PAPER 18
Corpora and English for Academic Purposes

Professor Hilary Nesi (PhD)

Coventry University, UK

h.nesi@coventry.ac.uk

Abstract
This paper provides an overview of four types of corpus currently available to EAP practitioners, and their strengths and weaknesses:

1. Corpora of "expert" writing – by far the most common, because expert writing is readily available in the public domain
2. Learner corpora - compiled to monitor the process of language acquisition, and their accompanying "control" corpora, used for comparative purposes to identify learners' overuse and underuse of lexical and grammatical items
3. Corpora of university student writing – much less common, because of the difficulties of obtaining balanced quantities of good quality text, but a useful source of information for EAP practitioners
4. Spoken academic corpora – multimodal resources which are expensive and time consuming to produce, but which are also of particular interest to the EAP profession.

The paper focuses particularly on the British Academic Written English (BAWE) corpus, and the British Academic Spoken English (BASE) corpus, but will also outline plans for the development of parallel corpora at UTM, possibly leading to the creation of a Malaysian Academic Spoken English (MASE) corpus.

1. Introduction

Corpora (collections of naturally occurring samples of language stored in electronic form) are playing an increasingly important role in our professional life as LSP practitioners. Multi-million word databanks have been created to represent a wide range of genres, including fiction, journalism, and academic publications, and these inform the design of all the recent major English dictionaries and descriptive grammars. The *Oxford Dictionary of English* (2003, 2005) and the *Oxford Advanced Learner's Dictionary* (1995, 2000, 2005), for example, draw on the 100 million word British National Corpus, and entries in the *Longman Grammar of Spoken and Written English* (1999) and the *Longman Student Grammar of Spoken and Written English* (2002) were derived from analysis of the Longman Spoken and Written English Corpus (over 40 million words).

Lexicographers and grammarians who seek to describe an entire language system, or significant parts of it, need to work with such large and representative collections of texts. Not all corpus research is of this nature, however. Smaller, more specialized corpora may be used to investigate the frequency and co-occurrence of lexical and grammatical features in specific genres, fields or institutions. As teachers of

languages for specific purposes, it is these small specialized corpora that interest us most. This paper will discuss various types of specialist academic corpora in terms of their relevance to the EAP practitioner. Some of these corpora contain spoken academic language, but most are primarily concerned with written academic language, produced by expert or semi-expert writers, language learners, and students.

2. Corpora of published academic writing

At the least refined level, it is possible to use the web as a kind of corpus, simply by searching for language items that seem relevant in an EAP context. We can consult a search engine to check the frequency of a grammatical structure or collocation, for example, although because of the varied nature of texts on the web the results are not always helpful as a guide to good writing. For example, a Google search for *different to*, *different from* and *different than* finds that *different to* is the most frequent, whereas Israel (1997) cites statistical evidence from the Collins Cobuild Bank of English which shows that general usage has a rather different distribution, varying according to mode and region. The distribution in texts on the web with an academic provenance, as reflected in ".edu" and ".ac.uk" domain names, differs again, supporting the COBUILD frequency figures by favouring *different from,* but suggesting that websites conform more closely to spoken than written norms (see Table 1).

Table 1: Percentage distributions of prepositions with *different*

Different ….	***from***	***to***	***than***
Bank of English U.K. writing	87.6	10.8	1.5
Bank of English U.K. speech	68.8	27.3	3.9
Bank of English U.S. writing	92.7	0.3	7.0
Bank of English U.S. speech	69.3	0.6	30.1
Google search	33.3	40.5	26.1
Google Advanced Search (.ac.uk)	68.2	26.3	5.5
Google Advanced Search (.edu)	55.7	10.2	34.1

An online web concordancer can offer more refined search routes than a standard search engine, although the varied quality of web material will still remain a problem. *WebCorp* at Birmingham City University (Kehoe & Gee, 2007; Renouf, Kehoe & Banerjee, 2007) can restrict searches by site domain (the whole or part of a URL), newspaper domains (e.g. UK broadsheets) and by broad topic (e.g. arts, business, sport etc.). The program also allows collocate searches, and the use of wildcards and square brackets (for example **the [ship|boat] s[a|u]nk** to match *the ship sank*, *the ship sunk*, *the boat sank* or *the boat sunk*). Another useful feature for the teacher is *WebCor's* ability to generate frequency or alphabetical wordlists for any given webpage. To make a single list of words from a collection of texts, however, it is necessary to run them through a concordancing program such as Scott's *WordSmith Tools* (latest version 2008) or the freely available *Antconc* (Anthony, latest version 2008).

Strictly speaking a corpus is more than just a selection of texts, of course. It might be perfectly sufficient for personal study to identify and analyze some interesting samples from the web, but researchers need copyright permission to transform

downloads into a more permanent resource, for use by other scholars. Moreover the process of corpus creation is normally a principled one, conforming to a design matrix so that due balance is created between various factors that might later affect the findings of analysis. Thought should be given to the disciplines and topics of the corpus holdings, and perhaps most particularly to their provenance as an indication of quality and communicative purpose. Early corpora were sometimes a bit haphazard in this respect, but as the science of corpus construction has developed, researchers are taking increasing care to decide from the outset what kinds of texts (and authors) they do and do not want to include, and in what proportion.

Specialized academic corpora might concentrate on just one genre (for example the research article) or aim to represent a wide variety of genres. Likewise, they might focus on a single discipline, or many. Most, however, tend to be made up of professionally edited and expertly written texts, because these kinds of text, although they may not all be available on the web, are in the public domain, and are therefore relatively easy to access.

The TOEFL 2000 Spoken and Written Academic Language (T2K SWAL) Corpus is a good example of a carefully constructed academic corpus for use in EAP scholarship (e.g Biber 2006, Biber & Barbieri, 2007). It claims to represent "the full range of spoken and written registers used at US universities" (Biber et al. 2002, p11) and alongside university webpages it contains textbooks, course packs, and similar expert sources. These are the kinds of texts university students are required to read, but they are not the kinds of texts students are required to write. Student writing is absent from the corpus, presumably because it was less accessible to the corpus compilers.

Small personal corpora of academic textbooks have also been created (cf. James, Ho & Chu, 1997 and Hyland, 1999a, 2000). Even more common are corpora of published research articles, as these are extremely easy to obtain online, for example via university library services. Collections of research articles were first used in the pioneering work of Swales (1981, without the benefit of computer analysis), and later by Gosden (1993), Hyland (1999b, 2000), Marco (2000), Williams (2006), El Malik & Nesi (forthcoming) and many others.

The writing of experts is of course an important area for research, as it constitutes a model to which all academic writers ultimately aspire, and which they will repeatedly encounter in their programmes of reading. EAP writing tutors might not find the analysis of textbooks and research articles particularly helpful in lesson planning, however, because novice writers do not begin by writing for publication, and their early attempts at academic writing are likely to be assessed texts produced in the context of a course of study. Although there are undoubtedly generic similarities between the student assignment and the published academic text, there are also great differences in their communicative purposes and rhetorical features.

Some corpus-based studies of academic writing have focused on the semi-expert writing produced by students in their final stages of study, at the end of a postgraduate programme. Pramoolsook (2005), for example, discusses dissertations at Masters level, while Charles (2006), Thompson (2000, 2005) and Thompson and Tribble (2001) have analysed aspects of PhD theses. This kind of writing is usually available for readership beyond the confines of the department in which it was prepared, and is

therefore more easily acquired for corpus analysis. The texts which novice writers are required to produce, on the other hand, are often inaccessible to all but a few interested parties in the students' own subject disciplines.

3. Learner corpora

Most corpora of student writing are "learner" corpora, consisting of texts produced by learners of a second or foreign language, and used to monitor the process of language acquisition. As with other types of corpora, decisions need to be made at an early stage in the compilation process regarding the types of text to include. Tono (2003: 800) divides these design considerations into three categories: language-related, task-related and learner-related (Table 2).

Table 2: Design considerations for learner corpora (adapted from Tono 2003)

Language-related	Task-related	Learner-related
Mode (written / spoken)	Method of collection (e.g. cross-sectional / longitudinal)	Internal – cognitive (age / cognitive style)
Genre (e.g. fiction / essay)	Method of elicitation (e.g. spontaneous / prepared)	Internal-affective (motivation / attitude)
Style (e.g. narration / argumentation)	Use of references (e.g. access to dictionaries, source texts)	L1 background L2 proficiency
Topic	Time limitation (e.g. fixed / free / homework)	L2 environment ESL/EFL / level of school)

Learner corpora usually contain texts that have been written in the context of English language courses, either in class, under examination conditions, or for homework. These tend to take the form of argumentative essays on personal or general topics which do not require any preparation on the part of the writer, although in some cases information about the topic, such as a graph, table or short text is provided for the writer as part of the task specification. This is because learner corpus research is more concerned with lexical and grammatical variation amongst contributors than in variation in adherence to generic conventions. Thus, although learner corpora provide some insight into the type of tasks language teachers set, they do not represent the type of writing undertaken outside the language classroom. In contrast to language learning tasks, writing for academic or professional purposes usually requires advance preparation, extensive referencing to extratextual sources or data, and accommodation to the norms of a particular discourse community.

The best known and probably the first learner corpus is the three million word *International Corpus of Learner English* (ICLE), developed at the Louvain Centre for English Corpus Linguistics in Belgium, largely in collaboration with other European universities but also with contributions from universities in Brazil, Hong Kong, Japan and South Africa (16 different mother-tongue backgrounds in all). ICLE consists of essays produced by learners in their third or fourth year of study in a non-English-medium environment. Typical essay titles are "Crime does not pay" and "The role of censorship in Western society". The *Longman Learners' Corpus* (LLC) and the *Cambridge Learner Corpus* (CLC) are larger resources, covering a wider range of

levels and language backgrounds. Longman offers dictionaries to teachers in return for contributions to LLC, whereas Cambridge University Press has compiled CLC from the thousands of exam scripts written by students taking Cambridge ESOL English exams around the world. The ICLE corpus and handbook are commercially available on CD-Rom, but Longman and Cambridge University Press restrict access to LLC and CLC to their own lexicographers, materials writers and, in the case of CLC, the staff at Cambridge ESOL.

The *Japanese EFL Learner Corpus* (JEFLL) under development at Tokyo University of Foreign Studies is not dissimilar to ICLE and LLC, but focuses on younger learners. It contains the essays of more than 10,000 Japanese school children. Teachers contributing to this project are advised to set in-class controlled writing tasks on topics such "my school festival" or "bad dreams". Again the purpose of the project seems to be to provide examples of lexical and grammatical errors commonly produced by learners, and on their website the JEFLL team admit that the corpus "may not be suitable for examining stylistic differences in L2 writing".

Two further learner corpora, the *Thai English Learner Corpus (TELC)* and the *Lancaster Corpus of Academic Written English* (LANCAWE) are made up of the writings of university students, but are typical learner corpora in that the writing tasks have been set by teachers of EAP and EFL, rather than by subject tutors. These corpora are made up of English language examination scripts and homework assignments, both primarily intended to practise and demonstrate language proficiency rather than subject knowledge and academic literacy.

Perhaps the most academic of the learner corpora is the *Hong Kong University of Science and Technology* (HKUST) *Learner Corpus*. This was developed in an English-medium university, and thus contains more examples of texts written primarily to inform, rather than to practise English language skills. In this respect HKUST functions partly as an English as a Lingua Franca (ELF) corpus (see Section 6) as well as a learner corpus, although it was established well before the ELF movement got under way. HKUST has provided data for Hyland and Milton (1997), Milton (2000), Flowerdew (1998), Green et al. (2000) amongst others. Unfortunately, unlike JEFFL and LANCAWE, the HKUST corpus does not have a website or offer downloadable files.

Learner corpora are often used as data for 'Contrastive Interlanguage Analysis' (Granger 1998), a technique which examines differences between native and non-native varieties of the same language, for example in terms of the overuse and underuse of lexical and grammatical items. For this purpose learner corpora are compared with 'control corpora' of essays on similar topics produced by native speaker students. The control corpus for ICLE is the 324,304 word *Louvain Corpus of Native English Essays* (LOCNESS), which contains examination scripts and essays by British and American university students and British A level students, on general and literary topics. Although these were all produced as part of assessed work the grades they received are not recorded, there is little information about the context in which they were written, and the collection does not provide a representative sample of university student writing.

In studies involving the HKUST corpus, a control corpus of A-level General Studies

scripts has been used. This is sometimes known as the *Cambridge Syndicate Examination* corpus. Though useful as a control against which to compare learners' use of English, the tasks did not require subject-specific knowledge, nor are they typical of those set in university departments.

4. Corpora of university student writing

Corpora of university student writing are distinct from most learner corpora and control corpora in that they contain texts produced for subject tutors, as opposed to writing tutors, and are intended to demonstrate skills and knowledge relevant to their discipline, rather than language proficiency. Such writing is very different from that produced under exam conditions or in the classroom because the writers are relatively free from time constraints, and in most cases are expected to consult and cite data sources.

Small personal collections of assessed student writing include those used by Woodward-Kron (2002a,b), who examined 58 assignments produced by trainee teachers in Australia, and Hyland (2002), who worked with a collection of 64 project reports written by final year Hong Kong undergraduates. There also exist larger collections of assignments, compiled for use by other students, rather than by researchers and EAP practitioners. These are the essay banks accessible through student associations at some universities (for example at York and Kent in the UK) and also via a number of less scrupulous commercial websites. Essay banks are informal, inadequately documented and unannotated. They are patchy in their coverage of discipline areas, are not monitored by academics, and do not necessarily represent suitable models of writing. They also encourage students to copy, rather than to critically evaluate. (Guided analysis of well-written excerpts from relevant genres, on the other hand, may actually help learners to avoid the temptation to cut and paste from an on-line source.)

Fully documented corpora of good quality assessed university student writing have only come on the scene very recently. In the USA, the Michigan Corpus of Upper-level Student Papers (MICUSP) is under development at the University of Michigan (and currently contains about 900,000 words), and the provisionally-named Viking Corpus of Student Academic Writing was recently launched at Portland State University (and currently contains about 700,000 words). In the UK, the British Academic Written English (BAWE) corpus[1] has just been completed, and quantitative and qualitative information about its contents are now being disseminated.

Most of the contributors to the BAWE corpus were native speakers of English, but this was not a criterion for contribution; assignments were accepted regardless of the first language of the writer provided that they had received a grade equivalent to an upper second or first class honours degree. Contextual details of every contributor were noted, however, including gender, first language, number of years of secondary education in the UK, department, assignment title, grade, and level of study.

The corpus was designed to fit a four by four matrix, with a roughly equal distribution

[1] developed with funding from the ESRC (RES-000-23-0800)

across levels (three or four years of undergraduate study, and taught Masters level) and across disciplinary groupings (Arts and Humanities, Life Sciences, Physical Sciences and Social Sciences). The design is summarized in Table 3, with details of the number of words collected for each matrix cell.

Table 3: BAWE corpus holdings

disciplinary group	Yr 1	Yr 2	Final year	Masters	Total
Arts & Humanities	468,517	583,914	426,832	235,428	**1,714,691**
Life Sciences	300,190	408,552	223,784	482,229	**1,414,755**
Physical Sciences	301,161	313,622	426,054	343,733	**1,384,570**
Social Sciences	371,753	475,959	429,427	723,621	**2,000,760**
Total words	**1,441,621**	**1,782,047**	**1,506,097**	**1,785,011**	**6,514,776**

Of particular interest to the EAP practitioner is the range of genre types that were identified in the corpus (see Table 4). The 'argumentative essay' is the focus of most academic writing programmes and EAP textbooks, and is the unit of collection for most learner corpora, but although the essay is the best represented generic type in the BAWE corpus there are twelve other types of text that occur across many disciplines, and which therefore also deserve consideration in EAP course materials.

Table 4: Genre families

Genre family	Frequency	Range*	Examples
Essay	1225	24	Commentary, discussion, exposition
Methodology Recount	347	15	field report, forensic report, lab report
Critique	319	24	academic paper review, film review, financial report evaluation
Explanation	198	15	methodology review, disease overview, system overview
Case Study	194	12	organisation analysis, patient case notes, tourism report
Exercise	114	15	Calculations, data analysis, stats exercise
Design Specification	92	7	building design, product design, website design
Proposal	76	15	building proposal, marketing plan, research proposal
Narrative Recount	72	14	Biography, reflective recount, urban ethnography
Research Report	61	17	research paper, topic-based dissertation
Problem Question	40	7	law problem question, logistics simulation, medical problem
Literature Survey	35	11	annotated bibliography, anthology, summary
Empathy Writing	32	11	information leaflet, job application, newspaper article

*Across the 24 departments where 50 or more assignments have been collected

The descriptions of the social purpose and typical components of genres in the BAWE corpus should make it easier for tutors to identify factors that result in lack of communicative success. For example, two genre families that are common in the hard applied disciplines are design specifications and proposals. Design specifications are

intended to demonstrate students' ability to design a product that can be manufactured, or a procedure that can be implemented. Texts of this type typically include a design brief and a design plan, and often include accounts of the way the design was developed and tested. Proposals, on the other hand, are intended to demonstrate the ability to make a case for future action, and include persuasive argumentation regarding the merits and purpose of the student's plan. Analysis of the BAWE corpus can help EAP tutors understand the structure and functions of the two types, the better to advise learners faced with design specifications or proposal tasks, and to explain the differences between their communicative requirements.

5. Spoken academic corpora

Spoken academic corpora are few and far between, largely due to the difficulty and expense of identifying, recording and transcribing suitable spoken academic texts. American and British academic spoken English is represented to a certain extent in the *Longman Spoken and Written English Corpus* and the *TOEFL 2000 Spoken and Written Academic Language Corpus* (neither of which are publicly available). More extensive coverage is available in the freely accessible *Michigan Corpus of Academic Spoken English* (MICASE) and the *British Academic Spoken English* (BASE) corpus[2].

These two corpora are made up of speech recordings, text transcripts, and a database of speaker and speech event information. They are similar in size (MICASE contains 1,848,364 words, BASE 1,644,942 words) and have roughly equal quantities of text in broadly similar disciplinary domains, but they have been designed according to different matrices. MICASE consists of smaller quantities of a broader range of speech events, including meetings, interviews, study groups and so on as well as large and small lectures, whereas BASE contains larger numbers of just two speech event types (see Table 5): 160 lectures (almost entirely monologic) and 39 'seminars' (highly interactive small class events, featuring student presentations and discussion).

Table 5: The BASE corpus matrix

Disciplinary Grouping	Lectures	Seminars
Arts & Humanities	40	10
Life Sciences	40	10
Physical Sciences	40	9
Social Sciences	40	10

An unusual feature of BASE is that, unlike MICASE and most other spoken corpora, the majority of the recordings are on digital video rather than audio tape.

Unfortunately, most published EAP listening materials still tend to be based on recordings of scripted or semi-scripted 'lecturettes', performed by actors (see Nesi, 2001). These bear little resemblance to real lectures produced by academics in their

[2] Developed with funding from BALEAP, EURALEX, the British Academy and the Arts and Humanities Research Council

disciplines, and therefore are less than ideal as a means of preparing learners for English-medium study. Similarly many EAP speaking activities do not fully reflect the kinds of demands that will be made of students in their disciplines. Some materials writers have drawn on corpus data, however, in particular Reinhart (2002) who refers to MICASE student presentations, and the EASE series of EAP materials on CD-Rom, with activities based around video excerpts from the BASE corpus (Kelly, Revell & Nesi 2000; Kelly, Richards & Nesi 2004; Kelly, Sharpling & Nesi 2006).

6. Corpora of English as a Lingua Franca

As can be seen from above discussion, British and American varieties of English predominate in current academic corpora. Some of the contributors to BAWE, BASE and MICASE are users of English as a lingua franca, but this did not factor in the design of the corpus matrices. Non-native speaker speech and writing was recorded, but without the intention of analysing its particular characteristics, or comparing non-native-speaker and native-speaker production.

Around the world, of course, the majority of academic and professional users of English are not native speakers, and English as a Lingua Franca (ELF) corpora have been developed to pay particular attention to the use of English by those who do not speak it as their mother tongue, but would not necessarily class themselves as English language learners. The best known and earliest example of a corpus with ELF components is the *International Corpus of English* (ICE), launched in 1990 to facilitate comparative studies of varieties of English used around the world. ICE contains a number of one million word subcorpora collected in L1 contexts (Australia, Great Britain, New Zealand) and in ESL contexts (Hong Kong, East Africa, India, Philippines and Singapore).

A small proportion of each ICE sub-corpus is made up of academic texts (80,000 words of academic writing, and 40,000 words of class lessons), but there is a need for ELF corpora with a greater academic focus. English is increasingly becoming the medium of instruction at university level, in EFL as well as ESL contexts, as institutions aim to attract greater numbers of international students, and seek to promote their research on the international stage.

The Finnish *English as a Lingua Franca in Academic Settings* (ELFA) corpus was created as part of the ELF movement in response to this need (cf. Mauranen (2003, 2006). ELFA currently contains 0.9 million words of transcribed speech recorded at the University of Tampere and Tampere University of Technology, and it offers an interesting model for development of other spoken corpora in ESL/EFL settings. The basic unit of sampling is the speech event type, and genres have been selected in terms of their prototypicality across disciplines, (for example lectures, seminars, thesis defences and conference presentations), their influence in terms of the number of participants, and their prestige in the discourse community(the corpus includes, for example, guest lectures, and plenary lectures at conferences).

ELFA includes both monologic and dialogic speech, but places greater emphasis on dialogue, reflecting the compilers' interest in pragmatics. In other university contexts,

however there might be more concern with ELF monologue, especially if this could lead to the development of teaching materials to support novice university lecturers and students new to English-medium instruction. This is the case in Malaysia, where no ELF corpus yet exists, yet where there is a clear need for relevant staff development, EAP and Study Skills materials.

Researchers at Coventry University and Universiti Teknologi Malaysia (UTM) are now beginning a British Council funded project (a PMI2 Connect Research Co-operation Award) involving the creation of a small corpus of academic lectures, modelled on the BASE corpus. We aim to film ten engineering lectures in each university, as far as possible on matching or similar topics, and to gain insights into English medium engineering discourse which can be put to immediate use in various student and staff development programmes at Coventry and UTM.

We also hope that the Malaysian component of this small corpus can serve as a pilot for the development of a full scale Malaysian Academic Spoken English (MASE) corpus, in the not too distant future. It is hard to conceive of any new large-scale materials writing project that would not make use of corpus data, and it is also hard to justify the sole use of L1 corpus data in a country like Malaysia, with its own educational practices and a thriving university sector.

References

Anthony, L. (2007) *Antconc 3.2.1* http://www.antlab.sci.waseda.ac.jp/software.html

Biber, D. & Barbieri, F. (2007) Lexical bundles in university spoken and written registers. *English for Specific Purposes,* 26 (3) 263-286

Biber, D. (2006) Stance in spoken and written university registers. *Journal of English for Academic Purposes,* 5 (2) 97-116

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M.. (2002). Speaking and writing in the university: a multidimensional comparison. *TESOL Quarterly, 36*, 9-48.

British Academic Spoken English (BASE). Corpus developed at the Universities of Warwick and Reading, UK.

British Academic Written English (BAWE). Corpus developed at the Universities of Warwick, Reading and Oxford Brookes, UK.

Cambridge Learner Corpus (CLC)
http://www.cambridge.org/elt/corpus/learner_corpus.htm

Charles, M. 2006. The construction of stance in reporting clauses: A cross-disciplinary study of theses. *Applied Linguistics* **27** (3), pp 492–518

Dudley-Evans, T. (1986). Genre analysis: an investigation of the introduction and discussion sections of MSc dissertations. In Malcolm Coulthard (Ed.), *Talking about text* (pp. 128-45). Birmingham: English Language Research, University of

Birmingham Press.

El Malik, A. & H. Nesi (forthcoming) Publishing research in a second language: the case of Sudanese contributors to international medical journals. *Journal of English for Academic Purposes*

English as a Lingua Franca in Academic Settings (ELFA) corpus, Universities of Tampere and Helsinki, Finland. http://www.uta.fi/laitokset/kielet/engf/research/elfa/

Flowerdew, J. (2002) Ethnographically Inspired Approaches to Academic Discourse. In: J. Flowerdew (ed) *Academic Discourse* London: Longman

Flowerdew, L. (1998). Integrating 'Expert' and 'Interlanguage' computer corpora findings on causality: discoveries for teachers and students. *English for Specific Purposes* 17 329-345

Gosden, H. (1993). Discourse functions of subject in scientific research articles. *Applied Linguistics, 14*, 56-75.

Granger, S. & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes, 15*, 17-27.

Granger S. (2008) Learner Corpora in Foreign Language Education. In Van Deusen-Scholl N. and Hornberger N.H. (ed.) *Encyclopedia of Language and Education. Volume 4. Second and Foreign Language Education* . Springer, 337-351.

Gilquin G., Granger S. & Paquot M. (2007) Learner corpora: the missing link in EAP pedagogy. In Thompson, P. (ed.) *Corpus-based EAP Pedagogy* . Special issue of *Journal of English for Academic Purposes* 6(4): 319-335.

Green, C., Christopher, E. & Lam, J. (2000). The incidence and effects on coherence of marked themes in interlanguage texts: a corpus-based enquiry. *English for Specific Purposes, 19*, 99-113.

Hyland, K. (2002). Directives: argument and engagement in academic writing. *Applied Linguistics, 23*, 215-239.

Hyland, K. (2000). *Disciplinary Discourses: Social Interaction in Academic Writing.* London: Longman Pearson Education.

Hyland, K. (1999a). Talking to students: metadiscourse in introductory textbooks. *English for Specific Purposes, 18*, 3-26.

Hyland, K. (1999b). Academic attribution: citation and the construction of disciplinary knowledge. *Applied Linguistics, 20*, 341-367.

Hyland, K. & Milton, J. (1997) Qualifications and certainty in L1 and L2 students' writing. *Journal of Second Language Writing, 6*, 183-205.

International Corpus of English (ICE). The Chinese University of Hong Kong and

University College London. http://www.ucl.ac.uk/english-usage/ice/

International Corpus of Learner English (ICLE). Université Catholique de Louvain http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm

Israel, M. (1997) "different to", "different than" *AUE FAQ document*, alt.usage.english newsgroup http://alt-usage-english.org/excerpts/fxdiffer.html

James, G., Ho, P.W.L. & Chu, A.C.Y. (1997). *English in biochemistry, biology and chemistry: A corpus-based lexical analysis*. Hong Kong: Language Centre, Hong Kong University of Science and Technology.

Japanese EFL Learner Corpus (JEFLL), Tokyo University of Foreign Studies, Japan http://jefll.corpuscobo.net/

Kehoe, A. & M. Gee (2007) New corpora from the web: making web text more 'text-like'. in Pahta, P., I. Taavitsainen, T. Nevalainen & J. Tyrkkö (eds.) Towards Multimedia in Corpus Studies, electronic publication, University of Helsinki.

Kelly, T., R. Revell & H. Nesi (2000) *Listening to Lectures.* EASE (Essential Academic Skills in English) Series, University of Warwick, UK http://www.ease.ac.uk/

Kelly, T., L. Richards & H. Nesi (2004) *Seminar Skills 1: Presentations.* EASE (Essential Academic Skills in English) Series, University of Warwick, UK http://www.ease.ac.uk/

Kelly, T., G. Sharpling & H. Nesi (2006) *Seminar Skills 2: Discussions.* EASE (Essential Academic Skills in English) Series, University of Warwick, UK http://www.ease.ac.uk/

Lancaster Corpus of Academic Written English (LANCAWE) Lancaster University http://www.ling.lancs.ac.uk/groups/slarg/lancawe/

Longman Learners' Corpus (LLC) http://www.pearsonlongman.com/dictionaries/corpus/learners.html

Louvain Corpus of Native English Essays (LOCNESS) http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/locness1.htm

Marco, M. J. 2000. Collocational frameworks in medical research papers: a genre-based study. *English for Specific Purposes*, 19, 63-86.

Mauranen, A. (2006) A rich domain of ELF – the ELFA corpus of academic discourse. *Nordic Journal of English Studies* 5.2: 145–59.
Mauranen, A. 2003. The Corpus of English as Lingua Franca in Academic Settings. *TESOL Quarterly* 37 (3), 513-527.

Michigan Corpus of Academic Spoken English (MICASE). University of Michigan, USA http://quod.lib.umich.edu/m/micase/

Michigan Corpus of Upper-level Student Papers (MICUSP). University of Michigan, USA http://www.micusp.org/home

Milton, John. (2000) *Elements of a written interlanguage: A computational and corpus-based study of institutional influences on the acquisition of English by Hong Kong Chinese students*. Hong Kong: HKUST.

Nesi, H. (2001) 'A corpus based analysis of academic lectures across disciplines', in: Cotterill, J. and Ife, A. (eds) *Language Across Boundaries***,** London: Continuum Press.

Pramoolsook, I. (2005) Field and effects of genre transfer: an investigation of changes in field from dissertation to research article in two disciplines. Paper presented at ESFLCW, London, July 2005

Reinhart, S. M. (2002) *Giving Academic Presentations*. Ann Arbor: University of Michigan Press

Renouf, A., A. Kehoe & J. Banerjee (2007) WebCorp: an integrated system for web text search" in C. Nesselhauf, M. Hundt & C. Biewer (eds.), Corpus Linguistics and the Web. Amsterdam: Rodopi.

Scott, M. (2008). *WordSmith Tools* (version 5) Oxford: Oxford University Press

Swales, J. (1990). *Genre analysis.* Cambridge: Cambridge University Press.

Swales, John. (1981). *Aspects of article introductions.* Birmingham: The University of Aston Language Studies Unit.

Thai English Learner Corpus (TELC).Assumption University, Thailand http://iele.au.edu/corpus/ (site currently unavailable)

Thompson, P. (2000). Citation practices in PhD theses. In Lou Burnard & Tony McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 91-101). Frankfurt: Peter Lang.

Thompson, P. 2005. Aspects of identification and position in intertextual reference in PhD theses', in E. Tognini-Bonelli and G. Del Lungo Camiciotti (eds) *Strategies in Academic Discourse*, pp 31-50. Amsterdam: John Benjamins.

Thompson, P. & Tribble, C. (2001). Looking at citations: using corpora in English for Academic Purposes. *Language Learning and Technology, 5*, 91-105.

Tono Y. (2003) Learner corpora: design, development and applications. In Archer D.**,** Rayson P.**,** Wilson A. **and** McEnery T. (eds.) (2003) *Proceedings of the Corpus Linguistics 2003 Conference (CL 2003). Technical Papers 16*. Lancaster University: University Centre for Computer Corpus Research on Language., 800-809.

Viking Corpus of Student Academic Writing, Portland State University, USA http://web.pdx.edu/~conrads/online_corpus.html

WebCorp (undated) Research and Development Unit for English Studies (RDUES), the School of English, Birmingham City University. http://www.webcorp.org.uk/index.html

Williams, I. 2006. Move, voice and stance in biomedical research article discussions: a pedagogical perspective. In: Neumann, C-P., Pérez-Llantada Auría, C & Plo Alastrué, R. (eds) *Proceedings of the 5th International AELFE Conference*. Prensas Universitarias de Zaragoza, 43-51

Woodward-Kron, R. (2002a). Disciplinary Learning through Writing: An investigation into the writing of undergraduate Education students. Unpublished PhD thesis. Faculty of Education, University of Wollongong, Australia.

Woodward-Kron, R. (2002b). Critical analysis versus description? Examining the relationship in successful student writing. *Journal of English for Academic Purposes* 1 121-143