ASSESSING THE CONTRIBUTION OF CORPORA TO EAP PRACTICE

Dr Paul Thompson
Department of Applied Linguistics, University of Reading
p.a.thompson@reading.ac.uk

Abstract

This paper reviews the uses of corpora in EAP, and assesses the contributions that corpus-based studies have made both to the general understanding of academic discourse, and to EAP teaching practice. Firstly, the various EAP-related corpora, and recent trends in corpus-based research, are described. Secondly, the issues involved in the direct application of corpus findings, and corpus analysis methodologies, to EAP classroom teaching, are discussed. It is observed that the direct application of corpus methodologies and corpus evidence in the classroom remains as yet limited. This is due in large part to a lack of familiarity amongst practitioners with 1) the methods of corpus analysis; and 2) with the criteria by which insights afforded by corpus-based linguistic research should be evaluated. The paper concludes with a set of proposals for taking the use of corpora in EAP practice forward.

1. Introduction

Dudley-Evans and St John (1998:4-5) define the absolute characteristics of ESP as being that it is “designed to meet the specific needs of the learner” and that it makes use of the “underlying methodology and activities of the disciplines it serves”. They go on to say that “ESP is centred on the language (grammar, lexis, register), skills, discourse and genres appropriate to these activities”. For the ESP practitioner, or learner, to be able to identify the linguistic features or generic elements of those communicative activities, it is clearly beneficial to have access to exemplars of those activities. The exemplars, be they written texts or transcripts of oral interactions, form a body of evidence upon which analyses can be conducted.

EAP can be seen to be a subset of ESP, and to have the same absolute characteristics that Dudley-Evans and St John propose. Corpora (collections of texts) can act as the bodies of evidence upon which analysis of linguistic, generic and discourse features can be conducted, and the purpose of this paper is 1) to review the contributions that corpus-based analysis has had on EAP practice, both from a research point of view and from a pedagogical perspective (the uses made of corpora in teaching and learning), and 2) to discuss the potential for further development of corpus-informed instruction.

A recent collection of papers entitled “Academic Discourse” (Flowerdew 2002) provides evidence of the mainstream role that corpora have come to play in current approaches to research into academic discourse. The book is divided into four parts: genre analysis studies, corpus-based studies, contrastive rhetoric work, and ethnographic approaches. Two out of the three genre studies make use of corpora, three out the four contrastive rhetoric, and, of course, all the chapters in the corpus-based studies section either are based on, or review, uses of corpora. It should be noted, however, that the notion of a “corpus” is variously understood: for several writers in this collection, a corpus is simply a collection of texts, or text samples, in any format, while for others it is a collection of texts that is held in electronic format.
In both cases, however, a corpus is a source of evidence for investigations into language use in particular contexts. In this paper, the focus is on the latter type of corpus, the electronically-held database.

This paper assesses the contributions that corpus-based approaches have made to EAP practice. Practice is taken in a broad sense, to encompass both direct teaching and also work that feeds into classroom teaching, such as materials development, syllabus description, and research into academic discourse. Our primary observation, that it is becoming increasingly common to make use of corpora in research into academic discourse, leads us to question the contribution of this research to what happens in EAP teaching, and also to ask whether corpora can make a useful contribution to EAP pedagogy. The editors of the English for Specific Purposes Journal predicted in 2000 (Volume 19, Issue 2) that, “although the successful use of corpora in classroom teaching of ESP work has been reported … it may well turn out that the most extensive and valuable use of computer-based corpora will be for ESP research purposes rather than for teaching purposes”. The review of literature that follows in Sections 3, 4 and 5 supports this assertion, but the discussion in Section 6 below suggests that this may simply be a consequence of limited access to corpora, and of limited skills and knowledge amongst EAP practitioners.

This paper presents a survey of the corpora available, some of the uses that have been made of the resources, and then examines the current direct uses of corpora and concordancing in EAP teaching. In the space available, it is not possible to give a comprehensive review of a vast and ever-growing body of literature, but it is intended that most of the major issues and initiatives will be touched on.

First, though, what is meant by “English for Academic Purposes”? As stated above, EAP can be seen to be a subset of English for Specific Purposes, a branch of English Language Teaching that is concerned with the teaching of English for students with specific goals. Those specific goals might be the needs of, say, police in a non-English-speaking country to learn sufficient English to cope with visiting football fans, or it might be immigrant health workers learning enough English to perform their jobs in an English-speaking country. EAP is concerned with the teaching of English for use in academic contexts, to students for whom English is an additional language, and who are preparing to begin a course of academic studies, or who are currently engaged on a course. In the UK, these are referred to as pre-sessional and in-sessional EAP courses. A second useful distinction is that made between English for General Academic Purposes (EGAP), and English for Specific Purposes (ESAP) (Blue 1988). Students studying on summer pre-sessional courses in the UK are often placed in mixed subject classes, with, for example, economics and engineering students working together, as the classes are created according to language proficiency level rather than subject, and this comes under the heading of EGAP, as the language taught is not specific to a particular subject area; in-sessional support within a department, on the other hand, such as a language teacher working in collaboration with a Highway Engineering lecturer (Dudley- Evans and Johns 1981), and dealing with subject-specific language demands, is an example of ESAP. These distinctions will be taken up in the following section in which the various types of EAP corpus are discussed.
2 EAP Corpora

A corpus can be defined as “a body of written text or transcribed speech which can serve as a basis for linguistic analysis or description” (Kennedy 1998: 1). A corpus should also act, to some degree, as a sample of a particular kind of discourse, or variety of discourse types, as Francis (1982:7) stresses: “a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis”. In EAP, as in ESP in general, corpora can constitute highly valuable sources of evidence about the kinds of texts or spoken genres that the students are likely to either have to produce themselves in the future, or to have to be able to understand, provided that the corpora consulted are reasonably representative of those texts or genres.

![Figure 1 The corpus ring, showing the different people that engage with corpora in the EAP context, with examples of the corpora developed, of the concepts derived, or of the texts produced.](image)

In language teaching contexts, a corpus can be used by a range of users, as shown in Figure 1. In the upper half of the circle, there are the people furthest removed from the classroom itself, such as the lexicographers, grammarians and language testers, who will have an influence on pedagogy through their descriptions of language, and through the specification of aspects of language knowledge and performance that are to be the subject of assessments, while those involved in direct teaching (and learning) are in the lower half. All of these users may have occasion to examine the data contained in the corpus, and they will have different aims and will follow markedly different approaches.

Section 3 below discusses some of the corpus-informed research work that has been done, that relates to the upper half of the circle, with the researcher included. Sections
4 and 5 look at the lower half of the circle, investigating direct uses of corpora in EAP teaching and learning, and also the uptake of corpora by EAP practitioners.

At the centre of the circle is the corpus (plural corpora) and there is a wide range of corpora that can be exploited in EAP work. One approach to the categorization of EAP corpora is as follows:

- **academic subsection of general corpora** (such as the British National Corpus)
- **EGAP corpora** (eg, the MicroConcord B Corpus, JDEST)
- **ESAP corpora** (eg, the Hyland corpus, and the Reading Academic Text corpus)
- **learner corpora** (eg, ICLE and HKUST Learner Corpora)

In large corpora which aim to represent the language in general, there is typically a section of the corpus that consists of texts from academic genres. In the 100 million token British National Corpus, for example, according to David Lee’s BNC World Index ([http://elixir.davidlee00](http://elixir.davidlee00)), there are five hundred written academic texts, running to nearly 15.8 million tokens. It should be noted too that in a large general corpus, these texts are often not complete texts, as the corpus design may dictate that lengthy texts are merely sampled; the relevance of this observation is that this makes full text analysis impossible, and thus restricts the range of analyses that can be conducted (eg, it is not possible to perform a full text genre analysis on such a text). There are also 31 lecture transcripts which amount to 300,000 tokens, and eighteen university tutorials and seminars, at 143,000 tokens. There are problems with representativeness, as ten of the lectures, for example, are Psychology lectures, of which 5 appear to have been delivered by the same lecturer; similarly, there are three economics lectures from one university. A general corpus, however, does not claim representativeness at the level of genre, and it should not be taken as a criticism of the BNC that it does not offer adequate data for the investigation of academic lectures, across disciplines, as this is not one of the aims of the corpus (cf, Aston 2001). It can however give an indication of the linguistic features of discourse in different domains, different genres, which will allow the analyst to form hypotheses which can be tested on different datasets.

The Longman Spoken and Written English corpus ([Biber et al. 1999](#)) contains just over 40 million tokens, of which academic prose texts contribute 530,000, in 408 texts. From this data, the researchers on the Longman Grammar project were able to extract statistics on the frequency of occurrence of lexico-grammatical patterns in what is referred to as the ‘register’ of academic prose, in comparison with the registers of conversation, fiction, and news. While the BNC is openly accessible to limited degrees through a web interface, or can be purchased by educational institutions, the LSWE corpus is restricted to use in a small number of research projects. In addition to the four-way distinction between types of corpus, then, a second differentiation can be made in terms of how accessible the corpus is, a point that will be returned to below.

The first EAP corpus in the world, an example of an EGAP corpus, was the JDEST corpus, constructed by Jiao Tong University in Shanghai in 1985 ([Yang 1986](#)). It concentrated on language in science and technology, and was constructed along the lines of the so-called first-generation corpora (eg, the Brown and LOB corpora) with approximately one million tokens, and employing a sampling approach to corpus...
construction: the original JDEST corpus comprised ten sections, each made up of two hundred 500-token samples of texts (the JDEST corpus has now been expanded to 4 million tokens).

One of the earliest publicly-available concordancers (a computer programme which retrieves from a corpus all examples of a given search term, each within a specified span of co-text) used in language teaching was MicroConcord (Scott and Johns 1993) and this was sold with a choice of two corpora: one, a general corpus (Corpus A), and, the other, a corpus of academic prose (Corpus B). Corpus B contains five 200,000 word mini-corpora of texts from different knowledge domains. Corpus B, as with the majority of EAP corpora, is minimally tagged; that is, there is very little annotation of the text contained within the corpus. In contrast, each token in the JDEST corpus is tagged for part of speech. A third distinction that can be made between types of corpora, therefore, is that of whether or not the text is tagged in any way, and also to what degree it is annotated.

The last fifteen years or so have seen an increasing tendency among researchers to study the language practices of specific disciplines, and with this trend there has been a growth in the number of ESAP corpora developed for particular research projects. Hyland, for his research into disciplinary discourses (Hyland 2000; 2002) has compiled a corpus of texts, from a variety of genres, from eight disciplines: Mechanical Engineering, Electronic Engineering, Physics, Biology, Marketing, Applied Linguistics, Sociology and Philosophy. His initial corpus was developed to represent genres but as his interest progressed to disciplinary variation, the corpus has evolved to represent discourse in eight different disciplines. This corpus has also been used by John Swales in the revision of his “Academic Writing for Graduate Students”, co-authored with Christine Feak (Swales and Feak 2004), and in the writing of Swales 2005. Another trend in recent years has been the development of corpora of student writing, as distinct from writing by ‘expert’ writers, such as textbooks and research articles, because, it is argued, learners need to work on data that is drawn from texts that are analogous to those that they are expected to produce, and researchers also need to analyse student texts. An example of such a corpus is the Reading Academic Text corpus which contains 35 theses written by native speaker students, from a range of disciplines (information available at http://www.rdg.ac.uk/app_ling/corpus.htm). One of the aims of the corpus developers was to create resources for contrastive analyses of writing practices at the doctoral level across disciplines. Both the Hyland and the Reading corpora aim to characterize genres as well, and they contain complete texts. This is in contrast with the partial text approach of the JDEST corpus, for example, where all text samples of the same number of words, and do not constitute full texts in any case. A fourth distinction, then, can be made between corpora that contain full texts, and those that contain samples of texts; the former is supportive of text linguistic analysis and genre studies, while the latter is designed for statistical sampling.

ESAP corpora can also be restricted to a single discipline, and one genre thereof. Gledhill (2000), for example, built a corpus of research articles in pharmaceutical science for his investigation of the discourse function of collocation in research article introductions. Marco (1999) established a corpus of 100 research articles from the New England Journal of Medicine and the British Medical Journal.
Learner corpora are collections of writings and transcripts of spoken data produced by learners of English. The International Corpus of Learner English (ICLE) established by Sylviane Granger and her colleagues at the Centre for English Corpus Linguistics, in Belgium, has developed into a truly international project with a range of sections compiled by researchers in five other countries increasing the variety of language backgrounds represented in the corpus. A comprehensive listing of learner corpora worldwide has been compiled by Yukio Tono (http://leo.meikai.ac.jp/~tono/). Learner corpora tend to be used primarily for contrastive analysis purposes: lexico-grammatical features of learner discourse are contrasted with those in similar performances by native speakers. Points of contention surrounding this work are that of whether the different corpora are strictly comparable, the degree to which the native speaker writings can be considered to be examples of good writing, and also whether or not native speaker writing should be considered the norm.

As L. Flowerdew (2002) has noted, there is a heavy imbalance in representation in the majority of academic corpora between spoken and written data. The figures for the BNC, for example, show that the ratio of written to spoken in the academic sections of the corpus is nearly 36:1. Much of this is, of course, due to the fact that it is time-consuming and costly to transcribe spoken discourse, in comparison to the relative simplicity of scanning published documents or of making use of pre-digitized texts. This imbalance is to be remedied to some extent through the development of three new corpora: the T2K-SWAL and MICASE corpora in North America, and the BASE corpus in the UK. The TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) corpus is a project sponsored by the ETS, which suggests that access to the corpus will be restricted, and the aim is to collect 2.8 million tokens of spoken and written text, taken from 6 academic disciplines at four US universities. The corpus is fully marked up for grammatical features using Biber’s (1988) tagger, and contains 174 academic lectures, amounting to approximately 1.2 million tokens (Csomay 2002). Though access to the corpus may remain restricted, the corpus should generate considerable numbers of studies to the general research literature on academic speaking.

A project that began at roughly the same time in the late nineties in the US is the MICASE corpus (Simpson et al. 2000). One of the main aims of this project is to provide an openly accessible research resource. The entire corpus of transcripts can be searched online (http://micase.umdl.umich.edu/m/micase/) through a customized interface that allows searches by a range of parameters (age range, gender, academic position, discourse mode, and so on). A similar project currently being conducted in Britain is the BASE (British Academic Spoken English) corpus (Nesi 2001; Thompson forthcoming). Whereas MICASE features transcripts of a wide variety of academic speech events, such as office hour sessions, service encounters, in addition to lectures and seminars, BASE concentrates on the latter two only, and contains 160 lectures and 40 seminars, equally divided between four broad domains of academic enquiry: arts and humanities; life sciences; physical sciences; social sciences. A large proportion of the recordings have been made on video and the video data will be made available, upon request, in addition to the transcripts.

A key aspect of both the MICASE and the BASE is that the researchers have obtained the permission of all participants to make the corpus available publicly. Copyright and data protection makes it extremely difficult in most cases to make corpus data openly
accessible; the first edition of the BNC, for example, could only be distributed within the European Union, and when the World Edition was released for distribution worldwide a number of texts in the corpus had to be removed because copyright clearance could not be obtained. The majority of corpora mentioned in this section are restricted to different degrees in terms of how freely people outside the research project are able to access the corpus, and in most cast cases, there is no such opportunity.

A feature of ESP work is that the needs of learners are often highly specific and teachers may have to develop materials and resources for the specific context. Tribble (1997) has described ‘quick-and-dirty’ means for practitioners to develop their own small corpora for teaching purposes; further discussion of exploitation of small corpora for teaching purposes can be found in Ghadessy et al (2001). Such corpora are not necessarily documented in a public research literature, and so it is difficult to ascertain how many there are, and it is also likely that many of these will be ephemeral – corpora that are devised for a particular set of purposes, and that may then become redundant once those purposes have been achieved. The same can be said for many of the corpora that are developed for individual research purposes: they are compiled following relatively idiosyncratic procedures, and are often not made available to other researchers for further research or for replication studies.

3. Corpus-based EAP research

L. Flowerdew (2002) has observed a shift in corpus-based research into academic discourse from the study of academic register, at the lexicogrammatical level, which is chiefly statistically driven, to a range of studies at a discoursal or genre level, which focus more on patterning, function and phraseology. The early studies of register (Barber 1962/1988, for example) were based on analysis of paper-based collections of texts, but Biber’s work has been conducted on large scale statistical analysis of electronic corpora (Biber 1988). Early corpora encouraged the description of the lexico-grammatical features of an academic register as they were composed of samples of text, rather than full texts (see comment above on the JDEST corpus). The investigation of an academic register continues, as in the research work conducted for the Longman Grammar, but the growth of discourse and genre studies since the 1980s has led to the development of corpora that contain whole texts, or complete comparable rhetorical sections of texts. Thompson (2000) tagged and then quantified citation types in different sections of PhD theses, and Bunton (1998) examined metadiscourse in the Introductions and Conclusions of theses. L. Flowerdew (1998) has argued for an approach to corpus analysis that incorporates a textlinguistic perspective, and Upton and Connor (2001) have utilized corpus analysis methods in the study of discourse moves.

A characteristic of work in ESP has been the use of specialist informants. The EAP teacher or researcher is in most cases not an expert in the subject(s) that the learners are specializing in, and the teacher therefore may choose to consult a specialist informant about disciplinary conventions and concepts. Both Hyland and Gledhill in the studies cited above interview subject specialists to complement, or confirm, their analyses. This combination of corpus analysis with qualitative data gained from interviews and observation is a distinctive feature of recent work in EAP research.
As stated above, corpora can be investigated by a wide range of parties interested in EAP. Both the MICASE and T2K-SWAL corpora have been funded to large degrees by language testing organizations, and another recent development is the UCLES/CUP Learner corpus (Boyle and Booth 2000), which contains over 15 million tokens of language produced by foreign language learners in UCLES test responses. Roughly a quarter of the corpus has been tagged, using an in-house Learner Error Coding system. These corpora will be used by the testing organizations to examine learner performances, to characterize the linguistic and pragmatic features of learners at different stages of proficiency, for example, as well as using corpora as sources of language data for item writing.

The Collins COBUILD project set a new standard for lexicographers, based as it was on investigation of the Bank of English, and it is now commonplace for all major dictionaries to be based on corpus study. As yet, there is no dedicated corpus-based dictionary of academic language, but the ELT dictionaries that are used in EAP classrooms are a sign of the influence that corpus-based lexicographical research has had on EAP teaching practice.

Again, it was the COBUILD Grammar that introduced a new approach to the description of the grammar of the English language, an approach which has been adopted and extended by the LGSWE. The latter, as discussed above, devotes considerable space to the description of the language features of a general academic register, and, while it is still too early to show evidence, this project is likely to have a powerful influence on the teaching of English for General Academic Purposes.

A common use of corpora is to generate wordlists, based on frequency counts. This is clearly of value both to syllabus writers and to materials developers. For the syllabus writer, the list provides a useful reference point for specifying the lexical component of a syllabus, although this information should be balanced by consideration of factors such as coverage, range, and learnability (White 1988). For the textbook writer, similarly, the list suggests which lexical items need to be focused on, and also acts as a starting point for investigation of the lexico-grammatical behaviour of these items. The most influential of academic wordlists have been the University Word List (Nation 1990), and the Academic Word List (Coxhead 2000) which has superseded the UWL. These two lists both deal with the lexis that is most frequent, beyond the two thousand items described in West’s General Service List (West 1953). The Academic Word List is derived from extensive analysis of lexical frequency and range statistics of the 3.5 million token corpus that Coxhead created for the purposes of the study. This corpus comprised extracts from journal articles, textbooks, manuals and course notes, divided into four broad faculty divisions. It should be noted that the word list therefore represents the language of academic prose. Thompson (forthcoming) examines the lexis of academic lectures, using the BASE corpus, and observes that a reduced set of the Academic Word List consisting of 230 word families, rather than the 570 that Coxhead identifies, is an adequate target for learning purposes for the task of listening to academic lectures.

The link between corpus analysis, wordlists and materials development can be see in the work of a group at the English Centre, University of Hong Kong, on a project entitled “Vocabulary for specific disciplines” (Bruce and Storey 2002). The disciplines under investigation are Law, Social Work, Business/Economics,
Engineering and Medicine, and corpora of approximately half a million tokens each have been created for each discipline. Each of these corpora is made up of first year textbooks, academic articles and some general newspaper/magazine articles, and these are used to compile word frequency lists, using Paul Nation’s Range program (http://www.vuw.ac.nz/lals/staff/paul-nation/RANGE32.zip), which then inform a series of vocabulary learning exercises. A similar project is being undertaken at Bilkent University, Turkey, where the first year textbooks for a range of courses have been entered into a corpus which is the basis for specification of the English Language course, and there are likely to be many other such projects at other EAP institutions in other countries.

Looking at descriptions beyond the word, we find a growing set of studies of what have been described as “lexical phrases” (Nattinger and DeCarrico 1992) or “lexical bundles” (Biber et al. 1999). Oakey (2002) used the MicroConcord B corpus and the academic texts in the BNC to investigate actual uses of various lexical phrases, and to contrast the evidence of the corpora with the descriptions found in EAP textbooks and in Nattinger and DeCarrico’s study. Siepmann (Siepmann 2001) constructed a trilingual corpus (English, German and French) and conducted contrastive analyses of the uses of lexical phrases across the three languages. He is now working on a set of textbooks for academic writing in each of the three languages, based upon a phraseological approach.

Marco (1999) prefers the term “collocational frameworks”, and follows Gledhill and Hyland in adopting a genre-based approach. The corpus investigated is a collection of medical research papers, and the collocational frameworks examined are related to functions that are suggested to be central to the discourse of medical reports, such as the pattern “be X to”, shown to express relation (is related to; is similar to), and other frameworks are found to be characteristic of certain rhetorical moves (for example, “determine/estimate the number/degree of” is typically used in the move “Occupying the niche” in the CARS – Create a Research Space - model).

As can be seen, the description of the linguistic features of various academic text types across different disciplines has contributed to the development of EAP tests, syllabii and teaching materials. A cautionary note, however, is sounded in Swales (2002). In this paper, Swales reflects upon his experiences of working with academic corpora both as a teacher and a materials writer. His preferred approach, as an EAP practitioner, is top-down, looking at discourse from the level of genre, before moving down to the lexicogrammatical level, and he argues that a corpus approach proceeds in the opposite direction, starting at the concordance line view and moving upwards. His experience, he claims, makes the reversal of procedure difficult to learn, and the outcomes, in terms of the materials produced, are both costly in investment of time, and also of dubious value. These criticisms of corpora as a resource in materials development, and teaching, are challenging and pertinent.

We can counter, however, that the causes of Swales’s disquietude may lie in the fact that EAP corpus studies and methodologies are still in a relatively young state. Recent research has argued for a joining of the discoursal and the lexicogrammatical, as shown above, and for the exploration of contextual features in conjunction with textual analysis, in corpus approaches to EAP. One implication of these developments is that there is a need for corpora that can easily be examined in terms of rhetorical
moves, at least in broad rhetorical sections, such as ‘Introduction’, ‘Methods’, ‘Results’ and so on. Language features need to be related to rhetorical choices, and the division of texts into rhetorical sections is a first step in this direction. Secondly, it should be remarked that Swales’s concerns about the effectiveness of drawing on corpus evidence and findings derive, to a certain extent, from his experience of working with spoken data from the MICASE corpus. The use of spoken language data in corpus-based language learning is relatively new, and the criteria by which to assess the value of the evidence are not determined. Swales comments on the triviality of some of the findings from the corpus, such as the frequency data that show that there are several hundred tokens of ‘I don’t know’ in the MICASE corpus. Such findings do indeed appear to have little pedagogical value, as they are unrelated to pragmatic factors, and there is no distinction made in such an observation as to who tends to use ‘I don’t know’, to whom, in what context, for what purpose, and so on, nor why we were interested in the use of ‘I don’t know’ in the first place. The work of Swales and his colleagues on the MICASE corpus is ground-breaking, and it may well be that their work necessarily involves the exploration of approaches that turn out to be dead-ends, before informed and tested approaches to the exploration of academic speech corpora are developed, approaches that guide researchers and practitioners to an awareness of what questions should be posed, what parts of the corpus to investigate, and by what methods. The work of Mauranen (2004a and 2004b, for example), in particular, has done much to establish theoretical frameworks for the analysis of spoken academic language data.

4. Uses of corpora in the classroom

The past ten years have seen a steady increase in conference reports of applications of corpora in language teaching. The Teaching and Language Corpora conference has become a regular biennial event for a chiefly European audience since its inception in Lancaster in 1994, and similar events have been held in Poland (the PALC conferences of 1997, 1999, 2001, 2003 and 2005), and in North America (the annual North American Symposia on Linguistics and Language Teaching, from 1999). A number of publications deriving from these conferences have since appeared (Botley et al. 1996; Lewandowska-Tomaszczyk and Melia 1997; Wichmann et al. 1997; Burnard and McEnery 2000; Simpson and Swales 2001) Other publications which feature papers related to the use of corpora in EAP pedagogy are journals such as ReCALL, Language Learning and Technology and System.

One of the pioneers for the use of corpora and concordances in the language learning classroom is Tim Johns, at the University of Birmingham, who coined the term “Data-Driven Learning” (abbreviated as DDL) to describe an approach to learning that was focused upon the analysis of corpus linguistic evidence. The preferred form of learning in this case is necessarily inductive: learners inspect the evidence and look for patterns in the data from which they can form generalizations. An interesting archive of investigations made by Johns with learners can be found at the ‘Kibbitzers’ website (www.eisu.bham.ac.uk/Webmaterials/kibbitzers/). The word ‘kibbitzer’ was originally used in middle Europe to describe an observer of a chess game, someone who watched and commented but did not play, and is extended by analogy to the role of a corpus analyst, one who observes the features and patterns of communication. In one-on-one EAP teaching, Johns would investigate language problems from the student’s own writing by searching for particular lexical items or phrases in a suitable corpus. An example of this is the kibbitzer on ‘Cause’ (kibitzer 24), which looks at a
sentence produced by a student: “Although economic improvement may be caused by tourism, the investment and operational costs of tourism must also be considered.” The word ‘cause’ was identified as sounding out of place and thus became the subject of a corpus query. A sampling of the lines that Johns shows appear below, with 5 lines each of the verb in active and passive constructions:

In all instances shown here, what is caused is a negative outcome. The student had used a positive outcome with ‘cause’ and this was the root of the problem. A further search was then conducted to identify an alternative verb to use, and one of the searches was for ‘lead to’ which produces results such as the following:

As can be seen from this small sample of lines, ‘lead to’ can appear with both negative and positive outcomes, and thus has what can be described as neutral semantic prosody (Sinclair 1991). Johns also observes that ‘lead to’ is not used with a human subject, and does not appear in the passive.

The kibbitzer is an example of how knowledge about language can be acquired through analysis of corpus data. In this case, the student works with a teacher to solve a problem. A stronger version of DDL would involve learners working directly with corpora to investigate language problems by themselves. This approach has been criticized by Owen (1996), on the grounds that learners may be overwhelmed by the quantity of information (it is possible to get thousands of concordance lines from a single query), and learners may lack sufficient linguistic and background knowledge to make inferences. Responses to these criticisms are that most learners will require some training in the analysis of concordance lines, that the amount of data can be restricted by the materials writer or the teacher, and that it is possible to explore the
contexts of concordance lines through reading of the source texts, or even through investigation of the wider social processes through which the texts were produced. The relation of language features in concordance lines to the functions that they play within whole texts can help to recontextualize those features. Weber (2001) provides an account of how he has tried to combine genre and concordance based approaches to the teaching of academic essay writing, for example.

The Advanced School in Modern Languages for Interpreters and Translators in Forlì (SSLiMIT), part of the University of Bologna, has pioneered the development of data-driven learning and useful publications are the various papers of Guy Aston (for example, 1995, 1997, and Gavioli and Aston 2001), and Silvia Bernadini (2000, 2004). Gavioli (2005) is the most comprehensive account and evaluation of the use of DDL in the ESP classroom, illustrating the approach through a series of corpus-based explorations that were conducted by students in a process of “search-and-discovery”.

Cook (1998) challenges the relevance of corpus-driven (rather than corpus-based, or corpus-informed) approaches to pedagogy, on the basis that corpus linguistics should not dictate to pedagogy what the content of learning programmes should be. Similarly, Aston (1995) proposes that teachers should not try to apply the methods and findings of corpus linguistics directly to pedagogy, but should decide what uses CL findings have for the classroom. This view is especially suited to the EAP classroom, where corpus analysis can be seen to be but one of the several means available to the teacher and learner in approaching the study of the target discourse types and language skills. Aston also presents a theoretical justification for the use of concordancing and corpus investigation work in language teaching: concordancing can assist in the development of schematic and idiomatic knowledge, referring to Sinclair’s (1991) notions of the open principle and the idiom principle, through the apprehension of patterning in text and the analysis of contextual variation.

Concordancing activities also have potential for use in self-access work (Aston 1997), and he advises that in worksheets for this purpose, as for other purposes, it is preferable to constrain the range of answers possible and the number of concordance lines to be analyzed, so that definite answers can be provided. McEnery and Wilson (1997) argue that corpus-based language activities in the language learning classroom best fit the principles of discovery learning, divergent learning, mediated learning, and directed learning. Which principles are invoked will depend on how the activities are presented. Conrad (2004) has spoken of how corpus study requires the analyst, whether it be researcher, teacher or learner, to feel comfortable with using the qualifying statement, “It depends”, and she indicates that this is a difficult attitude for many educators to adopt – they prefer a simple clear rule and believe that learners require clear rules too. The analysis of corpus data, therefore, may require a change in attitude towards language data on the part of the novice analyst (who might be a teacher or learner).

Exploring Academic English (Thurstun and Candlin 1997) is the first EAP textbook to integrate the study of concordance lines into its methodology. The theoretical and methodological issues underlying the development of the book are well set out in Thurstun and Candlin (1998). The book presents sets of concordances for selected lexical exponents of a set of “rhetorical functions”, such as “stating the topic”, “referring to the literature” and “reporting the research of others”, with exercises.
leading the student methodically through the analysis of the lines. As the concordance lines are presented in the book, there is no need for the learner or the teacher to have access to computers and corpora directly. One criticism of the book is that the reward for the learner, in terms of words learnt related to time invested, is low: each unit focuses on only three or four lexical items (Thompson 2001).

Warren (1998) describes a set of concordance-based remedial grammar activities, that she and Tim Johns used with international students at the University of Birmingham. Seidlhofer (2001) describes how students can build a corpus of their own writing (summaries of articles) and compare these with the language of the originals, using corpus analysis tools, to look at collocational, grammatical and phraseological features. She found that student motivation was greatly increased when they worked on examples of their own writing (work that they had a personal investment in) as well as with texts from outside the classroom.

Tribble (2002) demonstrates how a short text from a corpus of academic writing can be investigated using concordance analysis, and contrasted with similar analyses of a reference corpus. Similarly, Granger and Tribble (1996) use a contrastive approach in comparing learner usage of connectors with that of native speaker writers, to help learners see how their writing differs from that of expert writers, and to illustrate means of achieving greater variety in their use of connectors. Further examples of ways that concordancing can be used in the language learning classroom can be found in Stevens (1995), Tribble and Jones (1997) and Partington (1998).

Cobb and Horst (2001) explored the effectiveness of vocabulary learning on a concordance-assisted reading course, using experimental methods to measure learning. The results were positive, suggesting that students who do concordance-based extension activities make greater vocabulary gains than those who do not, but Cobb and Horst recommend that further empirical testing of the effects of concordance-based learning needs to be conducted. Sripicharn (2002) argues that such research should be qualitative as well as quantitative, and his study makes use of questionnaire data in exploring how both native speakers and Thai non-native speakers carried out a range of concordance analysis tasks and the reasons that they gave for their choices. His results indicate that the Thai learners made greater use of the data than did the native speakers, who tended to rely on their intuition rather than ‘trust’ the data, but the Thai learners lacked knowledge of the pragmatic and discoursal aspects of the language, and were restricted in their ability to interpret the context of each concordance line. These findings suggest that the materials writer and the teacher need to be sensitive to the constraints under which learners approach concordance data. In another qualitative study, Turnbull and Burston (1998) examined the uses that two learners of English made of concordance-based activities, in correction of their written texts, and they found that learner cognitive style and previous learning experiences strongly influence the learner’s readiness to adopt the inductive learning strategies required in concordance work.

5. Survey of BALEAP institutions
As may be seen from the previous section, while it is relatively easy to assess the uses of electronic corpora in EAP research, because this work is widely reported, it is more difficult to evaluate the extent to which corpora have been exploited in EAP teaching. In order to find out more about whether corpora are used in EAP teaching and
materials development in the UK, I sent out a questionnaire to the 75 BALEAP (British Association of Lecturers of English for Academic Purposes) member institutions in May 2002. The 75 member institutions (each at a different UK university) range in size from EAP support units with only two full-time members of staff to departments with well over 15 full-time staff, a large part-time teaching cadre, and a responsibility both for teaching EAP on pre-sessional courses, and for running Masters courses in TEFL. In some cases, then, the EAP unit may be working on severely limited resources, while in others, the unit may have access to good resources, and considerable research expertise.

The aims of the survey were to find out:

- how many EAP units in the UK have access to corpora
- whether they are developing their own corpora
- how many make use of corpora in some way, and
- how many staff know how to make use of corpora and use concordancing software.

The results suggest that the uptake of corpus-based approaches in EAP teaching is limited. Of the 30 institutions that returned responses to the survey, 13 said that they had no access to corpora at all. 16 said they do not use corpora in any way (by this, they meant that no members of staff in the EAP unit made use of corpora directly). A further 5 said that only one member of staff used corpora. Of the remaining nine, the only unit where all teachers make use of corpora is IALS at Edinburgh, where the two full-time staff members have access to a number of corpora and have also developed their own corpus of native speaker performances of the tasks in the Study Speaking textbook written by themselves (Lynch and Anderson 1992). At IALS, the use of corpora also appears to be well-integrated into practice: for the teachers, as a source of data for materials development, and evidence in feedback to students; for the students as a resource to draw on in their project work. This degree of integration is, however, clearly the exception, rather than the norm.

Seven EAP units said that they had started work on developing their own corpora, but two of these indicated that they had given up the venture. Of the remaining five, three are creating corpora of relevant texts in the special subject areas that their students are studying, and these corpora are being used in materials creation.

A lack of familiarity with technical resources available, and possibly also with the potential value of corpus evidence, was evident in the fact that many respondents stated that they did not make use of any online corpus resources, such as COBUILD, MICASE or the Web Concordancer. Of the 13 units that mentioned software that they used, the commonest responses were that they used Wordsmith Tools (7), or the COBUILD concordancer (3), by which it is assumed that they meant that they used the online COBUILD concordancer. One respondent mentioned the Longman Mini-concordancer (a rather dated programme) and there was only one mention of MicroConcord (Gavioli and Aston 2001 recommend this as the best program to use with learners, for its simplicity). WordPilot (Milton 2000), a powerful concordancer designed for language learners, and which fits into Microsoft Word as a plug-in, was mentioned by only one respondent.

Overall, then, use of corpora and corpus analysis methodologies at the time of the survey was clearly limited, and in many cases non-existent. The most common
reasons given by the respondents were that these institutions lack resources, and they lacked familiarity with both the resources and with potential applications. In several cases, they also claimed that their units were already overworked and did not have time to learn about corpora. In many cases, it was clear that use of corpora was limited to one or two members of staff.

In 2006, the situation has improved to a degree, although I cannot substantiate this with reference to a comparable survey. Within BALEAP institutions, I am aware of two large scale corpus development projects, one at Imperial College where the EAP unit has created huge corpora of web pages from each departmental website at the university, and these corpora have been used for developing specialized teaching materials to use in subject-specific in-sessional language support classes (Hartill and Lefevre 2003). The second is at the University of Nottingham, where a team is creating a set of corpus resources and a concordancer interface to facilitate direct use of a corpus by their EAP students. Using the terminology of Huberman (1973), these few users can be seen to be the enthusiasts and possibly also the early adopters of an innovation, and it will be some time (Huberman posits an S-curve rate of adoption) before the late adopters are prepared for change.

6. Discussion

On the basis of the review, it can be seen that corpus-based studies, through research reports and through reference materials, have already provided EAP practitioners with a number of insights into the language of academic discourse, and, to a lesser degree, into the diversity of textual practices between genres and disciplines. The survey, however, shows that the direct application of corpus methodologies and corpus evidence to the classroom remains as yet limited and this is due in large part to a lack of familiarity amongst practitioners with 1) the methods of corpus analysis; and 2) with the criteria by which insights afforded by corpus-based linguistic research should be evaluated. In addition, many practitioners lack easy access to adequate corpus resources.

In order to develop the direct use of corpora in EAP teaching and course design, it is proposed that the following are required:

- empirical research into the usefulness (or not) of corpora in direct teaching
- improved channels for dissemination of practice
- greater access to relevant corpora (MICASE/BASE; Learner corpora; corpora of relevant exemplar texts) and corpus analysis tools, and an increased number of corpora created for use by practitioners rather than primarily for researchers
- thorough training in corpus analysis techniques, and the evaluation of such methods, into EAP teacher development courses, especially within MA TESOL programs

The first point refers back to the caution that Cobb and Horst advised; there is as yet little empirical evidence of the value of concordance activities in language teaching. As Salkie (2002) comments, “this is ‘evidence-based teaching’ – but only in the sense that it uses authentic evidence in the classroom, not because it is based on evidence about the effectiveness of the teaching method”. Until more research has been conducted, it is not possible to claim with justification that corpora can make an effective direct contribution to the classroom. At the same time, however, it must be recognized that the application of corpus-based methodologies to EAP teaching requires a greater level of understanding of the concepts and technologies of corpus
work. Without a greater level of understanding of the purposes and potential for use of concordancing and corpus-based work in the classroom, the implementation of such approaches cannot be fairly assessed.

Secondly, there appears, from the lack of familiarity, among the respondents to the BALEAP survey with possible uses of corpora in EAP teaching practice, that there is a need for better means of dissemination of principles and practice. Furthermore, the fact that many of these institutions claimed to have no access to corpora suggests that access to relevant sources of data and to affordable data analysis tools may be a necessary step towards greater uptake of corpus-based approaches. As seen above, only limited access is available for many of the large corpora (the BNC and Bank of English online samplers, for example) unless the unit is prepared, and able, to pay for the purchase or subscription to the full corpus. In several cases, technical support is also required. The MICASE corpus is a welcome development in the area of large scale EAP corpora as it offers open access to all through its web interface, and it is to be hoped that is a sign of what is to come. By the time of publication of this paper, the BASE corpus will also be accessible online, using the Sketch Engine interface (for further information, see http://www.rdg.ac.uk/AcaDepts/lb/base_corpus/). This is soon to be supplemented by the BAWE corpus (British Academic Written English corpus, a substantial collection of assessed writing by students at three UK universities, in a range of different disciplines, which is currently under development, see www2.warwick.ac.uk/fac/soc/celte/bawe/).

Chambers and Kelly (2002) have suggested that concordancing could become a basic IT skill for language learners in the coming decade. They envisage a learning environment in which students can access corpora, both public and personal ones, from their own computers, to assist in their empirical investigations of language. For this to happen, corpus analysis tools need to be made more easily available, perhaps even integrated into, and packaged with, word processing software, in a manner similar to that of WordPilot (Milton 2000), a programme that appears as a plug-in to Word, so that users have simple and ready access to software, and to preset corpus links, when they are working in a larger programme that they use regularly. Johns (personal communication) has worked on a concordance program that also helps the user to produce language study exercises; although the program is not yet publicly available, it is to be hoped, as Gavioli & Aston (2001) have proposed, that concordancing programs which assist the EAP teacher and learner in exploiting corpora will be developed, and a preferred feature of such programs would be that they are cheaply priced, or even freely available, to encourage uptake of the technology. Examples of concordancing programmes that are free and relatively powerful are ConcApp (www.edict.com.hk/PUB/concapp/) and AntConc (www.antlab.sci.waseda.ac.jp/software.html).

As is evident from the survey reported above, there is a widespread lack of knowledge among EAP teachers about the principles of corpus creation, the criteria for evaluating the data that is obtained from corpus searches, and the potential uses of concordancing both in the language classroom and in materials development. This can be remedied to some extent by greater dissemination of practice and of research in journals and conference presentations, not only at conferences for enthusiasts but also at meetings of EAP practitioners in general, but provision of better coverage of such issues on training courses is another solution (the availability of guides such as
Bowker & Pearson 2002 will be of help). In-house staff development training sessions, and adequate coverage of corpus-based approaches on MA in TESOL programs would swiftly increase the numbers of EAP teachers that are familiar both with the techniques, and with the issues involved in exploitation of corpora in the teaching of English for Academic Purposes.

References


tlt.msu.edu/vol5num3/aston/default.html


5th International Conference in LAP/LSP, University of Macedonia, October 2005


Thompson, P. (forthcoming) “A corpus perspective on the lexis of lectures, with a focus on economics lectures” In K. Hyland and M. Bondi (eds) *Academic discourse across disciplines* Hamburg: Peter Lang


