ESP corpus construction: a plea for a needs-driven approach

Nesi, H.

ESP corpus construction: a plea for a needs-driven approach

Bâtir des corpus en anglais de spécialité : plaidoyer pour une approche fondée sur l'analyse des besoins

Hilary Nesi

Coventry University, United Kingdom

KEY WORDS

English for Specific Purposes, English for Academic Purposes, BASE corpus, BAWE corpus, Engineering Lecture Corpus.

ABSTRACT

This paper presents the case for compiling small, manually-sampled corpora, rich in contextual information, for research, teaching and learning in English for Specific Purposes (ESP) and English for Academic Purposes (EAP). Large datasets such as the British National Corpus, more recent web-derived corpora, and the web itself are major sources of information about the lexis and grammar of general English, and progress has been made in automatic corpus compilation methods; it is possible to mine the internet for documents on specified topics, within specified domains, and with the linguistic features associated with particular genres. Many types of text which are of interest to ESP and EAP practitioners are absent from these general corpora, however, or are under-represented. Their compilation methods do not capture all the background information ESP and EAP practitioners need in order to understand the linguistic choices speakers and writers make. This paper suggests that the process of designing a good corpus for ESP or EAP is similar to the process of needs analysis, and illustrates this process with examples from various corpus projects, showing how contextual information can be added to corpora in the form of textual annotations or as supplementary material.

MOTS CLÉS

Anglais de spécialité, anglais universitaire, corpus BASE, corpus BAWE, Engineering Lecture Corpus.

RÉSUMÉ

Cet article plaide en faveur de la compilation de petits corpus, rassemblés manuellement, riches en données contextuelles pour la recherche, l'enseignement et l'apprentissage en anglais de spécialité (ASP) et en anglais universitaire. Les grands corpus, comme le British National Corpus, des corpus plus récents tirés de la Toile et la Toile elle-même constituent des sources d'information essentielles sur le lexique et la grammaire de l'anglais général. Il est possible de recueillir sur la Toile des documents sur des sujets spécifiques, relevant de domaines précis et comportant des traits linguistiques associés à des genres spécifiques. Cependant, de nombreux textes intéressant les praticiens de l'ASP ou de l'anglais universitaire ne figurent pas dans les grands corpus ou bien sont sous-représentés, compilés sans prendre en compte les informations contextuelles dont les praticiens de l'ASP ou de l'anglais universitaire ont besoin pour comprendre les choix linguistiques des locuteurs et des rédacteurs. Cet article suggère qu'il convient de s'inspirer de l'analyse des besoins pour construire des corpus adaptés : à travers plusieurs exemples, il montre la façon dont des informations contextuelles peuvent être ajoutées aux corpus sous la forme d'annotations textuelles ou d'annexes.

1. Introduction

A huge amount of corpus data is now available for English language teachers and learners, making it difficult for them to decide which kind of corpus is most suited to their needs. The corpora that they probably know best are the large, publicly accessible general corpora which inform general reference works and textbooks; the ones that aim to cover the English language as a whole and to represent many different categories of text. The 100 million word British National Corpus (BNC), created in the 1990s, is a famous example of this type and has been used extensively for language learning purposes, but at 100 million words it is too small to contain large numbers of texts within each of its subcomponents, and it is now too old to represent modern usage in the more fast-moving fields. Corpora created more recently, such as the 440 million word Corpus of Contemporary American English (COCA), the 470 million word Synchronic English Web Corpus, the 550 million word WordBanks Online (formerly known as the Bank of English), or the TenTen corpus family, so-named because each corpus in the family aims at a target size of 10^{10} (10 billion) words (Jakubíček *et al.* 2013), have much more potential to supply examples of specific up-to-date usage. The 2013 English version of TenTen in SketchEngine (enTenTen13), for example, contains 985,220 instances of the term *search engine,* whereas there are only two instances of *search engine* in the entire BNC. Also, alongside these web-derived corpora, the web itself can be treated as a kind of corpus, in ways described by Gatto (2014) and Boulton (2015) amongst others.

All these types of general corpora have the potential to provide data for the teaching and learning of English for Academic Purposes (EAP) and English for Specific Purposes (ESP). Whether these types of corpora are the best kinds for EAP and ESP learners is open to question, however, in view of issues concerning quality, the representation of relevant genres, and the extent of the information provided about the circumstances surrounding text production.

2.The quality of corpus data

Quality was not an issue for the older, and smaller, general corpora, because texts for these corpora were mostly taken from published sources, and were chosen manually as good representative examples of their categories. The newer general corpora which rely much more heavily on web resources take varying approaches to selecting texts for corpus inclusion, and filtering out those that are unsuitable. The Synchronic English Web Corpus, for example, tries to ensure the quality of its holdings by choosing pages according to the DMOZ Open Directory system, which is run by human volunteers who select, evaluate and classify (according to topic) the websites they encounter. Some other web corpora employ automatic techniques, as described by Kilgarriff and Suchomel (2013), to try to filter out documents which are not exactly what the corpus builders want. Quality control is not entirely successful, however, and rogue documents find their way into web corpora; these may just be badly written and communicatively unsuccessful, or they may be reduplications, automatic translations or computer-generated nonsense

texts. The following, for example, are excerpts from concordance lines for spam texts found in enTenTen13. The search term was the rather infrequent word *toothed* (0.4 occurrences per million words):

- Whose valve can be *toothed* slow cooker beef fajitas
- *Toothed* months later, Dr Mongkol accelerated PLAVIX
- It has necessarily *toothed* my gene.
- *toothed* wheel glorious to be thither and hit it big in online casinos.

Nonsense concordance lines such as these might be dismissed as a minor problem, given that they are hugely outnumbered in enTenTen13 by meaningful lines which could provide ESP learners with useful information about the behaviour of *toothed* in biology, botany, medicine and other domains (e.g. "*Toothed* forceps are routinely used to hold skin to suture"). Also on the whole it seems that language teachers and learners are prepared to accept some loss of quality in exchange for size, modernity and accessibility. Gatto (2014: 101) recommends using web data as long as "cautionary procedures are adopted in submitting the query and interpreting the results", and for Data Driven Learning (DDL) purposes, Sha (2010: 377) concludes that the Google web search facility is superior to the BNC "in usability, search speed, the number of solutions and […] preference investigations". One way of dealing with possibly unreliable data might be to ask students to choose and study in detail only the best and most relevant results from their corpus queries, as did Watson Todd (2001).

Nevertheless poor quality data from the web can cause problems, especially for teachers and learners who have little exposure to authentic English language texts from other sources. In Korea, for example, learners rely extensively on e-dictionaries for information about the English language, and some of these dictionaries include sentences extracted from the web to illustrate word use. In theory web extraction should be an efficient way of supplementing traditional dictionary entries with authentic, modern examples, but the effectiveness of this approach depends very much on the quality of the data. Nesi (2012) lists dictionary examples such as the following, extracted from a discussion forum on the web and included in a dictionary entry to illustrate the collocation of "dark" with "coffee" (although learners should have been steered towards "black coffee" as the standard English usage):

- What did Francis arrive the cup before the dark coffee?
- It should change the dark coffee and arrive it through its monolith.
- She'd rather kick furiously than call with George's dark coffee.

3. The range of genres

A further and probably greater concern when using corpora or the web-as-corpus in ESP and EAP contexts is whether these resources can provide learners with examples of the right types of texts, or in other words the types that are most

relevant to their needs. It seems that none of the large general corpora that are available for teaching and learning contain authentic examples of such genres as emails, business meetings or job interviews, although these are likely to be of interest to students of Business English. Neither do they contain EAP genres such as student writing, examination questions or lecture slides. COCA is made up of roughly equal quantities of broadcast material, fiction, and articles from popular magazines, newspapers and academic journals. These categories are described as "genres", using a rather broader sense of the term than that applied in ESP. The Synchronic English Web Corpus and WordBanks Online categorise their holdings according to subject domains such as business and science. Distribution across domains may be fairly even, but the genres represented are not; more than half of WordBanks Online is composed of newspaper articles, for example, and most of the remaining texts come from books and magazines for the general reader, or from TV and radio broadcasts. The BNC, compiled in the days before huge quantities of text could be extracted from the web, contains more varieties of professional and academic discourse such as courtroom and classroom interaction and university lectures, but coverage is patchy, especially in the BNC Sampler – Lee provides a long list of the "Missing or unrepresentative genres" in the appendix to his article (2001: 63-64).

Despite its size, the internet is probably even less representative of human language output. We can never know the full extent of the web, but most kinds of authentic private interaction are unlikely to be available in any quantity, and scripts for spoken genres, if they are authentic, may not be exact representations of what was originally said (having been transcribed to reveal the message rather than linguistic features). It is still difficult to find webpages by genre rather than by content, or to search for lexicogrammatical patterns within texts. Gatto (2014) suggests some ways in which language learners and researchers can filter queries using the information encoded in website addresses. Searches can be restricted to particular regions by specifying national domains, for example, or can be made to focus on sites which contain academic content, from academic publishers, or with an academic domain name (all possible using the Google Advanced Search options). However, although Gatto (2014: 93) considers such filtering strategies to be "particularly useful in turning the web into a virtual corpus on demand", web query results do not facilitate language study because they are not laid out to show context, in the manner of concordance lines, and the categories of text that are extracted by these means are not as specific as one would want for ESP and EAP. Searching within the websites of named journals, or in a repository of PhD theses such as the British Library EThOS service, is probably a more effective way of finding ESP and EAP texts relating to specific professions or academic disciplines, although documents discovered in this way cannot be interrogated directly online, and have to be downloaded and analysed using an offline corpus query tool such as Laurence Anthony's *AntConc*. It should also be borne in mind that by no means all ESP and EAP learners need to write or even read theses or journal articles. The genres that they most need to engage with may be impossible to locate on the web.

Some progress is being made in the field of information and language

technology towards the automatic differentiation of texts according to their "genre", but technological experts understand this term rather differently from ESP practitioners. Whereas genre in ESP is generally considered from a sociological perspective, in terms of the communicative purposes specific to a given discourse community, as described by Swales (1990), information and language technologists perceive genre as "a stylostatistically or philologically observable objective characteristic of texts" (Karlgren 2012), and thus group texts into genres according to the statistical distribution of their structural and lexicogrammatical features. Some of the resulting genre classifications are very broad: zu Eissen and Stein (2004), for example, automatically extract texts that they classify as belonging to the "article" genre: this sounds relevant to EAP, but in fact this "genre" includes all longer passages of text on the web, be they research articles, reviews, technical reports, or book chapters. In contrast other classification schemes are extremely narrow: Temnikova *et al*. (2014), for example, describe a process of automatically identifying "sublanguages" which have their own distinct lexicogrammatical patterns, such as air traffic control talk or weather reports. Sublanguages are similar to genres as understood by ESP practitioners in that they arise recurrently in certain domains and are used by a community of specialists to discuss domain-related issues. However, many naturally-occurring texts which members of a discourse community would recognise as sharing a common purpose (and as therefore belonging to the same genre) are not sufficiently distinct in formal terms to be identified as genres by such automatic procedures.

Thus, for the time being at least, there is no easy way for ESP and EAP teachers and learners to find many examples of specific, narrow, genre categories in general corpora, or by using the web-as-corpus.

A central premise in applied linguistics is that "the meaning of a sentence is more than a combination of the meaning of the words it contains" (Channell 2009), and that for full understanding of a sentence or an utterance we need information about the situation in which it originally occurred. Sinclair's first criterion for corpus design (2005: 1) was therefore that corpus contents should be selected "according to their communicative function in the community in which they arise". As we have seen, this criterion is quite loosely interpreted in most large general corpora. It is true that research articles, and some types of newspaper article, have recognisable functions and can be associated with known discourse communities at a specific period in time, but for other types of corpus holdings the function and the circumstances surrounding text production may be difficult to determine. Very often, and especially in the case of web-derived corpora, "reality [...] does not travel with the text" (Widdowson 1998: 711-12). The original context of the text is irretrievable, having been lost in the process of corpus compilation, and so texts with very different communicative purposes come to be grouped into broad categories and treated as "part of one indistinguishable whole" (Mishan 2004: 220). This makes it difficult to investigate contextual variation in patterns of use and acceptability; if all types of newspaper text are lumped together, for example, the results of a query in a "newspaper" domain will not distinguish between editorials, financial reports, law reports, news reports and sports reports, despite the fact that

they are generically very different.

4. Needs analysis as a model for ESP/EAP corpus development

All these considerations point to the need for corpora which focus on the types of texts learners will engage with, annotated with information about their original contexts of use. Useful corpus categories for teaching and learning are those which shed light on communicative function and context rather than simply indicating a broad domain such as "Academic" (as in COCA) or "Business" (as in the Synchronic English Web Corpus and WordBanks Online).

Perhaps the best way to approach corpus design from an ESP perspective is to regard it as akin to the process of needs analysis, central to the practice of ESP. Halliday, McIntosh and Strevens (1964) first wrote of "English for special needs" rather than for special (or specific) purposes, but the needs they considered were solely linguistic and quantifiable, to be analysed in terms of register, much like the language technologists of today who count only the objectively observable features of texts. The notion of basing general language curricula on learners' communicative needs developed in the 1970s through the work of the Council of Europe (see, for example, van Ek's "Threshold Level" specification, 1975), and one of the first applications to ESP contexts was by Munby (1978). Munby's "communication needs processor" took into account both the learner's target situation (the settings, interaction between participants, and manner in which communication was carried out) and the texts produced in that situation (the communicative events, their purpose, and their medium, mode and channel). His detailed needs inventory was a "performance repertoire", as described by West (1994: 3); the activities and events to be identified were categories of real-world language use, but, unlike many later writers about needs analysis such as Hutchinson and Waters (1987), Munby was not really concerned with identifying learners' underlying competences or their requirements in terms of wants, lacks or learning styles. Course and syllabus designers criticise Munby's system for this reason, but it does not prevent *Communicative Syllabus Design* from being an excellent source of ideas for ESP/EAP corpus developers intent on collecting texts relevant to specific target situations.

Needs analysis is, of course, an art rather than a science, because of the choices that have to be made at the data gathering stage, when situations, participants and texts are selected, and the degree of interpretation required at the analysis stage, when meaning is attributed to the data. No two needs analysts are likely to arrive at the same list of syllabus items for the same group of learners, because each will have different priorities, ask different questions, and understand answers in at least slightly different ways. The same is true in corpus design, where no two developers will make the same selection of texts even if they are creating corpora for the same purposes, and every corpus developer will have different views about the kinds of contextual information to include, and the extent to which texts should be interpreted for the corpus user.

4.1. The process of ESP/EAP corpus development

As described by Munby (1978), ESP needs analysis involves analysing texts from the target situation together with the communicative situations in which they occur, and using the findings from this analysis to inform a teaching syllabus. There are choices at every stage in this process: texts have to be selected and prioritised, because it will probably not be possible to consider every type of text that occurs in the target situation, decisions have to be made about the type and quantity of participant information to include, because all kinds of personal information could potentially have some bearing on the way participants communicate, and methods of discovering the communicative roles of the texts have to be chosen, bearing in mind that the same text may be interpreted differently by different participants. The same kinds of choices have to be made when developing a corpus for use in ESP or EAP contexts, and in what follows I will discuss these decision-making processes with reference to the three academic corpora I have worked on: the British Academic Spoken English (BASE) corpus, the British Academic Written English (BAWE) corpus, and the Engineering Lecture Corpus (ELC).

4.2. The collection process

The plan for BASE, BAWE and ELC was to select the kinds of texts that university students regularly engage with, but of which there were few, if any, reliable examples already in the public domain, either on the web or in existing corpora. As an EAP practitioner I had taught academic listening and academic writing for many years, but my knowledge of lecture and seminar discourse was more or less limited to the lectures and seminars I myself had taken part in, my own experiences as a student (of English language and literature), and as an audience member for guest lectures and conference presentations, both of which are different from regular academic lectures in terms of their purposes and the speaker-audience relationships. Similarly my knowledge of student assignments was limited to the applied linguistics assignments I had set for my own students, and the assignments I had once written myself. My situation was, I think, little different from that of most teachers, students, and EAP textbook writers. Some genres in some disciplines might be accessible to some stakeholders sometimes, but even subject lecturers might not know much about the practices of their colleagues, and, contrary to good pedagogical practice, students are regularly required to produce genres they had never previously encountered. Even the texts and scripts in EAP textbooks often seem to reflect rather idealised notions of how students and university students communicate, giving advice on what the textbook writers think ought to happen, rather than what actually happens on degree courses in the disciplines. So the texts I wanted to include in the three corpora were "occluded" genres (Swales 1996), or in other words texts that are not easily accessible to researchers, teachers and learners by other means.

Munby's model requires the needs analyst to identify the purposes for which language is used in the target situation (the "purposive domain") and its medium, mode and channel ("instrumentality"). However, when working with occluded genres there are a number of constraints on what material can actually be collected,

and to what extent it is possible to control for contextual variables. Texts in a spoken academic corpus could be selected to represent different disciplines, degree programmes, stages of study, positions in a study sequence, class sizes and/or levels of interactivity, and might also try to be representative of the gender, age, experience and/or expertise of the participants. It would be very difficult to represent all these variables equally well, however, as a choice of one variable, for example discipline, might greatly limit the chance of finding lecturers of both genders, with greater and lesser degrees of experience, in small and large classes, with and without interactivity, for example. Yet these considerations are important, because if corpus holdings are going to be compared in terms of a variable such as discipline, the groups of texts in each discipline need to be broadly similar in terms of the other variables.

Spoken text collection also depends, of course, on participants' willingness to be recorded; attempts to record an entire sequence of lectures might be thwarted if the sequence is taught by several lecturers, and not all agree to take part in the project. In the end the collection plans for BASE and ELC were fairly simple. BASE was designed to represent lectures and seminars in equal quantities across four broad disciplinary domains: Arts and Humanities, Life Sciences, Physical Sciences and Social Sciences. Other variables were not controlled, although in hindsight the design could have been improved by selecting an equal number of lectures and seminars from each year of study on undergraduate and postgraduate courses. ELC has been designed to enable comparison of lecturing styles across different countries where English is used as a medium of instruction. For this reason only the lecture genre has been chosen; a decision was made to focus on only one discipline to prevent the project from becoming too complex. Roughly equal numbers of engineering lectures are being collected from different institutions – so far from universities in the UK, New Zealand and Malaysia. The corpus is expanding, so it may be possible to find very good matches across a large number of countries in terms of lecture topic and level. Cultural similarities and differences are already emerging (see, for example, Alsop *et al*. 2013, and Alsop & Nesi 2015).

In the BASE corpus the "seminar" turned out to be a problematic category. As in needs analysis, the process of text collection is often a journey of discovery: before entering the target situation not only were we unsure of the number of texts available for collection, but we were also unsure of the exact nature of these texts. There is no seminar category in the Michigan Corpus of Academic Spoken English (MICASE), which was developed over roughly the same time period as BASE. Only large and small lectures are represented, the small lectures being more interactive than the large ones. Within British university degree courses, however, lectures and seminars are usually treated as distinct events; lectures are largely monologic, while in seminars it is the students rather than the lecturers who do most of the talking. The way seminars were conducted varied widely across the BASE corpus collection contexts, however; sometimes large classes were divided into small discussion groups, sometimes the entire class discussed together, and sometimes students took turns to deliver presentations, informally around a table, or more formally at the front of the class with slides. It was difficult to plan in advance to collect a

representative sample of each of these formats, because we did not always know in advance of data collection what the format would be.

Similarly, we did not know what genre types we would discover during the process of collecting texts for the BAWE corpus. BAWE was developed with the intention of discovering the range of written genres students had to produce for degree programmes in Britain, but we could not plan in advance to collect a representative sample of each. Instead we designed the corpus to reflect the variables we were sure of from the start: we collected roughly equal quantities of assignments in each discipline and at each level of study (first year undergraduate to taught Masters level – labelled as Levels 1 to 4). It was only after collection had ended that we were able to map and describe the way genres were distributed across these levels and disciplines (see Nesi & Gardner 2012). This distribution might be taken to reflect the relative importance of each genre in each discipline, but any assumption of this kind must be very tentative, because we did not sample in a stratified way from the entire population of texts. It should also be noted that dissertations and theses were not included in the BAWE mix, on the simple grounds that these are not really "occluded" genres.

Strict statistical sampling methods proved unviable for all three corpora, and may be unviable for other EAP corpus developers too, if it is not possible to calculate how many lectures, seminars, or assignments are produced at any given time in the communicative context, just as it is impossible to calculate how many texts in any particular category exist on the web. Without knowing the total population size it is not possible to collect a proportional sample of the population, and with BASE, BAWE and ELC we have simply aimed to collect the quantity of texts we could afford to process, balanced across a certain number of categories (genre in the case of BASE, discipline and level in the case of BAWE, country of origin in the case of ELC). To represent additional categories in a balanced way would be far more costly, as it would entail collecting many more samples than were needed, and discarding those that were over-represented in any subcategory.

Moreover, although we started our collection procedures with a finite set of "disciplines", taught in the universities where we gathered data, assigning texts to disciplinary categories is not entirely straightforward. Boundaries between related fields of study are permeable, and within discipline-specific programmes there are often outlying modules, for example on the history of mathematics in a mathematics programme, or on business law for a degree in business. For the BASE corpus we decided to place speech events within domains according to their content rather than the organising department. Thus an ecology lecture delivered in a mathematics department was placed in the domain of Life Sciences, a linguistics lecture was placed in Physical Sciences because of its technical nature, and Philosophy and Typography speech events were placed in more than one domain. A less interpretative procedure was followed for the BAWE corpus, on the grounds that we intended to capture the student experience of genre production in specific disciplines. Texts were assigned to domains according to the department where the student was enrolled, and departments at different institutions were merged, ignoring slight variations in their names, if they offered degrees in the

same disciplinary areas.

The final contents of BASE and BAWE are outlined in Tables 1 and 2. Apart from the loss of one seminar in the Physical Sciences, we managed to collect as much BASE data as we had planned, spread across the four domains. The design of BAWE was more ambitious, and while we had originally aimed to collect 32 assignments from each year of study in each of the major disciplines, the final collection was slightly less evenly spread. We also found that we had to distinguish between assignments (the complete pieces of work submitted for assessment) and texts (the generically distinct texts included within assignments); figures for both are provided in Table 2. More details of the process of collecting texts for the BAWE corpus can be found in Alsop and Nesi (2009).

Table 1: Overview of BASE Corpus Holdings

	Lectures	**Seminars**	**Total**
Arts and Humanities Caribbean Studies; Classics; East Asian Studies; English; Comparative American Studies; Comparative Literary Studies; Comparative Cultural Studies; Film and Television Studies; French; History; History of Art; Philosophy; Theatre Studies; Typography	40	10	50
Life Sciences Agricultural Botany; Animal and Microbial Sciences; Biological Sciences; Ecology (in Mathematics); Medicine; Plant Sciences; Zoology	40	10	50
Physical Sciences Chemistry; Computer Science; Cybernetics; Economics; Engineering; Linguistic Science; Mathematics; Meteorology; Philosophy; Physics; Psychology; Statistics	40	9	49
Social Sciences Applied Linguistics; Business; Globalisation and Regionalisation; Economics; Education; Law; Linguistics; Internationalisation; Japanese Studies; Management; Politics; Psychology; Social Work; Sociology; Study of Women and Gender; Typography	40	10	50
Total	**160**	**39**	**199**

Table 2: Overview of BAWE Corpus Holdings

		Level 1	**Level 2**	**Level 3**	**Level 4**	**Total**
Arts and Humanities Archaeology; Classics; Comparative American Studies; English; History; Linguistics / English Language Studies; Philosophy; others	Assignments	239	228	160	78	**705**
	Texts	255	229	160	80	**724**
Life Sciences Agriculture; Biological Science; Food Science; Health; Medicine; Psychology	Assignments	180	193	113	197	**683**
	Texts	188	206	120	205	**719**
Physical Sciences Architecture; Chemistry; Computer Science; Cybernetics/Electronic Engineering; Engineering; Mathematics; Meteorology; Physics; Planning	Assignments	181	149	156	110	**596**
	Texts	181	154	156	133	**624**
Social Sciences Anthropology; Business; Economics; Hospitality, Leisure and Tourism; Management; Law; Politics; Publishing; Sociology	Assignments	207	197	162	202	***777**
	Texts	216	198	166	202	***791**
Total assignments		**807**	**767**	**591**	**587**	***2761**
Total texts		**840**	**787**	**602**	**620**	***2858**

*Including 9 of unknown level

4.3. Recording participant information

Munby's model requires the needs analyst to identify participants and describe any characteristics that might be relevant to communication in the target situation, for example age, gender and nationality. Some of this participant information might be confidential, however, and even if participants are willing to provide a great deal of personal information about themselves there may not be time to collect and record it all, or to find useful ways of categorising the information for future reference.

For BASE, only participant information about the lecturers was included. This was fairly easy to collect, as all the lecturers had public profiles. Students would have found it inconvenient to provide us with personal information on an individual basis, as data collection would have had to take place outside class time, so we contented ourselves with broad general information about their level of study and the size of the class. This information was included in the header for each file, together with the recording date, discipline, module name and the title of the

speech event.

The 1,039 student contributors to BAWE received payment for each assignment they submitted, but had to fill in a participant form giving their age, gender, department and degree course, to be added to the header information for each file together with the original submission date of the assignment and the module name.

In BASE, BAWE and ELC we did not differentiate between "native" and "non-native" speakers, and operated on the assumption that all the participants were communicatively effective users of English in their own contexts, as lecturers, seminar discussants, or writers (only assignments with above average marks were collected for the BAWE corpus). Nevertheless a certain amount of language information was collected for the three corpora. For BASE and ELC, lecturer first language has been noted (just a few lecturers were not L1 speakers of English). BAWE contributors were asked to state their first language and also the number of years of secondary education they had received in the UK. This enabled us to differentiate between L1 English speakers educated in the UK and those educated elsewhere, and between L2 English speakers educated in other countries and those educated in the UK. This language and education information can help us to identify distinctive patterns of use which are restricted to contributors with a particular first language. We have received a number of requests for information about the first languages of speakers in the BASE seminars, from researchers who would like to investigate their speech in a similar way. Unfortunately we did not keep a record of the first languages of the BASE seminar participants.

We are wary, however, of encouraging cross-corpus generalisations about "native" and "non-native" use; BASE, BAWE and ELC are not designed to facilitate comparisons between speakers of different first languages. This is partly because participants' stated first languages are not necessarily the ones in which they are most proficient for academic purposes, and partly because participants with different first languages are not evenly spread across domains and levels. In BASE and BAWE there are many more international students at masters level and in applied disciplines such as business, for example, than there are across the corpora as a whole. ELC is designed to facilitate comparison between lecturing styles in different parts of the world, but we focus on the cultural context rather than the L1 of the participants; it appears that engineering lectures in New Zealand are somewhat different from engineering lectures in the UK, for example, even though all the lecturers and almost all the students share the same L1.

It would of course have been possible to focus on many other aspects of participant identity while collecting corpus data. The possibilities are endless; it has been suggested that we might have included information about the students' ethnicity, region or socio-economic status, for example, and whether or not they had been privately educated, or were from rural or urban backgrounds. However, different corpora, with different basic designs, are needed to explore the implications of these different types of information.

4.4. Interpreting communicative roles

BASE, BAWE and ELC could not have been created without access to course documentation and without the support of subject specialists and students. Documentation was needed at the design stage to identify target departments and modules, and we then called upon module leaders to permit us to record their lectures and seminars for BASE, and to help us make contact with students who might contribute their assignments to BAWE. Sources were also useful as an aid to categorising the texts we collected, and interpreting their meaning.

The development of the BASE corpus went hand-in-hand with the development of three CD-Roms for the self-access study of academic speaking and listening (Kelly *et al.* 2000, 2004, 2006, now available online from the University of Warwick). These materials were based around hundreds of video clips of lectures and seminars recorded for BASE, but also included excerpts from video interviews with subject lecturers and students, which in turn informed our analysis of the corpus. In particular, as reported by Nesi (2001), the interviews revealed how departments perceived and differentiated the roles of lectures and seminars.

We sought the same sort of guidance to categorise genres for the BAWE corpus, but this was less successful, as many lecturers and students seemed to have greater difficulty distinguishing between assignment types. When they submitted their work to the corpus, students were asked to choose the most appropriate descriptor for each assignment from a choice of *case-study*, *essay*, *exercise*, *notes*, *presentation*, *report*, *review*, and *specified other* (see Alsop & Nesi 2009). We could not find much agreement, however, between different contributors submitting the same sort of assignment, in the same discipline, or between what contributors chose as the descriptor for their assignment and their own references to their work within the assignments themselves. For example, assignments identified as *essays* sometimes began with the words "In this report", or vice versa. One or two genres such as the problem question in law were well established and clearly defined by all participants, but for the most part the nomenclature did not seem to exist within departments to enable lecturers and students to distinguish between all the different types of assignments students were required to write.

It therefore largely fell to the corpus developers to identify the communicative purposes (or genres) of the BAWE texts. The identification process was informed by the objectively observable structural and linguistic features of the texts, and by outside sources–the course documentation and advice from participant interviewees, as discussed in Nesi and Gardner (2006), but it was essentially interpretative, and ultimately imposed on the data our own understanding of why the assignments had been written, and what skills and knowledge they intended to demonstrate. Each text in the corpus is marked as belonging to one of the 13 "genre families" described in detail in Gardner and Nesi (2012). This classification system is also the organising principle for our academic writing materials on the British Council website, where practice activities are supplemented by links to corpus evidence and interviews with teaching staff and students (see Nesi & Gardner 2015).

"Pragmatic" markup, concerned with textual meaning, has been taken to greater lengths in ELC. The corpus to date has been annotated manually to indicate

recurring phases in lecture discourse where lecturers tell stories, inject humorous elements, or summarise previous or future lecture content (Alsop & Nesi 2014; Alsop *et al*. 2013). Pragmatic markup goes beyond identifying "philologically observable objective characteristics" (Karlgren 2012) to consider what texts mean from a human perspective. There have been experiments with the automatic identification of speech acts in restricted text types such as business emails (see, for example, De Felice & Deane 2012) but few, if any, other corpora have been annotated manually to identify the functions of longer stretches of text in less formulaic genres.

There are arguments against this kind of markup, in that it imposes on the corpus interpretations that other users may not share. A final, incontestable decision about the communicative effect of an utterance, or string of utterances, may be impossible to achieve, as texts have as many interpretations as they have readers or listeners, and "the same utterance will mean something different to each person who hears it" as Sinclair (2004:157) points out. On the other hand, by annotating our corpora in this way, we have prepared the ground for future users so that they can, if they wish, narrow their investigations to specific genres in BAWE or specific lecture phases in ELC, without needing to repeat our initial laborious identification stages. Markup greatly facilitates the retrieval of texts and parts of texts, and makes it possible to reveal new distribution patterns across the entire corpus, for example by using the bespoke visualisation tool described by Alsop and Nesi (2014). Alternatively users can, of course, just ignore the pragmatic information, and focus solely on the observable objective characteristics of the texts.

5. Conclusion

This paper has made a case against the exclusive use of large, general corpora in ESP/EAP teaching and learning, and in favour of smaller more specific corpora which contain the types of texts that learner will need to engage with, but which they may have difficulty finding on the web or in general corpora. It has drawn an analogy between ESP needs analysis and ESP corpus development, in that both processes require the ESP/EAP practitioner to collect and subsequently analyse data relating to communication in the target situation, taking into consideration the circumstances under which such communication takes place, and details of communicative purpose and participant identity that affect the way texts are constructed. It has also discussed some of the issues surrounding the collection of contextual information and its inclusion in corpus resources.

Small(ish) corpora are being created all the time by ESP and EAP practitioners around the world, mostly for private use in class, and to inform local course and syllabus design. I believe that it is worth enhancing these resources, where possible, with the kind of information that the needs analyst arrives at through investigation of the context and interpretation of the textual evidence. In this way, we can provide our students with a genuinely useful supplement to the giants of the corpus world.

Bibliographical references

Corpora and corpus query tools referred to in this paper

AntConc <http://www.laurenceanthony.net/software/antconc>

British Academic Spoken English (BASE) <www.coventry.ac.uk/base>

British Academic Written English (BAWE) <www.coventry.ac.uk/bawe>

British National Corpus (BNC) <http://www.natcorp.ox.ac.uk/>

Corpus of Contemporary American English (COCA) <http://corpus.byu.edu/coca>

Engineering Lecture Corpus (ELC) <www.coventry.ac.uk/elc>

enTenTen13
<https://www.sketchengine.co.uk/documentation/wiki/Corpora/TenTen/enTenTen>

Google Advanced Search <http://www.google.com/advanced_search>

Michigan Corpus of Academic Spoken English (MICASE) <http://quod.lib.umich.edu/m/micase>

SketchEngine <http://www.sketchengine.co.uk>

Synchronic English Web Corpus <http://wse1.webcorp.org.uk/home/syn.html>

WordBanks Online (formerly known as the Bank of English) <http://www.collins.co.uk/page/Wordbanks+Online>

References

Alsop, Siân, Emma Moreton & Hilary Nesi. 2013. "The uses of storytelling in university engineering lectures". *ESP across Cultures* 10, 8–19.

Alsop, Siân & Hilary Nesi. 2009. "Issues in the development of the British Academic Written English (BAWE) corpus". *Corpora* 4/1, 71–83.

Alsop, Siân & Hilary Nesi. 2014. "The pragmatic annotation of a corpus of academic lectures". In Calzolari, N. K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.), *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation*. May 26-31 2014, Reykjavik, Iceland, 1560–1563.

Alsop, Siân & Hilary Nesi. 2015. "Introductions in engineering lectures". *Proceedings of the 8th International Corpus Linguistics conference (CL2015), Lancaster University, 21-24 July 2015*. 19–22. Abstract retrieved from <http://ucrel.lancs.ac.uk/cl2015> on 1 August 2015.

Boulton, Alex. 2015. "Applying data-driven learning to the web". In Leńko-Szymańska, A. & A. Boulton, *Multiple Affordances of Language Corpora for Data-driven Learning*. Amsterdam: John Benjamins, 267–295.

Channell, Joanna. 2009. "Language awareness: Pragmatics". *MED Magazine*, Issue 54. Retrieved from <http://www.macmillandictionaries.com/MED-Magazine/August2009/54-LA-Pragmatics.htm> on 1 August 2015.

De Felice, Rachele & Paul Deane. 2012. *Identifying Speech Acts in E-mails: Toward automated scoring of the TOEIC(R) e-mail task* (ETS Research Report No. RR-12-16). Princeton, NJ: ETS.

Gardner, Sheena & Hilary Nesi. 2012. "A classification of genre families in university student writing". *Applied Linguistics* 34/1, 1–29.

Gatto, Maristella. 2014. *Web as Corpus: Theory and Practice*. London: Bloomsbury Academic.

Halliday, Michael, Angus McIntosh & Peter Strevens. 1964. *The Linguistic Sciences and Language Teaching*. London: Longman.

Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý & Vit Suchomel. 2013. "The TenTen Corpus Family". Paper presented at the *7th International Corpus Linguistics*

Conference, Lancaster, July 2013. Abstract retrieved from <http://ucrel.lancs.ac.uk/cl2013> on 1 May 2015.

HUTCHINSON, Tom & Alan WATERS. 1987. *English for Specific Purposes*. Cambridge: Cambridge University Press.

KARLGREN, Jussi. 2012. "Genres on the web – what is a genre anyway?". IR Seminar, Centre for Language Technology, University of Gothenburg. Abstract retrieved from <http://clt.gu.se/node/2981> on 1 May 2015.

KELLY, Tim, Rod REVELL & Hilary NESI. 2000. *Listening to Lectures*. Centre for English Language Teacher Education, University of Warwick.

KELLY, Tim, Lynette RICHARDS & Hilary NESI. 2004. *Seminar Skills 1: Presentations*. Centre for English Language Teacher Education, University of Warwick.

KELLY, Tim, Gerard SHARPLING & Hilary NESI. 2006. *Seminar Skills 2: Discussions*. Centre for English Language Teacher Education, University of Warwick.

KILGARRIFF, Adam & Vit SUCHOMEL. 2013. "Web Spam". *Proceedings of the 8th Web as Corpus Workshop (WAC-8)*, United Kingdom, July 2013, 46–52.

LEE, David. 2001. "Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle". *Language Learning & Technology* 5/3, 37–72.

MISHAN, Freda. 2004. "Authenticating corpora for language learning: a problem and its resolution" *ELT Journal* 58/3, 219–227.

MUNBY, John. 1978. *Communicative Syllabus Design*. Cambridge: Cambridge University Press.

NESI, Hilary. 2001. "EASE: a multimedia materials development project". In CAMERON, K. (ed.), *C.A.L.L. – The Challenge of Change*. Exeter: Elm Bank Publications, 287–292.

NESI, Hilary. 2012. "Alternative e-dictionaries: uncovering dark practices". In GRANGER, S. & M. PAQUOT (eds.), *Electronic Lexicography*. Oxford: Oxford University Press, 357–372.

NESI, Hilary & Sheena GARDNER. 2012. *Genres across the Disciplines: Student writing in higher education*. Cambridge: Cambridge University Press.

NESI, Hilary & Sheena GARDNER. 2015. "Balancing old and new activity types on an academic writing website". In KAVANAGH, M. & L. ROBINSON (eds.), *The Janus Moment in EAP: Revisiting the Past and Building the Future*. Reading, UK: Garnet Education, 187–198.

SHA, Guoquan. 2010. "Using Google as a super corpus to drive written language learning: a comparison with the British National Corpus". *Computer Assisted Language Learning* 23/5, 377–393.

SINCLAIR, John. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

SINCLAIR, John. 2005. "Corpus and Text – Basic Principles". In WYNNE, M. (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, 1–16. Retrieved from <http://ota.ox.ac.uk/documents/creating/dlc/> on 1 August 2015.

SWALES, John. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

SWALES, John. 1996. "Occluded genres in the academy: the case of the submission letter". In VENTOLA, E. & A. MAURANEN (eds.), *Academic writing: Intercultural and Textual Issues*. Amsterdam: John Benjamins, 45–58.

TEMNIKOVA, Irina, William BAUMGARTNER, Negacy HAILU, Ivelina NIKOLOVA, Tony MCENERY, Adam KILGARRIFF, Galia ANGELOVA & K. BRETONNEL COHEN. 2014. "Sublanguage corpus analysis toolkit: A tool for assessing the representativeness and sublanguage characteristics of corpora". In CALZOLARI, N. K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. Moreno, J. ODIJK & S. PIPERIDIS (eds.), *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation*. May 26-31 2014, Reykjavik, Iceland, 1714–1718.

VAN EK, Jan. 1975. *The Threshold Level*. Strasbourg: Council of Europe.

WATSON TODD, Richard. 2001. "Induction from self-selected concordances and self-correction". *System* 29/1, 91–102.

WEST, Richard. 1994. "Needs analysis in language teaching". *Language Teaching* 27/1, 1–19.

WIDDOWSON, Henry G. 1998. "Context, community and authentic language". *TESOL Quarterly* 32/4, 705–716.

ZU EISSEN, Sven MEYER & Benno STEIN. 2004. "Genre classification of web pages: user study and feasibility analysis". In BIUNDO, S., T. FRÜHWIRTH & G. PALM (eds.), *KI 2004: Advances in Artificial Intelligence. Lecture Notes in Computer Science* 3238, 256–269.

Hilary Nesi is Professor in English Language at Coventry University. Her research activities largely concern corpus development and analysis, the discourse of English for academic purposes, and the design and use of dictionaries and reference tools for academic contexts. She was one of the developers of the business correspondence component of the JISC-funded BT e-Archive, and was principal investigator for the project to create the BASE corpus of British Academic Spoken English (2001–2005), and for the project to create the BAWE corpus: 'An Investigation of Genres of Assessed Writing in British Higher Education' (2004–2007). She is currently leading the development of the Engineering Lecture Corpus (ELC). She was Chief Academic Advisor for the EASE series of multimedia EAP speaking and listening materials, and has recently led an ESRC-funded project to produce online academic writing materials for the British Council 'Learn English' website. She is the co-author of *Genres across the Disciplines: Student writing in higher education* (Cambridge University Press, 2012).

h.nesi@coventry.ac.uk