A New Venture in Corpus-Based Lexicography: Towards a Dictionary of Academic English

Iztok Kosem[1] and Ramesh Krishnamurthy[1]

1. Introduction

This paper asserts the increasing importance of academic English in an increasingly Anglophone world, and looks at the differences between academic English and general English, especially in terms of vocabulary. The creation of wordlists has played an important role in trying to establish the academic English lexicon, but these wordlists are not based on appropriate data, or are implemented inappropriately. There is as yet no adequate dictionary of academic English, and this paper reports on new efforts at Aston University to create a suitable corpus on which such a dictionary could be based.

2. Academic English

The increasing percentage of academic texts published in English (Swales, 1990; Graddol, 1997; Cargill and O'Connor, 2006) and the increasing numbers of students (both native and non-native speakers of English) at universities where English is the language of instruction (Graddol, 2006) testify to the important role of academic English.

At the same time, research has shown that there is a significant difference between academic English and general English. The research has focussed mainly on vocabulary: the lexical differences between academic English and general English have been thoroughly discussed by scholars (Coxhead and Nation, 2001; Nation, 2001, 1990; Coxhead, 2000; Schmitt, 2000, Nation and Waring, 1997; Xue and Nation, 1984), and Coxhead and Nation (2001: 254–56) list the following four distinguishing features of academic vocabulary:

> "1. Academic vocabulary is common to a wide range of academic texts, and generally not so common in non-academic texts.
> 2. Academic vocabulary accounts for a substantial number of words in academic texts.
> 3. Academic vocabulary is generally not as well known as technical vocabulary.
> 4. Academic vocabulary is the kind of specialised vocabulary that an English teacher can usefully help learners with."

Researchers have often attempted to pinpoint academic vocabulary by means of word lists. The two most often mentioned lists of academic words are Xue and Nation's (1984) University Word List and Coxhead's (2000) Academic Word List.

[1] School of Languages and Social Sciences, Aston University
e-mail: kosemi@aston.ac.uk r.krishnamurthy@aston.ac.uk

Word lists have increasingly used corpus methodology. Campion and Elley (1971), and Praninskas (1972) were the first ones to base their word lists on corpus data. But it took almost three decades until another corpus-based word list was created[2], with Coxhead basing the Academic Word List on a 3.5-million word electronic corpus. Also worth mentioning is an English spoken academic word list (Nesi, 2002), which is the first word list based on a spoken corpus (the British Academic Spoken English – BASE – corpus).

Word lists are often used by English for Academic Purposes (EAP) teachers. In a limited amount of time, EAP teachers need to equip their students with the vocabulary used in academic setting, and having word lists that point to the more relevant words is undoubtedly useful.

The design of word lists for academic English usually has one major flaw – it excludes high frequency words (in many cases, specifically the first 2,000 words from West's 1953 General Service List), because it is assumed that the students should know these words already, before starting their studies, and that they need to focus on academic vocabulary. The problem is that several high frequency words are found in academic English with different meanings than in general English. For example, *say* is a high frequency word. However, it refers to speech in general English, but to writing in academic English, e.g. it is often used to introduce a quotation: "X et al (yyyy: pp) **say** …"

3. Dictionaries and Academic English

There are three broad groups of monolingual dictionaries in existence: native-speaker dictionaries, learners' dictionaries, and technical (or domain-specific) dictionaries. However, although academic English is clearly increasing in significance globally, there is currently no dictionary of academic English on the market.

All of the existing dictionaries available to students deal with general English, thus academic vocabulary receives less prominent treatment. College dictionaries, which are especially popular in the US, are aimed at native-speaker students and have a lot of the features of native-speaker dictionaries in general. In fact, the only characteristic – other than the target user – that distinguishes between 'student' and 'general' types of dictionary seems to be the number of entries (Béjoint, 2000). In the UK, the *Compact Oxford English Dictionary for University and College Students* is very similar to the *Compact Oxford English Dictionary of Current English.* The publisher's promotional material tries its best to make a clear distinction between the two dictionaries, but a quick inspection reveals that the only real difference lies in the additional material (e.g. sections on academic writing, how to write a CV, *etc.*) in the 'student' version, not in the dictionary macrostructure or microstructure.

Non-native speaker students are in even greater need of a dictionary of academic English. They currently rely heavily on advanced learners' dictionaries and bilingual dictionaries, however these dictionaries do not focus on academic words/meanings, but

[2] Xue and Nation's University List was not deemed corpus-based, as the authors have actually combined four previous word lists, of which only two were based on corpora (Campion and Elley, 1971; Praninskas, 1972).

on high frequency words. Thus, native-speaker dictionaries and sometimes technical dictionaries are used to fill this gap in vocabulary coverage. However, both of these types of dictionary contain much more demanding definitions. Technical dictionaries suffer from an additional problem: because they focus on a single domain, they lack the vocabulary of general academic English, and also the vocabulary from neighbouring subject fields.

The recently published *Longman Exams Dictionary* is the first dictionary to make use of an academic word list (the Academic Word List by Coxhead, 2000). Regretfully, the authors of the dictionary did not scrutinize Coxhead's approach carefully enough, and have simply labelled the words from her list. As a consequence, the dictionary has inherited several problems from the Academic Word List, such as:

a) Words that are among the 2000 most frequent words in the General Service List (West, 1953) are excluded from the Academic Word List, and hence are not labelled academic (AC) in the dictionary. However, many of these words, like *say*, *argue*, *note*, have special meanings in academic English, a fact which is evidently recognized, because they are dealt with in the Writing Handbook section of the dictionary, where the focus is on academic writing. But because they are not labelled as academic words in the dictionary, the user is not directly alerted to their importance in academic English – he or she needs to get to the Writing Handbook first.

b) The Academic Word List identifies only words, not senses. Therefore, polysemous words in the dictionary lack information on which senses or uses of the word are academic. For example, *abandon*, the very first word in the dictionary labelled as academic, has 6 senses (Figure 1). How is the user to know which of the senses are used in academic English? Are all the senses 'academic'? Or is the user to deduce, from Nation's suggestion (2001: 209) that words in academic vocabulary refer to abstract ideas, that sense 4 ("to stop having a particular idea, belief or attitude") is academic?

Of course, if this really was a dictionary of academic English, the academic senses would be given first.

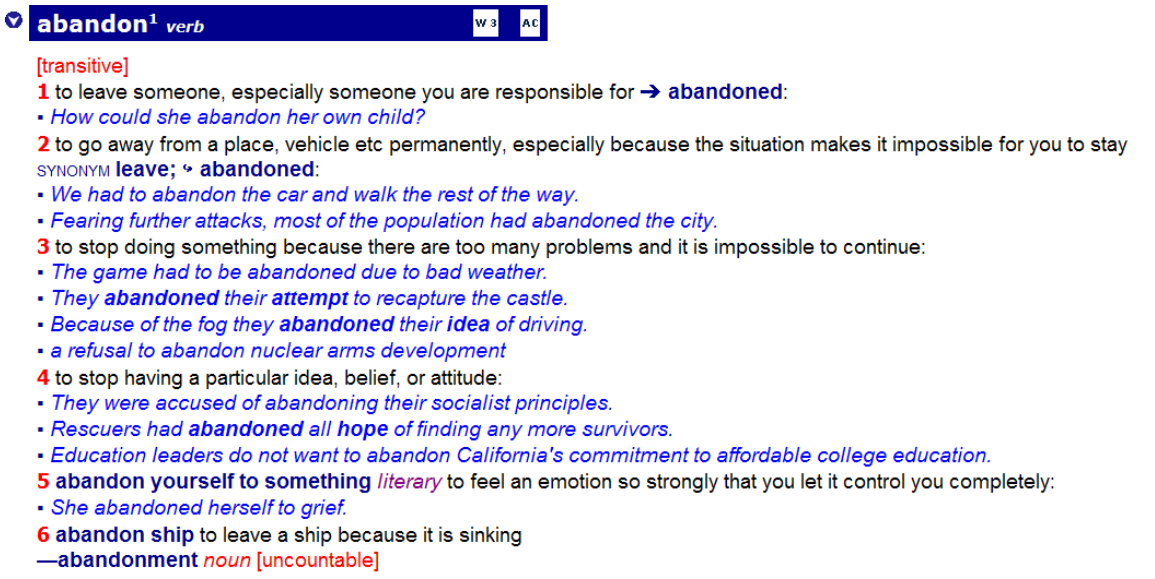

Figure 1: The entry for *abandon* in the Longman Exams Dictionary.

c) The Academic Word List is based on a rather small (3.5 million words) corpus, considering the wide range of disciplines that it attempts to cover (Arts, Commerce, Law, and Science, each sub-divided into 7 subject areas). Such an amount of data (*c.* 125,000 words per subject) does not meet lexicographic needs, and any claim that a dictionary based on this corpus made would carry little weight. The corpus also contains many old texts (e.g. the Brown and LOB corpora consist of texts from 1961) and incomplete texts (114 texts are listed as such, but this number does not include individual chapters extracted from textbooks). Furthermore, most of the texts (64 percent) were published in New Zealand (as against 20 percent in Britain, 13 percent in the US, 2 percent in Canada, and 1 percent in Australia), whereas surely most of the widely used academic texts (in English) are published in the UK or USA – although Coxhead (2000: 220) argues that at least some authors might not come from the country they publish in.

Examples in the dictionary also deserve some criticism. The dictionary entries offer few examples that are recognisably from academic texts (only the Writing Handbook does). For example, the majority of examples at the entry for *abandon* above seem to be from registers other than academic: informal conversation, journalism, or literature (indeed, sense 5 is marked *literary*). Many examples seem to feature weak collocates: in the whole entry for *abandon*, there is no example with *plans* (the commonest noun collocate by far in the Bank of English 448 million word corpus).

Hence university students, both native and non-native speakers, are still searching in vain for a single dictionary that will serve their academic English purposes, and are therefore compelled to learn how to use several dictionaries. While the *Longman Exams Dictionary* incorporates a corpus-based academic word list, it is clear that both a better corpus resource and better lexicographical treatment is needed.

4. Corpora and Academic English

In the last two decades, general English corpora have made a considerable contribution to lexicography. The beginning of the corpus era in lexicography was marked by the COBUILD project in the 1980s, which produced several dictionaries and grammar books, the *Collins Cobuild English Language Dictionary* being the best known of them all. It is worth pointing out that the COBUILD dictionary was one of the rare dictionaries that has actually used corpus "systematically and extensively" (Herbst, 1996: 322), i.e. it was corpus-driven.

The use of corpora was soon adopted by other EFL dictionaries, and by native-speaker and bilingual dictionaries, and was advertised as a unique dictionary feature, but corpora were in fact being used more as an additional resource, rather than as the driving force. Such corpus-based dictionaries have now become the norm in modern lexicography (Kilgarriff, 2000). Publishers no longer gain an advantage over their competitors by using corpus data, but they can find themselves at a serious disadvantage by not doing so.

Lexicographers use corpora to obtain information on the frequency of words and phrases, to discover their meanings, and to find examples of authentic usage. Some dictionaries, especially EFL dictionaries, make more extensive use of corpora to provide detailed information on collocations, grammar patterns, and (more recently) synonyms

and antonyms (via software such as SketchEngine). EFL dictionaries have consistently pioneered new ways of using corpus data. While other types of dictionary used only general corpora, EFL dictionaries initiated the use of learner corpora. For example, the *Longman Dictionary of Contemporary English* (1995) claimed to use learner corpora for its Usage Notes. Similarly, the second edition of the *Macmillan English Dictionary for Advanced Learners* (2007) used the International Corpus of Learner English (ICLE; see also Granger, 2003) to identify the most common errors for its "Get it Right" boxes.

Corpora have also had a more indirect influence on lexicography through their impact on linguistic theory, which has subsequently contributed to changes in dictionary microstructure; for example the notion of collocation, first mentioned by Firth (1951) but more fully developed by Sinclair (1991) after the emergence of electronic corpora, is now a common feature of many dictionaries.

The variety of dictionaries produced from corpus data is clear evidence of the many different ways in which it can be utilised: general dictionaries (e.g. EFL dictionaries, native-speaker dictionaries, and bilingual dictionaries), specialised dictionaries (e.g. dictionaries of collocations, production dictionaries, dictionaries of idioms, dictionaries of phrasal verbs), and thesauri. While corpora have been used to improve general dictionaries and thesauri, they are actually responsible for the creation of certain new types of specialised dictionaries, such as dictionaries of collocations.

But so far, dictionaries have only made use of general corpora (or subsets of data from general corpora) and learner corpora. They have not yet made significant use of specialised corpora. Sometimes the problem may lie in the nature of the dictionary itself. In the case of technical or specialist dictionaries, Landau (1974: 242) points out that "the meanings of scientific entries… are imposed on the basis of expert advice", which often results in a more prescriptive type of dictionary and militates against the use of corpus data as the basis for such dictionaries.

Similarly, corpora of academic English, or EAP (English for Academic Purposes) corpora, have not been used to produce a dictionary. They have been used mainly in EAP teaching (Johns, 1989, 1991; Flowerdew, 1998; Thurston and Candlin, 1998). They have also been used more recently to produce materials for university teachers and students; for example, the BASE corpus was used in the development of the Essential Academic Skills in English (EASE) series of multimedia CD–ROMs.

It is interesting to compare the different ways in which EFL and EAP have utilised corpus resources for teaching and learning. On the one hand, EFL initially benefited from corpora via lexicography, and only later produced corpus-based EFL materials and introduced corpora into teaching. In other words, lexicographers have presented the world of corpora to EFL material designers and teachers. By contrast, EAP material designers and teachers have had little help from lexicographers, but have accessed corpora directly.

Lexicographers have yet to realize the potential of EAP corpora for dictionary-making. Of course, publishers may argue that academic English does not have enough users, and/or that existing dictionaries already cater for these users' needs; however, it has been shown above that neither of these arguments is valid.

A totally different question is whether existing EAP corpora are adequate for the production of a dictionary of academic English. Academic subcorpora of general corpora represent the easiest available option, because they exist, and have already been used as

part of other lexicographic ventures. For example, the 100-million-word British National Corpus contains around 16 million words (15.8 percent) of academic texts (BNC World Index), including some transcripts of lectures. But as Thompson (2006) notes, academic texts in general corpora are often incomplete, because of the use of text extracts rather than complete texts in corpus compilation. In addition, there are issues concerning representativeness, as some disciplines tend to be better represented than others (which is probably the result of the availability of data rather than corpus design principles). Such deficiencies in academic subcorpora of general corpora, combined with the fact that many academic publications are now becoming available in electronic form, means that there is no longer any need to use general corpora for EAP purposes.

While existing EAP corpora are much smaller than general corpora, the variety of data they cover is impressive. The data covers written and spoken English, e.g. the British Academic Written English (BAWE) corpus, and the Michigan Corpus of Academic Spoken English (MICASE). The written data includes articles, essays, theses, monographs, textbooks, course packs and laboratory manuals. The spoken data includes a wide range of events, from lectures and seminars to tutorials and student presentations.

The authors of the texts can be language learners (e.g. ICLE), students (e.g. the Michigan Corpus of Upper-Level Student Papers – MICUSP), or academics (e.g. The Hyland corpus; see also Hyland, 2002); native speakers of English (e.g. The Reading Academic Text corpus) or non-native speakers of English (e.g. The Uppsala Student English Corpus; see also Axelsson, 2003).

Hence, lexicographers cannot complain about the lack of EAP corpus data. However, there are three issues that stand in the way of utilizing this data. Firstly, there is the problem of accessibility, or rather, inaccessibility. This is without a doubt one of the main shortcomings of many EAP corpora, especially the ones containing published material. The problem is that EAP researchers often compile corpora for their own research and therefore do not bother to obtain copyright permission. Hence, the considerable efforts put into the compilation of the corpora have benefits only for the compiler. Spoken EAP corpora seem to be more accessible. For example, BASE and MICASE are available both online and on a CD–ROM.

Secondly, the EAP corpora that are currently accessible – especially the written ones – seem to focus only on particular disciplines, whereas EAP corpora need to cover a variety of texts from a range of disciplines. The available spoken EAP corpora, more specifically BASE and MICASE, seem to contain a variety of speech events in several different disciplines (and a further advantage is that the compilers of both corpora have used the same classification system).

Thirdly, existing EAP corpora are just not large enough. They are much smaller than general corpora. This is especially true of written EAP corpora that contain target texts. Coxhead's (2000) Academic Corpus is the largest corpus of this kind, containing around 3.5 million words, but (as discussed earlier) even this corpus is inadequate for lexicography. The existing spoken EAP corpora are smaller still. A larger corpus of spoken academic English would be especially valuable for determining the differences between spoken academic English and spoken general English, and between spoken academic English and written academic English.

Our analysis has shown the inadequacy of existing EAP corpora for the purposes of compiling a dictionary of academic English; the main shortcomings are their

inaccessibility, lack of appropriate contents (coverage), and insufficient size. This suggests that a completely new corpus of academic English will need to be compiled to meet the needs of lexicographers.

Creating a new corpus of academic English will enable lexicographers to adopt a design more suitable for dictionary purposes. In the next section, some of the issues that need to be considered in the design process will be discussed, and as the ACORN project at Aston University includes the creation of an EAP subcorpus, some of the potential solutions will also be presented.

5. A New Corpus for a Dictionary of Academic English

5.1 Aston University: The Aston Corpus Network (ACORN) Project

The initial phase of corpus activities at Aston University in the mid-1990s focussed on English for Specific Purposes (ESP), creating business-, engineering- and health-related corpora and stand-alone concordancing software, and also worked on Computer Assisted Language Learning (CALL) and data-driven learning projects. However, as staff retired or moved on, the activities ceased. In 2005, the Aston Corpus Network project (ACORN: http://corpus.aston.ac.uk) re-launched corpus work at Aston.

The project obtained start-up funding in 2006–7 from HEFCE via the Flexible Learning Development Centre at Aston, hence the initial focus was on enhancing language teaching and learning for HEFCE-funded students, by providing corpora in English, French, German and Spanish, and parallel texts for Translation Studies, with web-based software to allow flexible access.

5.2 The EAP Subcorpus

The initial ACORN data collection included some academic texts by Aston staff and students (Master's assignments and dissertations, PhD theses, and articles written by staff), but they were not of particular significance within the overall aims of the project at the time. The arrival of a researcher whose focus was on EAP stimulated the acquisition of academic texts, but the lack of a suitable corpus specifically for EAP research prompted the decision to design and create a complete EAP subcorpus.

The EAP subcorpus project offered the unique opportunity, with associated challenges, to combine lexicographic, research and teaching requirements. In the past, general language corpora have usually been created initially for lexicographic purposes, and only later made available to researchers and teaching materials designers. By contrast, EAP corpora have been created primarily for personal research purposes. However, neither type of corpus was designed specifically for pedagogy, for the needs of teachers and students. Therefore, designing and creating an EAP subcorpus with all three aspects in mind (lexicographic, research, and pedagogic) would ensure greater usefulness and wider applicability.

5.3 EAP Subcorpus Design

Data in existing EAP corpora can be broadly considered in terms of four dimensions: domain (academic subject or discipline), mode (written, spoken), genre (e.g. lecture, journal article, textbook; essay, seminar presentation, exam script, thesis) and level (undergraduate, Master, PhD, 'expert'). Accessible EAP corpora usually cover two of these dimensions, some inaccessible ones claim to cover three, but none of them covers all four. The ACORN EAP subcorpus design will attempt to cover all four dimensions. In addition, we may need to find an appropriate parallel range of levels for non-native speakers (based on traditional EFL categories such as beginner, intermediate, and advanced; or European Framework categories; or University programme levels such as Junior Year Abroad, pre-sessional, undergraduate, Pre-Master's, *etc.*).

5.4 Domains

There is no universally agreed classification system for academic subjects. Institutions such as libraries and universities tend to use many categories. For example, the Joint Academic Classification of Subjects used by the Higher Education Statistics Agency (HESA, http://www.hesa.ac.uk/jacs/jacs.htm) lists 19 subject groups and 142 subgroups, or principal subjects. On the other hand, compilers of EAP corpora seem to prefer fewer categories, at least at the top level. The (Coxhead) Academic Corpus, MICASE, BASE corpus, and BAWE corpus all use only 4 top-level categories. One of the reasons for using broader categories is probably to disguise the unequal distribution of data, or lack of data from certain subject-domains. The Academic Corpus, for example, uses 4 top-level categories and 28 sub-categories, but none of them includes Medicine. A preliminary design for an Academic Corpus at Birmingham University in the early 1990s (Krishnamurthy, personal communication) had 7 domains: arts, education, social sciences, medicine, law, sciences, and engineering.

Some close parallels can be drawn between the sub-corpus domain classification issues faced by corpus compilers, and the lexicographic problems involved in devising labels to indicate technical vocabulary. A corpus with data from a wide variety of domains will lead to problems in deciding how many sub-corpora to create. Similarly, a native-speaker dictionary will have to select from a large number of possible technical or subject-field labels. The size of the corpus can be compared with the number of dictionary entries. A smaller corpus will yield fewer lexical items and potential technical terms, just as for example an EFL dictionary will have fewer technical entries, and therefore require fewer subject-field labels.

However, it is often difficult to determine the exact number or scope of technical labels used in particular dictionaries, as they do not always provide even a list of all the labels, let alone offer any explanation for the labels and abbreviations used as labels. For example, Table 1 shows the 42 subject-field labels found in the entries from *a* to *absolute* in the *Collins English Dictionary*. Clearly, some labels are broader than others, but this is a reasonable indication of the number and range of subject-field labels used in a typical native-speaker dictionary.

More information is usually available in EFL dictionaries. Bogaards (2003) identifies eight subject labels in the *Macmillan English Dictionary for Advanced Learners*: *business, computing, journalism, legal, linguistics, medical, science* and *technical*. It is clear from these labels that certain subject vocabularies are much better represented than others.

Accounting	Grammar	Optics
Archaeol	Judaism	Pathol
Astronomy	Law	Philosophy
Athletics	Linguistics	Phonetics
Bible	Logic	Physics
Biology	Maths	Printing
Bookmaking	Med	Psychoanal
Botany	Military	Psychol
Chem	Mormon Church	Rugby
Computing	Motor Racing	Sport
Cricket	Mountaineering	Statistics
Films	Music	Stock Exchange
Geography	Nautical	Theatre
Golf, basketball…	Old Testament	Zoology

Table 1: Subject-field labels in the Collins English Dictionary (2004) (entries *a–absolute*).

Considering the problems involved in the classification of academic subjects and the labelling of subject-field vocabulary, it would perhaps be prudent to use the same categories for both corpus and lexicographic purposes. This may require the adoption of a classification system with a lot of categories, similar to the one used by HESA. Further categorization, both higher-level and lower-level, would be useful not only to lexicographers, who could use lower-level categories as indicators in the entries, but also (or perhaps especially) to researchers and teachers.

Using such a design could contribute significantly to the user-friendliness of a dictionary of academic English, especially in an electronic version. For example, it could offer a customizable feature which would enable the user to arrange the information provided in the entries by selecting a specific subject-field label.

5.5 Modes

The ACORN EAP subcorpus was initially planned to focus mainly on written texts, due to lack of funding and the complexities of obtaining spoken data. However, many members of staff in the School of Languages and Social Sciences have recently expressed their interest in spoken academic English, and their willingness to assist in the collection of spoken data. Therefore, it has been decided to include as much spoken data as

possible, and perhaps supplement it by referring to existing and accessible EAP spoken corpora, such as MICASE and BASE.

5.6 Genres

Academic English consists of a wide range of different speech events and written texts. Existing spoken data in EAP corpora, for example in MICASE, range from more formal events such as lectures, seminars, and dissertation defences, to less formal ones such as 'office hours' conversations, and study groups. Written EAP corpora include articles, textbooks, manuals, essays, PhD theses, dissertations and many other types of texts. The preliminary design for an Academic Corpus at Birmingham University in the early 1990s (mentioned earlier: Krishnamurthy, personal communication) included the following genre categories: tutorial, lecture, seminar, conference paper (spoken), conference paper (written), undergraduate text, postgraduate text, manual, article, technical journal, popular journal, essay, thesis, technical report, handout, letter, exam paper.

Just as we discovered earlier with the classification of academic subjects, academic genres also lack a universal classification system. Part of the problem lies in the fact that "there is still only limited consensus about what the concept fundamentally entails" (Moore and Morton, 2005: 49). In addition, genre categories can differ across subject-fields. The difference may involve essential characteristics, such as structure, length, and level of formality, or may simply be a difference in nomenclature. For example, the concept of *assignment* in Linguistics may be totally different to the concept of *assignment* in Business; or an *essay* in one discipline may be called an *assignment* in another.

Decisions about genre categories are complex and should not be made in advance. In the ACORN project, there is fortunately no immediate need to do that, as categories can be determined retrospectively and later in the corpus compilation process, preferably with the help of informants from the various subject-fields. In addition, the analysis of corpus texts could provide internal linguistic criteria that would provide additional information to assist in the categorization.

5.7 Levels

The data will need to be collected at a wide variety of levels in order to adequately cover the whole range from student texts to expert texts. Expert texts, or target texts, are necessary for determining the level that the user of a dictionary of academic English needs to achieve. The notion of target text in academic English is rather problematic to define, as students are far from homogeneous, not only in their first language and cultural background, but also with regard to their goals. For example, the majority of students do not pursue studies beyond the undergraduate level; most leave university after their first degree, and proceed to work in non-academic jobs. Therefore texts such as academic books and journal articles might set too high a standard, one they would not be aiming to achieve. Hence, final-year undergraduate dissertations may be more suitable texts for their purposes.

Having said that, we need to bear in mind that all students do have at least one thing in common: they all need to read academic books and journal articles during their studies, and therefore require at least the decoding aspect of an academic dictionary. Such texts therefore do need to be collected, even for the benefit of undergraduates. However, as it is often difficult to obtain copyright permission for published academic books and journal articles, we may need to compensate for the shortage of published materials with PhD theses, MA dissertations, and final-year undergraduate dissertations.

A distinction will also need to be made between native-speaker data and non-native-speaker data. However, the non-native-speaker student data that we will be collecting is different from the data found in learner corpora such as ICLE. Non-native-speaker student texts in the ACORN EAP subcorpus will be produced by learners of academic English, while existing 'learner corpora' define their learners as general English language learners. The general language learner corpus texts probably do not contain citations and references, and are shorter, different in structure, and may belong to different genres (e.g. creative writing); and few (if any) learner corpora contain spoken data.

Defining a learner of academic English is not at all straightforward. Should one consider a non-native-speaker pre-sessional student as a typical learner of academic English, or perhaps a first-year undergraduate student? What about fourth-year native-speaker students who are still struggling with academic English? As there is no simple answer, we will need to collect the data now at all possible levels, and attempt to classify the data by levels later, perhaps using a combination of external (year of study, number of years of English medium education, *etc.*) and internal (linguistic) criteria.

5.8 Proficiency

If we are to discover the lexico-grammatical errors that learners of academic English make, in order to provide suitable information and help in the dictionary of academic English, we will also need to collect texts produced by less able students, unlike BAWE, which focussed on "student assignments *of a good standard*" (our emphasis).

5.9 Complete texts

We mentioned earlier that one deficiency in general language corpora, and in some (often related) EAP corpora, is that they include partial or incomplete texts, selected chapters from books, *etc*. We regard the collection of complete texts as an important policy principle, however we acknowledge that this will lead to other problems in the data. For example, should we include References sections and Bibliographies, or will they distort both frequencies and lexical patterns? Similarly, will quotations from 'expert texts' skew the content of 'student texts' in terms of both language proficiency and level of mastery of academic English conventions? We may need to consider tagging and annotating such elements and excluding them from analyses, if we decide to use internal linguistic criteria in any automatic processes of classification.

5.10 Data sources

While Aston University students and staff are expected to contribute a substantial proportion of the texts, other UK universities are welcome to contribute to the ACORN EAP subcorpus. ACORN has already acquired data from British Council-funded alumni, and other non-Aston students. Individuals may also donate their personal EAP research corpora, compiled by themselves alone or in collaboration with Aston. This will help to recycle collections made at considerable effort, but that will otherwise remain forever inaccessible to the rest of the EAP community. Aston MSc TESOL distance learning students have expressed their willingness to create datasets of texts written by students in the countries where they are currently teaching. Our on-campus MA TESOL students may also wish to participate when they return to their countries at the end of their course.

5.11 Data collection and obtaining consent

Initially, students in the School of Languages and Social Sciences (LSS) at Aston University were contacted directly, or through their lecturers. However the success rate of this approach was rather poor, so a new strategy was considered. Coincidentally, LSS recently instituted a policy of insisting on electronic submission of coursework by undergraduates, to facilitate plagiarism detection. This offered another opportunity for data collection, and an ACORN consent form was issued to undergraduates which allowed their work to be included in a subcorpus of student writing, intended for research. By seeking consent at the beginning of a student's course, and covering their entire future studies at Aston, we can obtain access to their work as soon as it is submitted, and the student only needs to give permission once.

As all the Schools at Aston now collect undergraduate students' material electronically, this corpus collection procedure can easily be extended to the other Schools. Similar consent forms have also been issued to all taught postgraduates in LSS at registration, and no problems have been experienced so far in obtaining consent. We have no intention at the moment of paying for students' work (as BAWE compilers have done), as that would require additional funding which we cannot afford.

5.12 Metadata

For data submitted by Aston students, metadata should be relatively easy to obtain. The student consent form includes permission to email and/or interview students subsequently, to gather information such as their mother tongue, programme of study, gender, date of birth, *etc*.

Every student electronic text will eventually have a filename which incorporates information such as the student-id, date of submission, and module code (which encapsulates year of study and subject domain).

6. Conclusion

The EAP Subcorpus aspect of the ACORN project at Aston University has grown from the incidental collection of academic texts as part of the creation of general language corpora in English, French, German, and Spanish, aimed at language learning and teaching, into a separate project with its own requirements, momentum, policies, and goals.

The research interest in compiling a corpus suitable as the basis for a corpus-driven dictionary of academic English has offered us the opportunity to design a corpus with a wide range of potential uses in lexicography, pedagogy (learning, teaching, assessment, progression monitoring, curriculum development) and materials design and production. The EAP subcorpus project is in its very early days, so we look forward to reporting on progress at future conferences and in future publications.

So far, our activities have been conducted with relatively small-scale internal funding, as part of the general revival of corpus activities at Aston University. However, if we are going to be able to achieve even a modest part of our ambitious aims set out above, we will need to attract a considerable amount of external funding from research councils and similar funding bodies.

References

Dictionaries

Collins Cobuild English Language Dictionary. (1987). Worthing: HarperCollins.

Collins English Dictionary (Desktop Edition). (2004). 1st edition. Worthing: Harper Collins.

Compact Oxford English Dictionary for University and College Students. (2006). Oxford: Oxford University Press.

Compact Oxford English Dictionary of Current English. (2005). 3rd edition. Oxford: Oxford University Press.

Longman Dictionary of Contemporary English. (1995). 2nd edition. Harlow: Longman.

Macmillan English Dictionary for Advanced Learners. (2007). 2nd edition. Oxford: Macmillan.

Oxford Dictionary of English. (2005). Oxford: Oxford University Press.

Corpora

Axelsson, M.W. (2003). The Uppsala Student English corpus (USE) – Manual. Uppsala University, Department of English. Available on-line from http://ota.ahds.ac.uk/ (accessed: 21 June 2007).

BASE: British Academic Spoken English corpus. http://www2.warwick.ac.uk/fac/soc/celte/research/base/ (accessed: 24 June 2007).

BAWE: British Academic Written English corpus. http://www2.warwick.ac.uk/fac/soc/celte/research/bawe/ (accessed: 24 June 2007).
ICLE: International Corpus of Learner English. http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm (accessed: 25 June 2007).
MICASE: The Michigan Corpus of Academic Spoken English. http://quod.lib.umich.edu/m/micase/ (accessed: 13 June 2007).
MICUSP: The Michigan Corpus of Upper-Level Student Papers. Available on-line from http://www.micusp.org/ (accessed: 21 June 2007)
The Bank of English Corpus. titania.bham.ac.uk. Accessed June 25–29th 2007.
The Reading Academic Text Corpus. http://www.rdg.ac.uk/app_ling/corpus.htm (accessed: 24 June 2007).
The Sketch Engine. Available on-line from http://www.sketchengine.co.uk/ (accessed: 25 June 2007).

Other Literature

Béjoint, H. (2000). Modern Lexicography: An Introduction. Oxford: Oxford University Press.
BNC World Index. Available Online from: http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/home/BNC_WORLD_INDEX.ZIP (accessed: 12 June 2007).
Bogaards, P. (2003). MEDAL: a Fifth Dictionary for Learners of English. *International Journal of Lexicography 16*, 43–55.
Campion, M. and W. Elley. (1971). An Academic Vocabulary List. Wellington: New Zealand Council for Educational Research.
Cargill, M. and P. O'Connor. (2006). Developing Chinese scientists' skills for publishing in English: Evaluating collaborating-colleague workshops based on genre analysis. *Journal of English for Academic Purposes 5*, 207–221.
Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly 34*, 213–38.
Coxhead, A. and P. Nation. (2001). The specialised vocabulary of English for Academic Purposes, in J. Flowerdew and M. Peacock (eds.) Research Perspectives on English for Academic Purposes. Cambridge: Cambridge University Press.
EASE: Essential Academic Skills in English (interactive CD–ROMs). http://www.ease.ac.uk/ (accessed: 24 June 2007).
Firth, J.R. (1951). Modes of Meaning, in J.R. Firth (ed.) Papers in Linguistics 1934–1951. Oxford: Oxford University Press.
Flowerdew, L. (1998). Integrating `expert' and `interlanguage' computer corpora findings on causality: discoveries for teachers and students. *English for Specific Purposes 17*, 329–45.
Graddol, D. (1997). The Future of English. London: British Council.
Graddol, D. (2006). English Next. London: British Council.
Granger, S. (2003). The International Corpus of Learner English: A New Source for Foreign Language Learning and Teaching and Second Language Acquisition Research. *TESOL Quarterly 37*, 538–45.

Herbst, T. (1996). On the way to the perfect learners' dictionary: a first comparison of OALD5, LDOCE3, COBUILD2 and CIDE. *International Journal of Lexicography 9*, 321–57.

HESA (Higher Education Statistics Agency). Joint Academic Classification of Subjects. http://www.hesa.ac.uk/jacs/jacs.htm (accessed: 26 June 2007).

Hyland, K. (2002). Activity and Evaluation: Reporting practices in academic writing, in J. Flowerdew (ed.) Academic Discourse, pp. 115–30. Harlow: Pearson.

Johns, T.F. (1989). Whence and whither classroom concordancing? in T. Bongaerts et al. (eds.) Computer Applications in Language Learning, pp. 9–33. Dordrecht, the Netherlands: Foris.

Johns, T.F. (1991). From printout to handout: grammar and vocabulary teaching in the context of data-driven learning, in T.F. Johns and P.K. King (eds.) Classroom Concordancing (English Language Research Journal 4), pp. 27–45. University of Birmingham: English Language Research.

Kilgarriff, A. (2000). Business Models for Dictionaries and NLP. *International Journal of Lexicography 13*, 107–118.

Landau, S.I. (1974). Of Matters Lexicographical: Scientific and Technical Entries in American Dictionaries. *American Speech 49*, 241–44.

Moore, T. and J. Morton. (2005). Dimensions of difference: a comparison of university writing and IELTS writing. *Journal of English for Academic Purposes 4*, 43–66.

Nation, I.S.P. (2001). Learning Vocabulary in Another Language. Cambridge: Cambridge University Press.

Nation, P. (1990). Teaching and Learning Vocabulary. Boston: Heinle and Heinle.

Nation, P. and R. Waring. (1997). Vocabulary size, text coverage and word lists, in N. Schmitt and M. McCarthy (eds.) Vocabulary: Description, Acquisition and Pedagogy, pp. 6–19. Cambridge: Cambridge University Press.

Nesi, H. (2002). An English spoken academic word list, in A. Braasch and C. Poulsen (eds.) Proceedings of the Tenth EURALEX International Congress, pp. 351–58. Copenhagen: Center for Sprogteknologi.

Praninskas, J. (1972). American University Word List. London: Longman.

Schmitt, N. (2000). Vocabulary in Language Teaching. Cambridge: Cambridge University Press.

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Swales, J. (1990). Genre Analysis: English in Academic and Research Settings. Cambridge: Cambridge University Press.

Thompson, P. (2006). Assessing the contribution of corpora to EAP practice, in Z. Kantaridou et al. (eds.) Motivation in Learning Language for Specific and Academic Purposes. Macedonia: University of Macedonia.

Thurston, J. and C.N. Candlin. (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes 17*, 267–80.

West, M. (1953). A General Service List of English Words. London: Longman, Green and Co.

Xue, G.Y. and I.S.P. Nation. (1984). A University Word List. *Language Learning and Communication 3*, 215–29.