Finding Textual Examples of Genres: 
Issues for Corpus Users

Chae Kwan Jung • Sue Wharton
(Inha University) (University of Warwick)

Jung, Chae Kwan & Sue Wharton. 2012. Finding Textual Examples of Genres: Issues for Corpus Users. Korean Journal of English Language and Linguistics 12-1, 129-148. There is a range of EAP corpora available to teachers and researchers today. Some of these offer classifications of their texts by genre, in order to allow corpus users who are interested in genre-specific writing to select appropriate sets of texts. However, the classification used in the corpus may not easily identify the genres which the corpus user wishes to research, making it necessary for corpus users to develop their own procedures for identifying suitable texts. In this paper, we discuss our own experience of locating texts from two genres, Reflective Accounts and Undergraduate Engineering Laboratory Reports, from the BAWE corpus. We evaluate the success of the procedures we developed, and suggest implications for other corpus users and for corpus developers.

Key Words: EAP Corpora, Genre-Specific Writing, Corpus Users and Developers

1. Introduction

An increasing number of corpora based on academic texts are becoming available to teachers and researchers of EAP, often developed by groups of academics in higher education institutions. Some corpora, such as BAWE (British Academic Written English) or MICUSP (Michigan Corpus of Upper-level Student Papers), offer classifications by genre of the texts they contain. However, the classification systems offered do not necessarily fit with the research questions or data search needs
of the corpus users; it can therefore be necessary for corpus users to develop their own procedures to identify appropriate sets of texts (e.g. reflective accounts and laboratory reports). In this discussion paper we discuss our own experience of developing such systems, and suggest implications for other corpus users and for corpus developers.

2. Corpora for EAP

An increasing number of researchers and practitioners discuss the potential usefulness of corpora to support English for Academic Purposes research and teaching (Cheng et al. 2003, Flowerdew, 2002; Hyland, 2006; Krishnamurthy and Kosem, 2007; Lee and Swales, 2006; Nesi, 2008). EAP corpora are becoming increasingly well known and available, though they differ markedly in terms of their size, the way they are organised, and the types of texts which they include. The survey from which the list is derived is discussed in Xiao (2008). Nesi (2008) offers an overview of types of corpus used in EAP (e.g. corpora of expert writing, learner corpora, corpora of university student writing, and spoken academic corpora), and discusses the advantages and disadvantages of each type.

Some examples of corpora containing a range of academic text types are Reading Academic Text Corpus (RAT) which consists of research articles and theses written by staff and students from Reading University; or The Academic Corpus at the Victoria University of Wellington which includes 414 academic texts (3.5 million words) including journal articles, book chapters, course workbooks, laboratory manuals, and course notes. The TOEFL

---

1An annotated list of available corpora can be found at www.lancs.ac.uk/fass/projects/corpus/cbls/corpora.asp.

2http://www.victoria.ac.nz/lals/resources/academicwordlist
2000 SWAL (Spoken and Written Academic Language) corpus, arising from a research project sponsored by Educational Testing Services, aims to represent both spoken and written speech events and texts which students will encounter at universities in the USA and include textbooks, course handouts, and university web pages (Biber et al, 2004).

Krishnamurthy and Kosem (2007) and Nesi (2008) point out that the number of EAP corpora which draw specifically on student writing is currently increasing, often initiated by students themselves. Some examples are LANCAWE (Lancaster Corpus of Academic Written English)\(^3\). LANCAWE is a collection of texts written by international students, non-native speakers of English, at the University of Lancaster; the Viking Corpus of Student Academic Writing, Portland State University USA\(^4\). This is an ongoing student initiated project, consisting of papers written by PSU students achieving at least a B grade.

Perhaps the most accessible corpus of student writing is the MICUSP (Michigan Corpus of Upper Level Student Papers). This is a 2.6 million word corpus consisting of 829 A-grade student papers, written by final year undergraduates or year 1-3 postgraduate students. It is designed to be used by researchers, teachers, and students directly (Romer and Swales, 2010). The corpus is directly searchable via a web interface called MICUSP simple (Beta)\(^5\).

A particular comprehensive corpus of university level student writing is the BAWE (British Academic Written English) corpus. BAWE is an approximately 6.5 million word corpus of proficient student assignments, produced and assessed at the universities of Coventry, Oxford Brookes, Reading and Warwick in the UK (Alsop and Nesi, 2009: 73). It was created between 2004-2007.

\(^3\)http://www.ling.lancs.ac.uk/activities/294
\(^4\)http://web.pdx.edu/~conrads/online_corpus.html
\(^5\)http://search-micusp.elicorpora.info
through a project entitled 'An investigation of genres of assessed writing in British Higher Education.' The BAWE corpus is available to researchers through the Oxford Text Archive6 as resource number 2539. When downloaded, this resource includes sets of corpus files in ASCII, UTF-8 and TXT format, as well as a PDF file entitled The BAWE Corpus Manual and an Excel file providing detailed information about each text included.

The corpus itself contains 2761 pieces of proficient assessed student assignments from a range of disciplines and courses. Assignments are categorized on a number of dimensions - such as course, level of study, language background of writer, and genre family - in order to allow researchers to group and select sets of texts with characteristics in common. This categorization information is encoded in the document headers of UTF files and is separately recorded in the Excel file 'BAWE.xls,' a data spreadsheet which lists each assignment and its information.

PDF files of the original assignments, with contributor information anonymized, are held by members of the project team and by contributing universities7. These PDF files have an additional use in that they enable researchers to see, for example, the physical layout of an assignment or an actual drawing within it. In other corpus file formats, such features are represented by a corpus 'tag' e.g. <figure>.

3. Genre categorization of texts within EAP corpora

There is a long tradition of research into EAP genres, including research into writing tasks set for university students. A range of different approaches have been used, including questionnaires and interviews with academic staff, and studies of course

---

6http://ota.ahds.ac.uk
7http://www2.warwick.ac.uk/fac/soc/al/research/collection/bawe
documentation or assignment prompts (Braine, 1989, Casanave and Hubbard, 1992; Cooper and Bikowski, 2007; Jackson et al., 2006). Genre based research can also be carried out using EAP corpora, but corpora vary a great deal regarding their approach to classifying the text types or genres which they include. This, of course, is because different corpora are compiled with different types of exploitation in mind.

For example, the Academic Corpus Victoria University of Wellington does not include genre classification at all – this was not its purpose. Rather, it was designed principally to allow the creation of an academic word list. A list has been developed containing 570 word families common in academic registers (Coxhead, 2000); this list is available via sidebar navigation on the corpus homepage.

In contrast, the MICUSP Simple (Beta) interface lists a range of paper categories: Argumentative Essay, Creative Writing, Critique/Evaluation, Proposal, Report, Research Paper, Response Paper. All of these paper categories appear in a range of disciplines. A corpus information page gives more details of the definition of each paper category, including rhetorical purpose, typical features, and examples of more specific paper types within each category (Romer and O’Donnell, 2011).

The Bawe project laid considerable emphasis on the issues of genre. A key aim of the Bawe corpus, from which our own examples are drawn, was to categorize assignments into genre families, so that they could be compared on this dimension across disciplines and years of study (Alsop and Nesi, 2009; Gardner 2011). Bawe has indeed formed the basis for published genre-specific research, e.g., Essays in Sociology and in English (Bruce, 2010); Essays in Social Sciences (Durrant and

\[\text{http://www.victoria.ac.nz/lals/resources/academicwordlist}\]

\[\text{http://micusp.elicorpoa.info/researchers/micusp-paper-classification/definitions-for-micusp-paper-classificat}\]
Mathews-Aydinli, 2011); and Reflective Accounts (Wickens, 2011; Nesi 2011). For a comprehensive list of research using BAWE, see Publications on Coventry University web site10.

In the BAWE corpus manual, 13 genre families (e.g. essays, critiques, case studies, and so on) are identified (see also Gardner, 2011; Nesi, 2011). The description of each genre family includes social purpose, components and genre networks; it also gives examples of genres within each genre family. Gardner (2011) argues that this classification is more manageable than a much longer list of genres (the BAWE corpus manual lists 96 genres).

4. Focus of paper: locating textual examples of genres within a corpus

In the light of the issue discussed above, we feel there is more to say about using a corpus such as BAWE for research on specific genres. These issues are not widely acknowledged in the literature; clearly, there is little space in an analysis based research article for an author to discuss processes of text selection in detail. Some writers, whose research is based on selections of texts from BAWE, allude to the complexities involved; for example Durrant and Mathews-Aydinli (2011) state that: "In particular, we looked at essays produced by students in social science MA courses... However, two texts (from the 'Drama and Theatre' discipline group) were eliminated from our investigation because we were not satisfied with their classification as 'social science' assignments (p. 62). Complexities are also apparent from comparisons of work by different researchers. For example Wickens (2011) and Nesi (2011) all work on a subcorpus of reflective writing within BAWE, but

10See http://wwwm.coventry.ac.uk/researchnet/BAWE/research/Pages/Publications.aspx
each selects a slightly different range of texts.

We argue, then, that criteria and processes of selection may be rather complex and deserving of attention in a discussion article. To attempt to illustrate the issues and prompt discussion, we focus on our attempts to locate texts in two specific genres within the BAWE corpus. In each case, we discuss the extent to which we were able to rely on BAWE’s own genre categorization information to locate target texts, and the extent to which we needed to develop our own procedures to identify them. We evaluate the success of those procedures, and discuss implications for future corpus design and corpus use.

5. First example: selection of reflective accounts

Reflective assignments are becoming increasingly common in UK higher education contexts, and BAWE is an obvious resource in which to locate textual examples for research. However, it is not possible to simply select those BAWE texts which were labelled as reflective accounts, because the term ‘reflective account’ is not one of the genre family labels used within BAWE documentation for categorizing texts. We therefore began with our own working definition of a reflective account, and then examined the BAWE documentation of genre families (BAWE Corpus Manual, Appendix 2, pp. 45-50) as a starting point to discover how such accounts were most likely to have been categorized.

Our definition of assessed reflective accounts is “writing which is produced for formal evaluation in an educational course and which requires the writer to: narrate personal experience; comment on associated feelings; appraise their performance; discuss what they have learned; and relate the learning to some aspect of future action.”
The most likely genre family in which to locate reflective accounts is that of 'narrative recount' whose social purpose is described as "to develop awareness of motives and/or behaviour in individuals (including self) or organizations" and which has "reflective account" as one of its sub-genre examples (BAWE Corpus Manual, Appendix 2). Reflective accounts do not appear as an example of any other genre family. For these reasons, we began by identifying those texts which are labelled as narrative recounts in BAWE, 61 assignments in all.

To ascertain which of these assignments should be classified as reflective accounts, we read each of them in the light of the working definition above, and rejected any which did not fit. For example, BAWE Text ID 0317a is entitled 'History of Maths Essay 1.' In it, the writer reviews and reflects upon the contribution of a nineteenth century mathematician to the development of Calculus. BAWE Text ID 6002a is entitled 'Famous Microbiologists: Alexander Fleming.' It narrates the events which led Fleming to his discovery of penicillin. Both of these texts were narrative recounts, but neither was a reflective piece on our definition; they were therefore not included in our subcorpus.

After considering all 61 narrative recounts in this way, we were left with a set of 47 texts. The procedures followed thus far reduced the risk of including inappropriate texts, but still left the possibility of inappropriate exclusion. BAWE could contain reflective accounts which had not been labelled as narrative recounts, either through human error or because they possessed characteristics which had been judged by the corpus developers as associating more closely with a different genre family. Rather than read all the remaining texts in BAWE to try to identify such cases, we went to the BAWE.xls spreadsheet and read the contextual information, in an attempt to identify and combinations of assignment titles, courses, and so on which
might suggest a reflective account. We then checked any potential texts. This procedure led us to add 6 further texts to our sub-corpus, meaning that the final BAWE reflective writing subcorpus consisted of 53 texts.

Reflective accounts, then, were relatively easy to find by using the information systems provided in BAWE documentation. Our working definition did not require us to specify work from any particular discipline, year of study, context of production, etc. And yet despite this relative simplicity, it was not possible for us to choose texts only on the basis of the contextual information provided in the corpus. Text selection was still dependent to some degree on our own qualitative judgements as researchers.

We will now go on to discuss the procedures followed to identify text examples of a second genre, the Undergraduate Engineering Laboratory Report. This genre label involves more specific meanings in terms of level of study, academic discipline, event leading to text production, and resulting text. For this reason, we need to develop a more complex set of procedures to identify texts in this ‘occluded’ (Swales, 1996) genre.

6. Second example: selection of Undergraduate Engineering Laboratory Reports (UELRS)

The BAWE corpus is intended to be well balanced across both disciplinary areas and assignment types (Alsop and Nesi, 2009), and is therefore a good source to locate science and technology writing which differs from the traditional academic essay. Again, the genre label which we ourselves wished to use, ‘Undergraduate Engineering Laboratory Report’, is not used in BAWE for categorizing texts, and so we started from our own working definition. In this section, we will explain various aspects of our definition and the issues that each aspect brought
for selecting appropriate texts from BAWE.

The term, Undergraduate, in the UK higher education context, conventionally refers to students in their first three years of higher education and working towards a Bachelor’s degree. BAWE xls gives information about the degree course in which an assignment was produced (field AB) and our first instinct was to select assignments produced on Bachelor’s courses. However, examination of the spreadsheet suggested that the work of students following four year courses had in some cases been categorized as belonging to a Masters course, even when the assignments had been produced in the first three years of study. So for example, seven assignments contributed by student number 0347, produced in years 1 and 2 of university study, were labelled as belonging to a Masters degree course. In the light of this we decided that for our purposes, course title – Bachelors or Masters – was not a suitable criterion to identify an assignment as having been produced at undergraduate level. We therefore based our selection on year of study only, choosing assignments produced in years 1 to 3.

The term, Engineering, relates to a specific academic discipline, which may cover an entire degree program syllabus, or may be taught in an interdisciplinary combination. For example the Warwick School of Engineering works with Warwick Business School to offer a degree in Business and Engineering. BAWE xls gives information about the module and the discipline in which a given assignment was produced. Some of this information was self reported by student contributors during the assignment submission process (Alsop and Nesi, 2009) and it seemed that different students had followed different practices. For example, some students assigned a discipline based on the overall degree course they were following (as was intended by the corpus developers), whereas others assigned it on the basis of a specific module in which the assignment was produced. This potential
confusion meant that it was impractical for us to identify ‘Engineering’ assignments written by students not following an Engineering degree. We therefore decided to limit our selection to assignments produced on BEng or MEng degrees.

The term, Laboratory, is arguably the most critical term which distinguish between the UELR and other types of report which may appear in BAWE. For us the term laboratory includes, but is not limited to, a place or facility equipped for scientific experiment such as a laboratory room. It refers to practice-based or practical work using technical or engineering tools and techniques e.g. machinery, robot arms or computer software packages. Therefore, ‘laboratory’ sessions are primarily practice-based sessions. There is no field in BAWE xls which provides information about whether texts were produced in connection with practice based sessions; we were therefore unable to use the document to identify texts on this criterion.

The term Report in this definition acquires its meaning in conjunction with the term Laboratory. It is the compound term ‘laboratory report’ which is the genre label; we would not argue that laboratory reports are a sub-genre of an overarching ‘report’ macro-genre. Laboratory Reports are an account of what an undergraduate engineering student has done before, during and after one of her or his ‘laboratories’. The resulting text is a structured piece of writing which is likely to include procedural description under sequential sub-headings, such as ‘Introduction - Method - Results’. It is written for assessment in an educational context.

To sum up, there are four dimensions to our working definition, namely the specific level of study, the degree course followed, the pedagogic event and the form of the resulting text. Clearly, these relate to both textual and contextual approaches to genre identification. The information provided in BAWE xls might in principle have allowed us to identify examples on the first two dimensions, but we had become unsure about the
consistency of this information. The second two dimensions could only be considered by examining individual texts in detail.

6.1 The UELR Flowchart System
We decided to examine all of the BAWE assignments qualitatively, looking in each case at the assignment PDF file in conjunction with the BAWE.xls file. We put each assignment through a flowchart system as follows:

Fig. 1. UELR flowcharting system
The first two stages were determined by cross-checking BAWE xls with the information on the assignment cover. We selected only assignments produced in years 1 to 3 of a BEng and MEng course. The third stage, of ascertaining whether the assignment was a Laboratory Report, was more complex. We needed to find indicators, whether within the text itself or within the contextual information, of the kind of pedagogic event through which the text had been produced. We arrived at the following series of indicators:

- Exophoric references to lab activities e.g. "...In this materials and production assignments associated with the mechanical testing laboratory, six types of metal specimens were available to test with the Hounsfield Type W Hand Tensometer. The aim of the experiments was to develop an understanding of the standard tensile test, to study the mechanical properties of some important engineering materials, to obtain values for the yield stress..." (BAWE Text ID 0254c).
- First person (or first person plural) pronoun (i.e. 'I' or 'We') + active verb forms e.g. "...I calculated the current to be 22 mA from the simulated circuit characteristic (Figure 1 - Appendix) with a 240 resistor was in series with the diode and the supply voltage was 6V" (BAWE Text ID 0342b).
- Authentic graphs and tables that students created using a specialist software after a series of 'laboratories'.
- Authentic photographs which were taken by students using their own camera during 'laboratories'.
- Authentic 2 or 3 D (dimensional) drawings – manually or using computers during and/or after 'laboratories'.
- Authentic mathematical calculations by students using their data which they collected during a series of 'laboratories'.
- Authentic computer (or machine) codes which have been
generated by students using specialist machines or computer software packages such as MATLAB and SIMULUS.

By considering the extent to which any of the above indicators were present, we were able to make a qualitative judgement about whether a particular assignment should be included in our subcorpus of Undergraduate Engineering Laboratory Reports.

6.2 The UELR subcorpus

The resulting subcorpus consists of 99 UELR examples from 11 specific Engineering disciplines. Assignments were written between 2002 and 2007 by 29 students (19 male and 10 female). 29 assignments are from year 1, 48 from year 2, and 22 from year 3. In more detail, the subcorpus is made up as follows (Table 1).

<table>
<thead>
<tr>
<th>Course</th>
<th>No. of assignments</th>
<th>No. of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Civil Engineering</td>
<td>2</td>
<td>2,044</td>
</tr>
<tr>
<td>Computer Systems Engineering</td>
<td>4</td>
<td>10,495</td>
</tr>
<tr>
<td>Electrical Engineering</td>
<td>1</td>
<td>1,110</td>
</tr>
<tr>
<td>Electronic and Communications Engineering</td>
<td>2</td>
<td>3,425</td>
</tr>
<tr>
<td>Electronics Engineering</td>
<td>4</td>
<td>12,751</td>
</tr>
<tr>
<td>General Engineering</td>
<td>22</td>
<td>43,865</td>
</tr>
<tr>
<td>Engineering Business Management</td>
<td>3</td>
<td>2,675</td>
</tr>
<tr>
<td>Engineering Design and Appropriate Technology</td>
<td>2</td>
<td>7,409</td>
</tr>
<tr>
<td>Manufacturing Engineering and Management</td>
<td>2</td>
<td>4,352</td>
</tr>
<tr>
<td>Mechanical Engineering</td>
<td>55</td>
<td>118,006</td>
</tr>
<tr>
<td>Systems Engineering</td>
<td>2</td>
<td>1,887</td>
</tr>
</tbody>
</table>

This subcorpus formed the basis for a detailed study of the
6.3 Identification of discrepancies in BAWE documentation

The above procedures not only allowed us to identify a subcorpus of UELR assignments, but also allowed us to identify some apparent categorization discrepancies in BAWE itself. We found a small number of instances where the information given in BAWE.xls regarding year of study or course did not correspond to the information on the front cover of the PDF assignment file. One example (BAWE Text ID 0250d) was recorded as a year 3 assignment, whereas the front cover suggests that it should be year 2.

A second example would be the case of student 6159. This student contributed ten assignments from years 2-4. For all assignments, the course is listed in BAWE.xls as an MSc in Renewable Energy, and indeed 3 of the 10 PDFs indicate this degree course. However, the remaining 7 indicate Mechanical Engineering.

When we found these apparent discrepancies, we chose to revise our contextual information so as to reflect the PDF covers. We also chose to use information on the PDF covers to categorize Engineering courses in more detail than had been done in BAWE.xls. For example, in BAWE xls 5 assignments contributed by student 0250 are given the discipline label ‘Engineering.’ Using the information on the PDFs, we recategorized these as ‘Mechanical Engineering.’ Our table above reflects the changes which we made.

Our purpose in mentioning these apparent discrepancies is not to criticize BAWE; we would argue that a margin of human error is inevitable in a project of this size, and we suspect that similar discrepancies exist in all large corpora. Rather, we want to emphasize the responsibility of corpus users to make detailed, qualitative checks of any selection of texts that they make.
7. Implications for corpus design and use

7.1 Assigning genre labels – Information for users

Corpora such as MICUSP or BAWE, which aim to make it possible for researchers to locate texts from particular genres, need to consider the extent to which they explain the development of their categorization system to users. The MICUSP information page\textsuperscript{11} states that the procedure was data driven, with random samples of papers being classified independently by a range of people to give an initial set of categories, which was then refined several times before arriving at the final list of seven categories.

The BAWE corpus manual does not provide information about how genres and genre families were identified and grouped, and commentary on the issue can be confusing. Gardner (2010: slide 22) suggests that this was a bottom-up process, involving the reading of every text in the corpus and assigning it to a genre, then grouping genres with similar purpose and similar structure into genre families. On the other hand, Gardner (2011) seems to suggest an approach in which the broader categories were identified first: "We identify 13 genre families, each with their own exclusive, specific purpose, generic staging, and genre networks. Within each of the genre families we have identified around 6 distinct genres, and more are possible" (PDF handout, p1). We would argue that it would be useful for clear information about these processes to be included in corpus documentation.

Both MICUSP simple (Beta) and BAWE documentation seem to present paper categories or genre family labels as rather factual. Both the MICUSP simple (Beta) interface and the BAWE.xls spreadsheet present this label on the same level as information such as year of study, gender of contributor, which

\textsuperscript{11}http://micusp.elicorpora.info/researchers/micusp-paper-classification
are simple facts. The assignment of a genre category is of course not a fact, but a research finding which is in principle disputable, or which might differ according to the criteria used. We would argue that it would be useful for corpus designers to make this clear, so as to encourage caution among users who base their research on the validity of such labels.

A related issue is that of the relationship between paper categories and examples (in MICUSP) or genre families and genres (in BAWE). We will illustrate this point by considering the genre label ‘report’ in BAWE. For example, according to the BAWE Corpus Manual (Appendix 2, pp. 45-50), various genres of report are listed under a wide range of genre family headings: 4 under case study, 1 under critique, 7 under methodology recount, 1 under narrative recount. This grouping seems to us to imply that the report genres grouped under the same family are at once similar to each other, and different from those reports placed under other genre families. This seems to us to be an important claim and it would be helpful to know more about the dimensions on which such decisions were made. We note that the BAWE documentation does not assign assignments or texts to genres, only to genre families. This means it is difficult for corpus users to develop their own view about the appropriacy of these categorization and grouping decisions.

7.2 Labelling genres – A corpus user decision?

Both MICUSP and BAWE offer genre category labels which are independent of discipline. Indeed, this is an important aspect of the design of both corpora, since it enables users to make selections of texts based on either genre, or discipline, or a combination of both. However, any decision about genre categorization raises all the issues about genre identification which have been, and continue to be, debated in the field
(Askehave and Swales, 2001) discuss some of the issues. In particular, we would suggest, a decision to offer genre categories independently of discipline categories raises the issue of the relative importance of sameness and difference. Is a case study in for example xxx really so similar to a case study in xxx as to warrant the same genre label?

It seems to us that such decisions must be taken by individual corpus users, based on their particular research questions. This means that many corpus users, like us, will wish to search academic corpora to find textual examples of genres which do not coincide with the labels pre-planned by the corpus designers. We hope that our description here of some of the procedures which we followed and difficulties which we encountered will highlight the issue of genre selection within a corpus, and encourage other researchers to offer their own experiences.

References


Nesi, H. 2011. The trajectory of reflective writing, paper to Reflection


Dr C.K. Jung
Department of English Language Education
Inha University, 253 Yonhyun-Dong, Nam-Gu, Incheon, Korea
E-mail: c.k.jung@inha.ac.kr

Dr Sue Wharton
Center for Applied Linguistics
University of Warwick, Coventry, CV4 7AL, United Kingdom
E-mail: s.m.wharton@warwick.ac.uk

Received: 2012.01.17
Revised: 2012.03.06
Accepted: 2012.03.08