# Towards the compilation of a corpus of assessed student writing
## An account of work in progress

*Hilary Nesi, Sheena Gardner, Richard Forsyth and Dawn Hindle*
CELTE,
University of Warwick

*Paul Wickens, Signe Ebeling and Maria Leedham*
ICELS
Oxford Brookes University

*Paul Thompson  and Alois Heuboeck*
SLALS
University of Reading
BAWE@warwick.ac.uk
http://www.warwick.ac.uk/go/bawe

## 1. Background to the project

This paper reports on the first stages of a three-year ESRC funded project to investigate genres of assessed writing in British higher education. The project aims to identify the characteristics of proficient student writing, and the similarities and differences between genres produced in different disciplines, and at different stages of university study. At the heart of the project is the creation of a corpus of British Academic Written English (BAWE), containing between three and four thousand samples of student assignments.

Most corpora of written academic text in the past have focused on professional and semi-professional writing in the public domain. The TOEFL 2000 Spoken and Written Academic Language Corpus, for example, which claims to represent "the full range of spoken and written registers used at US universities" (Biber et al. 2002, p11), contains textbooks, course packs, university web pages and similar expert sources, but no examples at all of student writing. Some collections of assignments do exist in the form of essay banks and small private corpora, but these have generally proved to be inadequate to address questions of generic variation across disciplines. Some lack adequate contextual data, for example, and many are concerned with single disciplines at a single level of study.

Our challenge is to create a corpus which reflects the broad range of writing activity at undergraduate and taught postgraduate level. In many departments the essay predominates, but a recent survey of Education, English, and Engineering has recorded sixty-four additional types of writing, including texts as diverse as business plans, web sites, lesson plans and scientific posters (Ganobcsik-Williams, 2001). We cannot hope to represent fully every possible genre in every discipline, but we aim to record major genres across a large number of degree programmes, reflecting practice in the humanities, social sciences, and physical and life sciences.

In the initial design of the corpus we have consulted a variety of information sources, including documentation at the three universities, key academics and specialist informants within departments, and evidence from a corpus of 500 assignments we developed as part of a pilot study (Nesi, Sharpling and Ganobcsik-Williams 2004).

The experience of the pilot study has also informed our decisions regarding sampling and collection policy, but not our mark-up scheme, as the pilot study did not involve the encoding of texts. We have consulted with the Text Encoding Initiative Consortium, but even within the TEI framework we have had to confront some difficult decisions in developing guidelines for the treatment of non-linguistic elements such as formulae and figures.

Our decisions regarding the different aspects of corpus compilation are discussed and explained below.

## 2. Representation in an academic corpus: the division of knowledge domains

The first manual for the Brown Corpus was published in 1964 (Kucera & Francis, 1967), which means that corpus linguistics is at least 40 years old. In fact, it is arguably more than 50 years old, as Charles C Fries based his work on the structure of English on analysis of a corpus comprising over 250,000 words recorded from telephone conversations (Fries, 1952). With such a venerable tradition, someone planning to compile a new corpus might expect to be able to turn for guidance to a well-established consensus on what constitutes good practice in corpus compilation.

### 2.1 Some Sampling schemes

However, as far as sampling procedures are concerned, the guidance that does exist is sparse. Very few corpus linguists would argue that the principles of statistical sampling are completely irrelevant to their task; but it is clear that theoretical statistical concerns have seldom been uppermost in the minds of corpus compilers. Many eminent practitioners have cast doubt on statistical sampling as an appropriate option in practice. Three short extracts follow.

> “It follows then that the LOB Corpus is not representative in a strict statistical sense. It is, however, an illusion to think that a million-word corpus of English texts selected randomly from the texts printed during a certain year can be an ideal corpus.” (Hofland & Johansson, 1982: 3.)
>
> "Unfortunately, the standard approaches to statistical sampling are hardly applicable to building a language corpus." (Atkins et al., 1992: 4.)
>
> "For language studies, however, proportional samples are rarely useful." (Biber et al., 1998: 247.)

At this point, it is probably wise to clarify some terms used by statisticians in order to highlight the main differentiating features of various widely used sampling techniques. Table 1 provides a glossary; for further information see Barnett (1991).

| **Term** | **Description** |
|---|---|
| (Simple) Random Sampling | In this process a subset of the target population is chosen at random under conditions that ensure that every subset of the same size has an equal chance of being selected. In the simple version, every member of the population has an equal chance of being selected. |
| Stratified Sampling | In this procedure, the population is first divided into *strata,* and then random (sub)sampling is performed within each stratum. |

| | |
|---|---|
| | The strata should, between them, cover the whole population. |
| Cluster Sampling | In this procedure, the population is divided into subgroups (e.g. in localized geographic regions), then only some of those subgroups are chosen for sampling. Sometimes each whole subgroup is selected; in other cases, only a (random) subset of each subgroup is selected. The clusters don't necessarily cover the whole population. |
| Quota Sampling | In this procedure, the population is cross-classified along various dimensions deemed important and quotas are established for each of the subgroups defined by the cross-classification (e.g. male smokers between 40 and 49 years of age). Then individuals are picked to populate the grid thus defined until each cell is filled. Note that human judgement is used to pick individuals; thus this is not a random process. |
| Opportunistic Sampling (aka Convenience Sampling) | This is the technical name for just taking whatever members of the population are easy to get hold of. It is unreliable for drawing formal inferences about the properties of the target population. |
| Purposive Sampling (aka Judgemental Sampling) | This is a procedure where a person, ideally an expert in a relevant field, chooses cases on the basis of a judgement as to how well they, between them, exemplify a particular population, rather in the manner that the editor of an anthology might choose poems to represent (for example) British Poetry of the First World War. It is unreliable for drawing formal inferences about the properties of the target population |

**Table 1: Glossary of major sampling techniques**

It should be noted that stratified sampling requires (1) that the strata jointly cover the whole population, and (2) that cases from each stratum are chosen randomly. If these conditions do not hold, the process should properly be called cluster or quota sampling.

Statisticians warn of the temptation to extrapolate from a sample to its parent population, once some data has been obtained and analyzed, even though we may have originally gathered the sample without intending to draw firm conclusions about the population it comes from. This is acknowledged by Atkins, Clear and Ostler, in the article cited earlier:

> When a corpus is being set up as a sample with the intention that observation of the sample will allow us to make generalizations about language, then the relationship between the sample and the target population is very important. (Atkins et al., 1992: 5.)

It therefore behoves us as corpus compilers to be explicit about the composition of our corpus, and our sampling techniques. From a statistical perspective, the use of one of the four "lower" (non-random) sampling methods can only be justified on practical grounds.

In our project we have decided to employ a method in which the population of scripts is cross-classified according to year and disciplinary grouping. This gives a 4-by-4 matrix (four years: first, second, third year undergraduate, and taught post-graduate; four disciplinary groupings: biological & health sciences, physical sciences & engineering; social sciences & education; humanities & arts). Each cell in this 4-by-4 grid is to be filled with approximately the same number of scripts. However, although the scripts are grouped into strata, the sampling method will not be random; cells in the grid will be filled with scripts that are either gathered opportunistically or selected purposively from those made available to us.

Our choice of a non-random sampling system relates to the intended purpose of the corpus (as a means of identifying genres of assessed student writing) and to practical considerations. Simple random sampling, in which every member of the population (of assignments) has an equal chance of being selected, is out of the question. We do not want proportionally more assignments from those modules which attract large numbers of students, because we are equally interested in all genres regardless of quantity of output, and because the distribution of students in any case varies from year to year, and differs from one institution to another in the UK. There is also no practical way of accessing a random sample of all the proficient assessed writing produced within the three participating universities. Even if we identify instead a random sample of students, we have no means of ensuring that they will come forward to submit their work, or that they will have retained in electronic format an appropriately random sample of assessed work which also meets the agreed proficiency criterion (all assignments in the corpus must have been awarded the equivalent of at least an upper second degree mark).

Similarly, if we divide our population into strata it is practically impossible for us to identify a random sample of assessed work at the required proficiency level from each stratum, because we are dependent on students to submit this work, and they may not be able to produce the random sample that we require. Moreover, we suspect that if assignments were selected entirely randomly, from a very wide range of modules, some disciplines would be represented so patchily that it would be impossible to make general claims about the genres they employ. Our experience with the pilot project supports this belief (Nesi, Sharpling and Ganobcsik-Williams 2004).

We are aware, however, that the cells in our 4-by-4 matrix do not represent categories that are absolute and mutually exclusive: not all real student writers or modules fall neatly within one of the four years of study or disciplinary groupings. As noted by McEnery & Wilson (1996: 65), strata, as typically defined by corpus linguists, “are an act of interpretation on the part of the corpus builder because they are founded on particular ways of dividing up language into entities such as genres which it may be argued are not naturally inherent within it”. We will discuss some different perspectives on genre and sampling in the following sections.

### 2.2. Subjects and strata

Ideally academic disciplines, being ever-evolving and permeable, are best regarded as bundles of feature-values, i.e. objects with many attributes, among which similarity judgements can be made, but which do not fit tidily into any tree-structured taxonomic hierarchy.

For practical purposes academic subjects have to be classified, however, and many different classification systems have been proposed. Librarians developed the (mutually incompatible) Dewey Decimal and the Library of Congress systems, Commercial information providers like the World of Learning (www.worldoflearning.com) classify academic activities into disjoint domains, and in Britain UCAS has developed a classification scheme called JACS (Joint Academic

Classification of Subjects) with 19 high-level groupings. Meanwhile, the UK Research Assessment Exercise recognizes 25 major subject fields.

In the present context, we were faced with the problem of trying to ensure that our corpus contains a relatively well-balanced sample from all the main disciplinary groupings in British universities. As a first step to wards clarifying these intentions it is instructive to look at the high-level groupings of academic subjects used by some earlier corpus compilers. Table 2 lists the labels used in four major projects (five, strictly speaking, but LOB deliberately copied Brown).

| Brown (& LOB)<br>Category J | LSWE<br>disciplines<br>(academic books) | MICASE<br>academic divisions | T2K-SWAL<br>disciplines |
|---|---|---|---|
| 1. Natural Sciences<br>2. Medicine<br>3. Mathematics<br>4. Social & Behavioral Sciences<br>5. Political Science, Law, Education<br>6. Humanities<br>7. Technology & Engineering | 1. agriculture<br>2. biology/ecology<br>3. chemistry<br>4. computing<br>5. education<br>6. engineering /technology<br>7. geology /geography<br>8. law/history<br>9. linguistics /literature<br>10. mathematics<br>11. medicine<br>12. psychology<br>13. sociology | 1. Biological & Health Sciences<br>2. Physical Sciences & Engineering<br>3. Social Sciences & Education<br>4. Humanities & Arts | 1. Business<br>2. Education<br>3. Engineering<br>4. Humanities<br>5. Natural Science<br>6. Social Science |
| Kucera & Francis (1967);<br>Hofland & Johannson (1982). | Biber et al. (1999). | http://www. lsa.umich.edu/eli/micase /MICASE_MANUAL.pdf (2003) | Biber et al. (2004). |

**Table 2: Top-level academic groupings used by four major corpora.**

It is evident from this table that there is no standard way of slicing up the academic world. As Becher (1990: 335) says: "discipline …. cannot be regarded as a neat category". This conclusion is reinforced by the fact that the compilers of the British National Corpus used a different approach altogether:

> "target percentages for the eight informative domains were arrived at by consensus within the project, based loosely upon the pattern of book publishing in the UK during the past 20 years or so." (Aston and Burnard, 1998: 29.)

Informative texts were in fact placed in one of nine different "domains" in the BNC (the ninth being a miscellaneous category), as shown in Table 3.

| Number | Domain |
|---|---|
| 1. | Arts |
| 2. | Belief & thought |
| 3. | Commerce & finance |
| 4. | Leisure |
| 5. | Natural & pure science |
| 6. | Applied science |
| 7. | Social science |

| | |
|---|---|
| 8. | World affairs |
| 9. | Unclassified. |

**Table 3: Non-fiction domains in the BNC**

Faced with this confusion, we have chosen to follow the division of academia into four high-level groupings, as used in MICASE and in the corpus of British Academic Spoken English (BASE) which was recently compiled at the Universities of Warwick and Reading (Nesi, 2001, 2002). This system has the merit of allowing some degree of comparability between corpora, and is broad enough to accommodate many university modules which might straddle more highly specified groupings. Nevertheless a number of problem cases remain. Readers may want to try to categorise the five modules listed in Table 4, which are all currently offered at the University of Warwick. Of course there is no ‘right’ method of categorisation, but in practice we have been guided by the two letter prefix to the module code. EC indicates, for example, that Mathematical Economics is taught by the economics department, and therefore belongs within the broad category of Social rather than Physical Science, while Physics in Medicine is taught primarily to students in the physics department, and might therefore be regarded as a Physical Science module.

- ■ CS231 Human Computer Interaction
- ■ EC221 Mathematical Economics 1B
- ■ MA235 Introduction to Mathematical Biology
- ■ PS351 Psychology & the Law
- ■ PX308 Physics in Medicine

**Table 4: Some Specimen Modules**

Clearly, categorisation depends to some extent on a consideration of the context in which assignments are produced, and by extension the demands and expectations of the academics who assess them.

## 3. An emic perspective on assessed genre: gathering contextual evidence

In later stages of our project we will conduct multivariate analysis of the corpus as a whole (taking a **register perspective**), and qualitative analysis of various subcorpora (taking an **SFL genre perspective**). It also important from the very start of the project, however, to take an **emic perspective**. This entails analysing information provided by members of the various discourse communities that make up the academic community as a whole.

The provision of full contextual information on all texts in a corpus of any size must be an impossible – or at least impractical – task. Decisions have to be made about what to collect, and then what to include in the corpus. Users will expect and need metadata, or information to contextualise the texts. However, as “it is far less clear on what basis or authority the definition of a standard set of metadata descriptors should proceed” (Burnard 2004: 1), and as there are no similar corpora of student writing for us to emulate, we have some flexibility in the decisions we make. Our decisions are based initially on meeting the needs of the current research, but also on anticipating possible future uses. They take into account what might be useful in an ideal world, and are modified for practical reasons such as time and money, for technological reasons such as what can easily be stored or retrieved, and for legal reasons related to data protection legislation, copyright and potential plagiarism.

### 3.1 Issues in constructing categories for contextual information

In order to gather information for an emic perspective on assessed genres we rely on three main sources:

a) departmental documentation
b) tutor interviews and surveys
c) information from student disclaimer forms

Departmental documentation, including print and on-line handbooks and module outlines, enables us to build a profile of each department with information such as lists of modules with tutors, assignment briefs, assessment criteria, and advice on academic writing. From this we develop an initial list of assessed writing. For example, in Sociology, the types of assessment referred to in the undergraduate module descriptions include:

- Essays (most frequently)
- Book Reviews
- Book Reports
- Projects
- Urban Ethnography Assignment
- Fieldwork Report
- Dissertations

Tutor interviews play a crucial role in affirming items in the departmental documentation. They also provide us with a list of assignment names with brief descriptions, information on academic attitudes, and practical information to help with the subsequent collection of assignments. These interviews enable us not only to identify assignment types by name and module location, but also to begin to understand their intended nature. We have adopted a semi-structured approach, using a series of open questions designed to capture general perceptions and factual detail (see Appendix 1 for our academic interview guidance notes).

We also plan to use the findings from these interviews to construct an electronic questionnaire, for much wider distribution. The questionnaire, to be sent to all tutors at the three universities, will elicit further information about types of assessed writing, progression observed from 1st year to taught masters level, and tutors' attitudes to various features of the texts students produce.

The third approach to establishing emic genre types involves identification by the student. When collecting assignments from students we ask them to complete a disclaimer form which asks them, *inter alia*, to identify the type of assignment they are submitting (see Appendix 2).

Thus we have three perspectives from the discourse community on the different types of assessed writing. Ideally these will converge, and where they do not, clarification will be sought from appropriate departmental sources. For instance, it remains to be clarified in the case of Sociology whether a distinction is being made between Book Reviews and Book Reports; and whether an Urban Ethnography is one particular type of Field Report, or whether the terms as used in the department are mutually exclusive. Only through investigating the discourse community and its assessment practices can we answer these questions and arrive at a satisfactory description of assessment genres from an emic perspective.

A distinction is generally drawn between contextual and linguistic information. "The social context (that is, the place, time, and participants) within which each of the language samples making up a corpus was produced or received is arguably at least as significant as any of its intrinsic linguistic properties – *if indeed the two can be entirely distinguished*" (Burnard 2004:8, italics added). In a similar vein, Atkins, Clear and Ostler make a distinction between external and internal criteria for constructing a corpus for linguistic analysis (1992:5):

> The internal criteria are those which are essentially *linguistic*: for example, to classify a text as formal/informal is to classify it according to its linguistic characteristics (lexis/diction and syntax). External criteria are those which are essentially non-linguistic. … These attributes are all founded on extra-linguistic features of texts (external evidence). Of course, the internal criteria are not independent of the external ones and *the interrelation between them is one of the areas of study for which a corpus is of primary value*. In general, external criteria can be determined without reading the text, thereby ensuring that no linguistic judgements are being made. (italics added)

In view of this external – internal dichotomy, we would class our emic categories as external, ascribed by the discourse community, whereas the multivariate analysis and SFL genre categories are internal, resulting from linguistic analysis of the texts. As it will indeed be of value to consider the interrelation between these internal and external categories, all three criteria will be recorded in the metadata (see below). However, this seems to contradict the assumptions made by Atkins, Clear and Ostler later in their article, where they appear to adopt a more realist account of the criteria by saying, for instance, that text attributes from external criteria to be included in metatext would include 'mode', 'style' and 'region'. An example of 'mode' would be *written*, and as this is arguably a linguistic feature, it seems that external features can be linguistic, but should not require close reading or detailed analysis of the text. Examples of 'style' are prose, verse, and rhyme as determined by "*surface features of text or author's claims*", suggesting an underlying assumption that the author's claims will concord with an analysis of surface features. What if they do not? 'Region' relates to the regional type of the language of the authors, a parameter which, according to Atkins et al "will be refined by internal evidence" (p. 8). Here we see explicit slippage between or convergence of external and internal criteria and a realist assertion that there will be one 'correct' category for each text attribute – denying future users the chance to study the interrelation between external and internal criteria earlier deemed to be of great interest. Presumably, then, we can exercise discretion concerning which categories we conflate, and which we retain for interrelational studies.

The problems with conflation arise, we would suggest, when we try to conflate external, emic categories used in the discourse community, or self-reported attributes (such as first language), with internal categories resulting from our linguistic analysis. As suggested by Gail Forey (2004) who combined in her analysis of business texts an external social perspective approach with an internal text analysis approach, certain linguistic choices are construed and interpreted differently by members of the business discourse community and EFL teachers. We would argue, therefore, that the corpus should include not only categories from the discourse communities, but also from the analysis of surface features and from the more qualitative genre analysis. These are ontologically different entities – and therefore slippage between them is not possible. The labels and realisations may well correspond in many instances, but it would be foolish at this stage to assume that there is one ‘correct’ assignment type that would satisfy the discourse community, the multivariate analysis and the SFL analysis. Indeed the whole point of the research is to discover where they do coincide and where they do not. How similar are essays in history and English when viewed from an analysis of register and genre? What about essays in philosophy, business or biology? We assume by looking for comparisons across disciplines that there will be differences, and that the same assignment label may be given to texts from different disciplines with very different characteristics. It is therefore essential, we would argue, that these categories are not conflated.

One final point is worth noting here: Linguistic analyses are not carried out in a contextual vacuum. The choice of texts and the variables analysed are all selected on the basis of assumptions about how texts might be characterised. Thus even the multivariate analysis relies to some extent on contextual decisions, and for the SFL analysis this relationship is much more explicit: meanings are construed in context, language reflects context, so in order to analyse the genres from an SFL perspective we must have contextual understanding. We start from our own understandings of academic writing and assessment, informed by experience and reading in the field, but these must be supplemented by contextual information from the discourse community. Without this external contextual information, interpretation of linguistic analysis is impossible. The question remains, how much contextual information might future external users need if they also wish to conduct genre analyses of the subcorpora?

### 3.3 Identifying progression through undergraduate to masters level

In an oversimplified world, 2nd year students would take 2nd year modules in their specialist subjects in their second year of university study, having progressed from school to university with no gap year or longer break, and would write all their assignments in the same month each year. In reality, of course, degree courses may be three or four years in length, and students may take modules from outside their specialisation or from another year of study, or may take an intercalated year in the middle of their programme. Assignments are submitted and returned to students at different times of year in different departments.

To inform our study of progression we try to capture data not only concerning the ‘year’ of the student, but also concerning the home department of the student and the location of the module within courses. Because of the practical difficulties of collecting texts we are not limiting our collection to full-time students on particular courses, or to 3rd year students only in 3rd year modules, for instance, but we are recording the contextual data that should allow the creation of subcorpora to reflect progression through the academic years.

### 3.4 Writing in or across the disciplines?

A further complication arises when we consider modules taken by students from different home departments. For instance one of the law tutors we interviewed was aware that even very good students from the English department would be at a disadvantage in a law module, because they would not be immersed in legal ways of writing, and that this raised assessment issues for the department. Such students might still get good marks, but their writing might not be considered 'good' legal writing. To flag this up, we are collecting students' self reported *home department*.

### 3.5 Language and educational background

Although as applied linguists we are interested in the writing of non-native speakers, we have not designed the corpus to include a balanced sample of native vs non-native writing, as this is not a major focus for the project. We appreciate, however, that proficient writing in British universities does not emerge from a uniform linguistic or educational background, and we have felt it important to note whether the student would self-identify as a native speaker of English, and whether they were educated in the UK for the bulk of their secondary education. This would allow us to exclude texts written by students from China, Germany or the United States, for instance, if we wished to identify assignments produced by more typically British students.

### 3.6 The role of assessment

As we progress with interviews and data collection, our categories are increasingly refined. For example we have addressed the core question of what counts as assessed writing. A key distinction at Warwick is between formative assessment, which will not contribute to the final grade for the module, and summative assessment, which will determine the grade for the module (these are clearly emic categories, wide open to challenge by the specialist assessment literature, for example Rea-Dickins and Gardner 2000). Initially it was anticipated that we would only be interested in summative assessment. However, following interviews with academics, we have found that formative work is viewed as equally important in the development of assessed writing, and may include genres not summatively assessed. As we are less concerned with whether the piece of work counts towards the final degree, and more with genre, it was agreed to include formative assessments. However, as we are only accepting work above a pre-defined proficiency level (IIi/65%+) it is important that a grade has been given to the work in every instance.

### 3.7 Fiction and foreign language texts

We have also been forced to reconsider what counts as written academic English. Faced with posters, models of theatre stages and other multi-modal pieces of assessed work, we have restricted the corpus to those where the written component is the major contribution that is assessed. So, for instance, we have decided to omit powerpoint presentations – which are certainly text that we could analyse – on the grounds that what was assessed was the whole presentation, including the spoken and visual complements to the powerpoint text, so we could not guarantee that the written slides represented the bulk of the assignment.

Following a similar rationale, we are excluding texts written primarily in a foreign language (such as French) or in formulae (such as algebraic equations). Texts which are predominantly English but include foreign language or formulae are included.

Fiction is another borderline area, initially excluded on the grounds that it was written for literary purposes rather than academic assessment. This view was challenged in an interview with the director of undergraduate studies in Sociology at Warwick:

> "we're quite a traditional department in that we still use mainly essays, we're very conscious that we would like to, and perhaps need to, do something about that" (Mike Neary, 26/04/2005)

Dr Neary went on to describe 'crime fiction', an assessment innovation in the Crime and Deviance module where students write a crime short story which is assessed by academic criteria. This was a new type of task, set for the first time this year, and so the criteria are still evolving, but it is illustrative of how the nature of assessed writing is constantly changing. Assignment types such as these are impossible to anticipate and their inclusion will be instrumental in challenging assumptions about the genres of assessed writing currently occurring in British higher education.

## 4. The collection process: dealing with students and their texts

The pilot study for the BAWE corpus indicated that the collection of the texts was likely to be one of the more time consuming and potentially frustrating processes in the construction of the corpus. Three main issues emerged from consideration of the pilot study:

1. the need to promote the project to the right people to ensure a supply of texts that meet our criteria
2. the need to provide a level of monetary incentive that ensures that the corpus can be collected within a reasonable timeframe and budget
3. the need to make the collection process both efficient and as user friendly as possible.

As we have seen, the sampling frame is based around the construct of 'disciplinary groupings' and specifically through a selection of modules based within a range of departments. The benefit of this sampling frame from a text collection perspective is that it enables access to the students and their assessed writing through the administrative systems of the institutions. Once contacts are established with course and module leaders, permission can be sought to contact students through lectures and/or cohort email lists. Other administrative tasks such as the validation of the marking criteria can be readily achieved once such contacts have been established.

Presentations in lectures are supported with a flyer explaining the overall research project, and a cohort specific handout detailing our criteria for text collection and contact details. Students' questions have highlighted the need for face to face clarification of specific questions concerning, for example, our collection criteria, the overall purpose of the corpus, and the degree of writer anonymity we can provide. Whilst we have had some success with a 'cold call' email to a cohort, collection rates were much better where we have established face to face contact. Cohort email lists may be a good strategy to fill in gaps in the sampling frame with groups that have already been contacted face to face.

In the light of findings from the pilot study (where contributors were paid first £1 and subsequently £2 per assignment), we set the rate of £3 for each assignment we accepted. With multiple submissions (maximum of nine) a student can receive a reasonable amount of money and this has probably contributed to our initial success in the collection process. The payment enables us to select and reject texts and motivates contributors to complete submission forms for each assignment.

Students are strongly encouraged to submit their work via email and we select those texts that meet our sampling frame. In our reply we indicate where and when submission can be made, and also attach the submission form. Typically we collect assignments on campus, in the students' own departments, but we have also designated fixed weekly times when we can receive assignments in our offices. Some submissions are collected on the spot but most are pre-arranged through email. This system is beneficial as it allows students to fill in forms in their own time and check in advance the amount of money they will receive. From our point of view, it allows time to match up offers of texts with the sampling frame (avoiding overspend) and time to perform any necessary checks on the submitted information forms.

We are particularly concerned to ensure an efficient paper trail at all times. A spreadsheet related to the sampling frame records expected submissions (received via email) and final submissions (permission forms signed and paid). It is important that this is regularly up-dated to keep pace with email applications, as we sometimes build up large numbers of expected submissions. The BAWE email is dealt with by one person to ensure consistency but is copied to other team members at the same university in case of absence. The contextual data on the information sheet is transferred to a spreadsheet and anonymised as soon as possible. From here the text moves forward to the generation of the header and the tagging of the text.

## 5. Text encoding issues

The corpus will be marked up according to the TEI guidelines (TEI P4) and will be stored in XML format. Since one of the main objectives of the project is to recognise distinct genres of student writing, special care will be taken to encode *layout* and *structural* features. Contextual information such as the author's gender, age, first language, course and level of study, as discussed in section 3 of this paper, is stored in the header of each file.

The encoding of *layout* ensures an objective representation of the text. Special attention is paid to *formatting information*, which potentially may serve a wide range of purposes (not to be analysed further at this stage). Generally we focus on the encoding of kinds of "highlighting" by character format: bold and italic type, underlining, a change in text colour, sub- and superscript.

The encoding of *structural features* is essential for locating phenomena within a text. Different parts of the document are of unequal importance and the "main" part, *running text*, may be seen as the text proper. Thus, a three-fold document structure arises: the running text, constituting the *body*, is separated from everything before (the *front*, including for example the title page and epigraphs) and after (the *back*, including for example the bibliography and any appendices).

Some elements embedded in the body are not considered as part of the running text proper, however. We here identify *tables*, *figures* and *notes*. While tables and figures are taken out of the text, leaving only an ID attribute pointing to the original text, notes are fully marked up. The "body" most importantly consists of *sections*, which in turn contain *paragraphs and sentence-like ("s"-) units*. Because they constitute a particular means of text progression, *lists* are also marked up.

Many contextual items and elements that are typically found in academic writing also play a role in defining the structure of the text. It is thus worthwhile marking up some *specific section types* characteristic for academic discourse, including *abstracts* (which may occur before or after the running text), a *bibliography section* and *appendices*. Additionally, some distinctive functional items may occur within running text sections and will be encoded: *formulae* and *block quotes*.

The general structure of a BAWE document is illustrated in Appendix 3. It should be noted that our mark-up strategy takes practical matters into account: the cost (implied workload) of mark-up can be a decisive argument for or against its inclusion. The encoding of information is, at the present stage of the BAWE corpus, guided by the overall concern of transposing from source (DOC, RTF) to target document format (XML). It is crucial to encode all information judged relevant that would otherwise be lost or retrievable only at great cost once formatting information has been removed from the document.

The encoding strategy presented here is not a detailed account and by no means complete; the BAWE document model will, very likely, be further refined. We anticipate, for example, that the corpus will be tagged for part of speech when we conduct multivariate analysis at a later stage in the project, in consultation with Doug Biber’s research team at Northern Arizona University.

## 6. Conclusion

As explained earlier, we are as yet only in the first stages of our project ‘An investigation of genres of assessed writing in British higher education’. In due course we expect to collect between three and four thousand samples of student work, amounting to about ten million words, which we will then subject to multivariate analysis and also sample for more detailed genre analysis. In this paper, before the corpus is complete, we think it is fitting to focus on our corpus compilation policies. In so doing we hope to justify decisions that have sometimes been difficult, and which have taken us many months to reach.

We also hope that this paper will provide sufficient procedural detail to encourage others to replicate our project, in whole or in part, in other discourse communities which we are not ourselves able to reach. We would be delighted to be able to compare our corpus findings with those of other corpora of student writing, compiled in other countries or at other levels of study.

We look forward to future exchanges with other corpus linguists on this topic.

## Acknowledgement

The pilot for this project was funded by the University of Warwick Research and Teaching Development Fund. The current project (2004- 2007) is funded by a research grant from the ESRC (RES-000-23-0800).

## References

Aston, G. and Burnard, L. (1998) *The BNC Handbook* (Edinburgh: Edinburgh University Press).

Atkins, S., Clear, J. and Ostler, N. (1992) Corpus design criteria. *Literary & Linguistic Computing*, 7, 1-16.

Barnett, V. (1991) *Sample Survey Principles and Methods* (London: Edward Arnold).

Becher, T. (1990) The counter-culture of specialisation. *European Journal of Education*, 25, 333-346.

Biber, D., Conrad, S. and Reppen, R. (1998) *Corpus Linguistics: investigating language structure and use* (Cambridge: Cambridge University Press).

Biber, D., Johansson, S., Leech, G., Conrad S.and Finegan, E. (1999) *The Longman Grammar of Spoken and Written English* (London: Longman).

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

Biber, D., Conrad, S. Reppen, R. Byrd, P. Helt, M. Clark, V. Cortes, V. Csomay, E. and Urzua, A. (2004) *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus.* TOEFL Monograph Series. (Princeton, NJ: Educational Testing Service).

Burnard, L. (2004) Metadata for corpus work. Available on-line from http://users.ox.ac.uk/~lou/wip/metadata.html (accessed June 6th 2005)

Forey, G. (2004) Workplace texts: do they mean the same for teachers and business people? *English for Specific Purposes*, 23, 447-469

Fries, C.C. (1952) *The Structure of English* (New York: Harcourt and World).

Ganobcsik-Williams, L. (2001). Teaching writing at university: a survey of staff perspectives. Paper presented at *Teaching writing in higher education: an international symposium*. (Warwick Writing Programme, the University of Warwick. March 2001).

Hofland, K. and Johansson S. (1982) *Word Frequencies in British and American English* (Bergen: Norwegian Computing Centre for the Humanities).

Kucera, H and Francis, W. (1967) *Computational Analysis of Present-day American English* (Providence, RI: Brown University Press).

McEnery, T. and Wilson, A. (1996) *Corpus Linguistics* (Edinburgh: Edinburgh University Press).

Nesi, H., Sharpling, G. and Ganobcsik-Williams, L. (2004) The design, development and purpose of a corpus of British student writing. *Computers and Composition,* 21, 439-450

Nesi, H. (2002) An English spoken academic word list, in: A. Braasch and C. Provlsen (eds.) *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002 Volume I.* (Copenhagen: Center for Sprogteknologi), 351-358

Nesi, H. (2001) A corpus-based analysis of academic lectures across disciplines, in J. Cotterill and A. Ife (eds) *Language Across Boundaries: BAAL 2000 Proceedings.* (London: BAAL in association with Continuum Press), 201-218

Rea-Dickins, P. and Gardner, S. (2000) Snares and silver bullets: disentangling the construct of formative assessment. *Language Testing,*.17, 215-243

Simpson, R., Lee, D. and Leicher, S. (2003) *MICASE Manual* (Ann Arbor: University of Michigan).

# Appendix 1 - Academic Interview Guidance Notes

- What role does assignment-writing play in your department?

- Can you tell us what different types of written assignment you set your students?
    - Could you tell us more about ZZ?
    - Are there other types of assignment task that you haven't mentioned?
    - Do you set assignments of type [pre-existing genre] as well?
    - Do you use other formats, e.g. non-written assignments, such as videos?

- What are the main ways in which the various types of assignment you set differ?
    - How could we tell a YY from an XX?
    - e.g. an experimental report from a case-study?
    - e.g. a critical review from an essay?

- What sort of differences do you expect to find between the written work of first & second year students on the one hand and final-year undergraduates & masters-level students on the other hand?

- What do you value most in student written work?

- What are the sorts of things that you most dislike finding in student's written work?

- In your opinion, how much does presentation matter?

- How do the various assignment tasks that you set enable you to judge whether a student has shown evidence of the qualities you value?

- Do you find that overseas students have particular problems with written assignments, compared to native English speakers?
    - Do you have any particular ways & means of helping them overcome these problems?

- Who should we talk to about collecting assignments?
- Is there a good time to collect assignments on module MM999?
    - [Opportunity to explain that we're hoping to get 5 or 6 good-quality assignments from 2 or 3 modules at each level (years 1-3 & masters).]
- Are there any modules in your dept which you think we should definitely include in our sample?
    - if so, which are they? and why?

## Appendix 2 - BAWE Corpus of British Academic Written English Submission form

Surname: ........................................ Forename: ....................................
Male/Female Date of birth: ................................
Preferred email address: …....………......……………………………………….......
Your first language: ..................................................................................
Your secondary education (since 11 years old but before university) was:

| | | |
|---|---|---|
| All in UK. | All overseas. | Some in UK, some overseas.<br>Please state number of years in UK ........ |
| | | |

Your year of study (when you wrote the assignment):

| | | | | |
|---|---|---|---|---|
| first year under-graduate | second year under-graduate | third year under-graduate, no intercalated year | fourth year under-graduate, with intercalated year | post-graduate (masters level or diploma) |

Other (please specify): ....................................................................
Your home department: ............................................................................
Course of study: .......................................................................................
Brief Title of assignment: ..........................................................................
Year & Month when assignment written: ..................................................…
Module title: ...............................................................................................
Module tutor’s name: ..................................................................................
Module code: .......................... Grade/Mark received: .............................

Please indicate the type of assignment, according to your understanding of the task, by choosing 1 of the options below:

Case-Study / Essay / Exercise / Notes / Presentation / Report / Review
/ none of the above (please specify):

# Appendix 3 - The general structure of a BAWE document

```
<TEI.2 id="BAWE_text">
<teiHeader type="text">
...
</teiHeader>
<text>
<front>
<titlePage>
<docTitle>
<titlePart type="main">...</titlePart>
</docTitle>
</titlePage>
<div1 type="abstract">
<p><s>...</s></p>
</div1>
</front>
<body>
<div1>
<head>..</head>
<p><s>...</s> <note place="foot" n="BAWE-note1"><p><s>...</s></p></note>
<s>...</s></p>
<figure id="BAWE-fig1"/>
<quote>...</quote>
<list>
<item><p><s>...</s></p></item>
</list>
<p><formula notation="" id="BAWE-form1"/></p>
<table id="BAWE-tab1"><row><cell/></row></table>
</div1>
</body>
<back>
<div1 type="bibliography">
<head>...</head>
<p>...</p>
</div1>
<div1 type="appendix">
<head>...</head>
<p>...</p>
</div1>
</back>
</text>
</TEI.2>
```