or more software tools with which to analyse corpus material. Details of such tools are again available in the standard texts (see e.g. McEnery & Wilson 1996, 189-92, Biber et al 1998, 285-6) and from the web sites devoted to corpora mentioned earlier. So once more, I will confine myself to a few remarks about the most commonly used software packages.

Without doubt, the most comprehensive and useful readily available set of analysis tools is WordSmith Tools, written by Mike Scott of the University of Liverpool and marketed by Oxford University Press.[17] I will present examples using WordSmith Tools later in this talk. The program allows the production of word lists, concordances sorted in various ways, distribution plots, collocations, a range of text statistics, and also lists of words which are 'keywords' within any given text or group of texts, as judged by the high frequency relative to some larger reference corpus. Work by Tribble (1998) has demonstrated how extremely useful the keyword technique can be in the context of teaching to write within specific genres.

Also very useful is MonoConc for Windows,[18] which allows searching of several million words for words, parts of words or phrases, and the production of frequency lists and concordances with resorting facilities. Lists of collocates at positions one or two words to right or left of the headword can also be produced. A more advanced version, MonoConc Pro, intended for use in linguistic research, has recently been released. A version of MonoConc for the Apple Macintosh is also available.

A further useful tool is TACT, written at the University of Toronto, and available as freeware.[19] TACT produces frequency lists, concordances and distribution plots, and can assess the strength of collocations statistically. A disadvantage is that the user must first convert the raw corpus text into a specific database form using software provided with TACT. Ready indexed TACT databases are available for the LOB, Brown, London-Lund and Helsinki corpora are available on the ICAME CD-ROM.

Also popular is Wordcruncher,[20] with similar functionality to TACT, and again requiring the corpus to be converted to a special form before processing. The ICAME CD-ROM has Wordcruncher versions of the LOB, Brown and London-Lund corpora.

A program which has been available for some time now is Micro-OCP, the PC version of the Oxford Concordance Program for mainframe machines, marketed by Oxford University Press. It has a wide range of options, but is perhaps not so user-friendly as some of the other tools available.

MicroConcord, which is in many ways the predecessor of WordSmith Tools, was specifically written for use by teachers in the DDL environment, and provides quick and easy access to word counts and concordances, with some collocational information.[21]

A simple tool which has found some favour with ELT teachers (see e.g. Kettemann 1995) is the Longman Mini Concordancer, for use with texts of less than 50,000 words.[22]

17. Details can be obtained from Mike Scott's own web site at *http://www.liv.ac.uk/~ms2928/wordsmit.htm* or from the OUP site at *http://www1.oup.co.uk/elt/catalogu/multimed/*
18. URL *http://www.nol.net/?athel/mono.html*
19. From the TACT website at *http://www.chass.utoronto.ca:8080/cch/TACT/tact0.html*.
20. Contact Johnson and Company, P.O. Box 446, American Fork, UT 84003, USA
21. MicroConcord is no longer marketed by OUP, but is still available from Athelstan. See the URL in fn. 10
22. Contact Longman Group UK, Longman House, Burnt Mill, Harlow, Essex CN20 2JE, UK

Finally, for those who work with Apple Macintosh machines, Free Text Browser[23] and Conc[24] are available, in addition to MonoConc for Mac.

6. What areas of ELT have benefited from a corpus-based approach?

There is now ample evidence that a corpus-based approach can be useful in a wide range of areas within ELT, and across various levels of linguistic patterning. One of the most important insights to come out of corpus linguistics, especially in the work of Sinclair and his colleagues, is the complex interdependence of grammar and lexis. Words have their own grammatical patterning, which may be partially different from that of even inflectionally related forms. We often find, therefore, that work in which corpora are exploited for language learning enriches the student's insight into the behaviour of words in their co-texts and contexts of production, bringing in and integrating aspects which would traditionally be labelled as grammar, vocabulary (including collocational patterning), etc. For examples of such work, I refer you to the collections of papers in Johns & King (1991), Wilson & McEnery (1994), Botley et al (1996), and Wichmann et al (1997).

A further key feature to emerge from corpus studies is the importance, especially but by no means exclusively in spoken language, of multi-word sequences. Indeed, such is the qualitative and quantitative importance of such sequences that Sinclair has proposed that the traditional view of language, in which the structure of a stretch of language is viewed in terms of choice from the patterns allowed by the grammar, needs to be supplemented by a different model which he has dubbed the 'principle of idiom':

> The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments. (Sinclair 1991, 110)

Work on English and Spanish, reviewed in Butler (1997, 1998), strongly supports this view, and has clear implications for language teaching and learning (for discussion, see Butler, forthcoming). Examples of corpus-based work relevant to the teaching of phraseology can be found in Magee & Rundell (1996) and Gledhill (1996).

Corpus-based approaches have also been productive in teaching related to genre and other aspects of variety in language, particularly in relation to English for Special Purposes (ESP) and English for Academic Purposes (EAP). I have already mentioned the work of Tribble (1997) with students beginning to write professionally-oriented texts in particular genres; an account of the potential of corpora in the teaching of academic writing, with particular reference to dissertations, can be found in Carne (1996). J Flowerdew (1993) has provided instructive examples of how the study of corpora based on the language to which learners will be exposed can help in the design of an ESP course for Arabic-speaking biology students. Flowerdew, like Tribble, emphasises that a corpus made up of texts specific to a particular field is usually of much greater utility in ESP work than a general English corpus. He provides a useful list of specialised ESP corpora developed for various applications (J Flowerdew 1996, 101).

Studies of literary style and critical literary appreciation can also benefit from a corpus orientation, as has been demonstrated in the work of Kettemann (1994), Jackson (1997), Louw (1997) and Tribble (forthcoming).

Finally, corpora such as the Helsinki Corpus are of inestimable value in the study of the history of the language, as is the CD-ROM version of the Oxford English Dictionary (see e.g. Facchinetti 1996, Knowles 1997).

23. Contact ICAME.
24. Contact International Academic Bookstore, Summer Institute of Linguistics, 7500 W. Camp Wisdom Road, Dallas, TX 75236, USA, or visit the SIL web site.

7. Bi- and multilingual corpora: parallel concordancing

An area which has seen considerable expansion very recently is the production and exploitation of corpora containing material from more than one language.[25] A distinction is often made between **parallel** corpora, which consist of sets of translations of a text in some source language, and **comparable** corpora, consisting of texts in different languages which are not translationally equivalent in any sense, yet do have a common communicative function. Peters et al (1996, 69) point out that both are of use to the language learner: parallel corpora are of interest to the average student of a second or foreign language because they provide data on different ways in which a particular word or construction may be translated into the foreign language; while comparable corpora are of more use to advanced students, particularly those with an interest in languages for special purposes. The ability of concordances to show, at a glance, multiple instances of the translation of a particular word or structure provides a particularly attractive tool for the learner, as has been demonstrated, for example, by Barlow (1996) in relation to the translation of English reflexive forms into French.

Analytical tools have recently been developed for the automatic alignment of parallel texts and for the production of parallel concordances, i.e. concordances giving not only the source language word in its various contexts, but also the translation in each of these contexts. Two such tools are Multiconcord, developed at the University of Birmingham as part of the Lingua project (see King & Woolls 1996), and ParaConc, developed by Michael Barlow (Barlow 1995a, 1995b).[26]

One problem with parallel concordancing is the current shortage of suitable parallel texts, though the situation is fast improving. MultiConcord comes with a small set of texts consisting of proceedings in the European Parliament, in English, French, German, Spanish and Portuguese. Other European Parliament documents can be downloaded from the project's Parallel Texts Library.[27] Parallel texts on topics concerned with health matters in English/French and English/Spanish can be downloaded from the World Health Organisation site.[28] Other parallel texts are available for purchase from the European Language Resources Association (ELRA)[29] and from the Linguistic Data Consortium (LDC).[30]

8. Learner corpora

A further recent development is the collection of corpora of productions by language learners. Foremost among these projects is that concerned with the International Corpus of Learner English (ICLE), which has been in progress since 1990 at the Centre for English Corpus Linguistics at the Université Catholique de Louvain, Belgium, under the leadership of Sylviane Granger.[31] The ICLE forms part of the British section of the International Corpus of English mentioned earlier, and contains over a million words of English written by learners from 11 different language backgrounds. Work based on ICLE, together with other work on learner corpora, can be found in Granger (1998).

Others working on learner corpora include Kojiro Asao and colleagues at Tokai University, Japan on a corpus of English by Japanese learners,[32] Gui

25. For details see Michael Barlow's web page at URL *http://www.ruf.rice.edu/~barlow/para.html*
26. For more on Multiconcord, including details of ordering, see *http://sun1.bham.ac.uk/johnstf/l_text.htm* and the links from that page, and for ParaConc see *http://www.ruf.rice.edu/~barlow/parac.html*.
27. URL *http://web.bham.ac.uk/johnstf/multdata.htm*.
28. URL *http://www-pll.who.ch/programmes/pll/cat/cat_resources.html*.
29. URL *http://www.icp.grenet.fr/ELRA/cata/tabtext.html*.
30. URL *http://www.cis.upenn.edu/~ldc*.
31. For an introduction to the project, see *http:///www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/cecl.html*. A useful bibliography of work on learner corpora is also available from this site.
32. For details, together with links to other learner corpora sites, see *http://www.lb.u-tokai.ac.jp/lcorpus/*.

Shichun on Corpus-Based Analysis of Chinese Learner English (CBACLE) at the Guangdong University of Foreign Studies, Guangzhou, and John Milton of the Hong Kong University of Science and Technology on a corpus of writing of more than 10 million words by Cantonese-speaking Hong Kong students (see Milton 1996). L Flowerdew (1998a, 1998b) has compared cause and effect markers in a 40,000-word subsection of the Hong Kong corpus with those in a similarly sized corpus of learner assignments, and has been able to generate results which can be used to inform course materials.

9. Conclusion

In this brief paper, I hope to have given some idea of the potential of corpus-based work in language teaching and learning, and to have indicated the range of the exciting projects which are now in hand in this area. I hope that some readers who teach EFL, but are not at present using corpora as one of the weapons in their armoury, will want to explore further the extensive literature on the topic, and to try some of the techniques for themselves.

Works cited

ASTON, Guy & Lou BURNARD. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh: Edinburgh UP, 1998.

BARLOW, Michael *A Guide to ParaConc*. Houston: Athelstan, 1995a.

—ParaConc: "A concordancer for parallel texts." *Computers and Texts* 10 (1995b): 14-16.

—"Parallel texts in language teaching." Botley *et al.*, 1996. 45-56.

BAZERMAN, Charles. *Constructing Experience*. Carbondale: Southern Illinois UP, 1994.

BIBER, Douglas. "Representativeness in corpus design." *Literary and Linguistic Computing* 8 (1993): 243-257.

BIBER, Douglas, Susan CONRAD & Randi REPPEN. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge UP, 1998.

BOTLEY, Simon, Julia GLASS, Tony MCENERY & Andrew WILSON, eds, *Proceedings of Teaching and Language Corpora 1996*. Lancaster: UCREL, Lancaster U, 1996.

BUTLER, Christopher S. "Repeated word combinations in spoken and written text: some implications for Functional Grammar." *A Fund of Ideas: Recent Developments in Functional Grammar*. Eds. Christopher S. Butler, John H. Connolly, Richard A. Gatward & Roel M. Vismans. Amsterdam: IFOTT, U of Amsterdam, 1997. 60-77.

—"Collocational frameworks in Spanish." *International Journal of Corpus Linguistics* 3.1 (1998): 1-32.

CARNE, Chris. "Corpora, genre analysis and dissertation writing: an evaluation of the potential of corpus-based techniques in the study of academic writing." Botley *et al.*, 1996. 127-137.

FACHINETTI, Roberta. "The exploration of diachronic English software by foreign language students." Botley *et al.*, 1996. 150-159.

FLIGELSTONE, Steve. "Some reflections on the question of teaching, from a corpus linguistics perspective." *ICAME Journal* 17 (1993): 97-109.

FLOWERDEW, John. "Concordancing as a tool in course design." *System* 21.2 (1993): 231-244.

—"Concordancing in language learning." *The Power of CALL*. Ed. Martha Pennington. Houston: Athelstan, 1996. 97-113.

FLOWERDEW, Lynne. "Concordancing on an expert and learner corpus for ESP." *CAELL Journal* Spring 1998 (1998a): 3-7.

—"Integrating 'expert' and 'interlanguage' computer corpora findings on causality: discoveries for teachers and students." *English for Specific Purposes* 17.4 (1998b): 329-345.

GARSIDE, Roger, Geoffrey Leech & Tony McEnery, eds. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman, 1997.

GAVIOLI, Laura. "Exploring texts through the concordancer: guiding the learner." Wichmann *et al.*, 1997. 83-99.

GLEDHILL, Chris. "Science as collocation: phraseology in cancer research articles." Botley *et al.*, 1996. 108-126.

GRANGER, Sylviane, ed. *Learner English on Computer*. London and New York: Addison Wesley Longman, 1998.

HADLEY, Gregory. "Sensing the winds of change: an introduction to data-driven learning." To appear in *Insights* 2.

JACKSON, Howard. "Corpus and concordance: finding out about style." Wichmann *et al.*, 1997. 224-239.

JOHNS, Tim. "Should you be persuaded – two examples of data-driven learning materials." Johns & King, 1991. 1-13.

—"From printout to handout: grammar and vocabulary teaching in the context of data-driven learning." *Approaches to Pedagogic Grammar*. Ed. T. Odlin. Cambridge: Cambridge UP, 1994. 293-313.

JOHNS, Tim. & Philip KING, eds. *Classroom Concordancing*. Birmingham English Language Research Journal 4. Birmingham: U of Birmingham, 1991.

KENNEDY, Graeme. *An Introduction to Corpus Linguistics*. London & New York: Addison Wesley Longman, 1998.

KETTEMANN, Bernhard. "Concordancing in stylistics teaching." *Festschrift zum 30-jährigen Bestehen des Instituts fur Anglistik der Universität Salzburg*, Salzburg, 1994.

—"On the use of concordancing in ELT." *TELL&CALL* 1995.4, (1995): 4-15.

KING, Philip & David Woolls. "Creating and using a multilingual parallel concordancer." *Translation and Meaning* Part 4 (1996): 459-466.

KNOWLES, Gerry. "Using corpora for the diachronic study of English." Wichmann *et al.*, 1997. 195-210.

LEECH, Geoffrey. "Teaching and language corpora: a convergence." Wichmann *et al.*, 1997. 1-23.

LEECH, Geoffrey & Steven Fligelstone. "Computers and corpus analysis." *Computers and Written Texts*. Ed. Christopher S. Butler. Oxford: Blackwell, 1992. 115-140.

LOUW, Bill. "The role of corpora in critical literary appreciation." Wichmann *et al.*, 1997. 240-251.

MAGEE, Stephen & Michael RUNDELL. "The role of the corpus-based 'phrasicon' in English Language Teaching." Botley *et al.*, 1996. 17-28.

MCENERY, Tony & Andrew WILSON. *Corpus Linguistics*. Edinburgh: Edinburgh UP, 1996.

MILTON, John. "Exploiting L1 and L2 corpora for computer assisted language learning design: the role of an interactive hypertext grammar." Botley *et al.*, 1996. 233-243.

MINUGH, David. "All the language that's fit to print: using British and American newspaper CD-ROMs as corpora." Wichmann *et al.*, 1997. 67-82.

MURISON-BOWIE, Simon. *MicroConcord Manual: An Introduction to the Practices and Principles of Concordancing in Language Teaching*. Oxford: Oxford UP, 1993.

PARTINGTON, Alan. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam and Philadelphia: John Benjamins, 1998.

PETERS, Carol, Eugenio Picchi & Lisa Biagini. "Parallel and comparable bilingual corpora in language teaching and learning." Botley *et al.*, 1996. 68-82.

SINCLAIR, John M. *Corpus, Concordance, Collocation*. Oxford: Oxford UP, 1991.

—"Shared knowledge." *Linguistics and Language Pedagogy: The State of the Art*. Ed. J. E. Alatis. Washington D.C.: Georgetown UP, 1992. 489-500.

—"Corpus evidence in language description." Wichmann *et al.*, 1997. 27-39.

SKEHAN, Peter. "Second language acquisition research and task-based instruction." *Challenge and Change in Language Teaching*. Eds. Jane Willis & David Willis. Oxford: Heinemann, 1996. 31-41.

STEVENS, Vance. "Concordancing with language learners: Why? When? What?" *CAELL Journal* 6.2 (1995): 2-10.

STUBBS, Michael. *Text and Corpus Analysis*. Oxford: Blackwell, 1996.

TOGNINI BONELLI, Elena. *Corpus Theory and Practice*. Birmingham: TWC, 1996.

TRIBBLE, Christopher. "Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching." *PALC '97 Proceedings*. Eds. J. Melia & B. Lewandowska-Tomaszczyk. £ódz: U of £ódz, 1997.

—"Genre, keywords, teaching: towards a pedagogical account of the language of project proposals." Paper presented at TALC 98, 1998.

—"Counting things in texts you can't count on: a study of Samuel Beckett's Texts for Nothing 1."

Proceedings of the Linguistics of Literature Seminar 1996. Ed. D. Malcolm. Gdansk: U of Gdansk P, forthcoming.

TRIBBLE, Christopher & Glyn JONES. *Concordances in the Classroom: Using Corpora in Language Education*. London: Longman, 1990. New ed. Houston, TX: Athelstan, 1997.

WICHMANN, Anne, Steven FLIGELSTONE, Tony MCENERY & Gerry KNOWLES, eds. *Teaching and Language Corpora*. London & New York: Addison Wesley Longman, 1997.

WIDDOWSON, Henry G. *Learning Purpose and Language Use*. Oxford: Oxford UP, 1983.

WILSON, Andrew & Tony MCENERY. *Corpora in Language Education and Research: A Selection of Papers from TALC94*. UCREL Technical Papers 4. Lancaster: UCREL, Lancaster U, 1994.