Encoding document information in a corpus of student writing

The experience of the British Academic Written English (BAWE) corpus

Signe O. Ebeling, Oxford Brookes University, UK

s.o.ebeling@ilos.uio.no

Alois Heuboeck, University of Reading, UK

a.heuboeck@reading.ac.uk

to appear in: Corpora 2 (2), 2007

Encoding document information in a corpus of student writing

The experience of the British Academic Written English (BAWE) corpus

ABSTRACT

The information contained in a document is only partly represented by the wording of the text; in addition, features of formatting and lay-out can be combined to lend specific functionality to chunks of text (e.g. section headings, highlighting, enumeration through list formatting etc.). Such functional features, although based on the ‘objective’ typographical surface of the document, are often inconsistently realised and encoded only implicitly, i.e. they depend on deciphering by a competent reader. They are characteristic of documents produced with standard text-processing tools.

The present paper will discuss the representation of such information with reference to the *British Academic Written English* (BAWE) corpus of student writing, currently under construction at the universities of Warwick, Reading and Oxford Brookes. Assignments are usually submitted to the corpus as Microsoft Word documents and make heavy use of surface-based functional features. As the documents are to be transformed into XML-encoded corpus files, this information can only be preserved through explicit annotation, based on interpretation. The choices made in the BAWE corpus are presented and practical requirements for a tagging interface discussed.

1. INTRODUCTION

The *British Academic Written English* (BAWE) corpus of student writing is currently under construction at the Universities of Warwick, Reading and Oxford Brookes. Work on the corpus was started at Warwick in 2001, when a pilot study was carried out over a period of eighteen months (see Nesi et al., 2004). The resulting pilot corpus stands at about 500 student assignments and is the forerunner to the current corpus project.

The BAWE corpus will be the main source of data for a cross-disciplinary research project entitled *An investigation of genres of assessed writing in British Higher Education*.[1] On the basis of material from the corpus, the project aims to identify salient characteristics of student assignments. In particular, the nature of proficient student writing will be investigated and we hope to be able to tell what characterises such writing, especially with regard to genres as represented across disciplines. In the future, the corpus will also be available as a resource to other researchers who wish to carry out studies on student writing in Britain.

In this article we will mainly focus on the part of the research project that concerns the development of the corpus. We will give an overview of some of the most important steps involved in the compilation of the BAWE corpus, including the encoding and mark-up process. The main bulk of this paper, focusing on encoding and mark-up issues, will be preceded by a brief description of the BAWE corpus and its structure in section 2.

Although the most time-consuming part of the corpus compilation process will be to collect the student assignments and to prepare them for inclusion in the corpus, the collection process will not be dealt with here; for details see Nesi et al. (2005).

Most of the assignments are submitted as Microsoft Word files. However, this format turns out to be impractical for efficient processing of a corpus. The documents need to be converted to a plain text format, which in turn involves preprocessing them in order to avoid loss of relevant information. To facilitate this stage, some effort has gone into the development of computer tools enabling semi-automatic and automatic processing of the assignments collected. The tools comprise a series of scripts: Visual Basic scripts for the semi-automatic processing of the Word documents and Perl scripts for the automatic post-processing. The steps involved in the preparation of the texts will be dealt with more fully in Sections 3 and 4.

2. CORPUS STRUCTURE

The core BAWE corpus aims to include 3,500 student assignments from four disciplinary groupings, viz.:

Arts and Humanities	Applied Linguistics, Archaeology, Classics, English Studies, History, History of Art, Philosophy, Theatre Studies
Life Sciences	Agriculture, Biochemistry, Biological Sciences, Food Science and Technology, Health and Social Care, Medical Science, Psychology
Physical Sciences	Architecture, Chemistry, Computing, Cybernetics & Electronic Engineering, Engineering, Mathematics, Physics
Social Sciences	Anthropology, Business, Economics, Hospitality, Leisure and

Tourism Management, Law, Publishing, Sociology

Although the disciplines in the list above have been assigned to one of four disciplinary groupings, the overview does not reflect a rigid division in this respect. While it is recognised that the initial assigning of disciplines to disciplinary groupings is not unproblematic, in that for instance some disciplines span more than one disciplinary grouping, or fall between two (Nesi and Gardner, 2006), this initial assignment of disciplines to disciplinary groupings represents our best understanding of the nature of the disciplines as taught in our sample institutions.

The corpus will comprise both undergraduate work (typically three years of study) and postgraduate work (typically one year of study) of a high quality, i.e. all assignments are graded II.i (B+) or above. To ensure that all four levels of undergraduate and postgraduate studies are represented in the corpus, we aim for a specific number of assignments per year of study per discipline.

In order to limit the contribution by individual students, each student can submit up to ten assignments per course of study, with a limit of five per year of study (and a maximum of three per module). Since the assignments vary in length, the exact size of the corpus can only be determined after completion. However, based on experience from the pilot corpus, the 3,500 assignments will amount to a corpus of around 10 million words.

For each assignment a set of metadata is recorded, including student gender, year of birth, first language, course of study, year of study, module name/ code, etc. This information is stored in the document header (cf. section 3 and 4).

The structure of the BAWE corpus as outlined above is meant to capture variety, both in terms of cross-departmental practices, assignment types and individual style. At the same time, the material represents homogeneity in that similar assignments at the same level are included, both within a given department and across different departments. (For a discussion of the BAWE corpus structure and sampling strategy, see Nesi et al., 2005.)

As pointed out by Biber et al. (1998: 177) ‘representativeness in corpus design is a crucial concern for all corpus-based studies, particularly in seeking a comprehensive description of language variation’. The BAWE corpus seeks to contain a representative collection of proficient student writing as produced by university students in Britain in the early 21st Century.

Having outlined the overall corpus composition, we shall now turn to the structure of individual corpus items, addressing the issue of representing the information contained in them. Specific transformations of the document as well as the question of an overall document model will be discussed.

3. ENCODING

Concerning the encoding of the BAWE corpus, a decision was made to apply the encoding standard proposed by the *Text Encoding Initiative* (TEI). This has implications on three levels:

First, the corpus files are to be encoded in plain text format. In this format, text is represented as a plain and linear sequence of characters. The encoding chosen for the BAWE corpus is

Unicode (UTF-8); however, an ASCII encoded version will also be available. Second, BAWE corpus files are also XML files, thus allowing a neat and unequivocal separation between text and meta-text (mark-up, annotation). Finally, the TEI standard proposes one particular specification of a document model (Sperberg-McQueen & Burnard, 2004), formally explicated in the TEI *Document Type Definition* (DTD). Since it is the ambition of TEI to provide markup for any feature of any type of (spoken or written) text, only a small fraction of the tags available is actually used in the BAWE encoding. The various features encoded will be discussed in the following; a list of the TEI elements used is given in the appendix.

Apart from providing a range of tags for encoding features of text, TEI imposes a strict hierarchical organisation of the text. Thus, the TEI conformant file contains ‘text’ as opposed to the ‘header’; the document header is a convenient place for storing ‘basic information about the nature of the text’ (McEnery & Wilson, 1996: 30; cf. next section). Within the text, a distinction is made between ‘front’, ‘body’ and ‘back’; the body, as the main part, grossly corresponding to the document’s *running text* (which is only a part of what the full source text contains).

One source of complication stems from the fact that the incoming ‘raw’ corpus documents are usually in Microsoft Word format (doc, rtf etc.). The necessary transformations raise not only practical, but also fundamental conceptual issues.

4. TRANSFORMING WORD DOCUMENTS TO ANNOTATED XML FILES

We noted above the contrast between the input and output format of our corpus documents, which are normally submitted as Word files (a *de facto* text processing standard) and will be represented in XML format in the end. Due to the complexity of information contained in a Word document, its conversion to XML involves not only adding tags to the text, but recoding the entire document. This is not a merely technical issue – on the contrary, technically speaking, the matter is quite simple, as Word (since MS Office 2003) itself allows saving files in XML format. In particular, we must take into account the fact that part of the information contained in a Word document is related to on-screen display and thus not necessarily encoded explicitly at all, which makes it necessary to consider and evaluate the information contained in the source document before deciding whether and how it should be represented in our final corpus files.

4.1. Information and formats of encoding

Information contained in or associated with a written document can be of different kinds: on the one hand, *textual information* is represented by the (usually) linear sequence of characters that constitute the document text; on the other hand, *metatextual information*, i.e. information about text, may intend to capture properties of either a specific chunk of text or the text as a whole.

Thus, we refer to *metadata* as 'the kind of data that is needed to describe a digital resource in sufficient detail and with sufficient accuracy for some agent to determine whether or not that digital resource is of relevance to a particular enquiry' (Burnard, 2005: 30) and allows the text to be situated within some text typology (Atkins et al., 1992: 6-9). Typically, this kind of data appears as 'header information' (cf. sections 2 and 3).

Information related to delimited chunks is meant to highlight some property of this chunk; it is worth noting that the nature of its content is basically unrestricted. We may distinguish linguistic features (cf. Leech, 1997 & 2005), 'underlying structural features' (e.g. quotations, lists, headings, front/ back matter etc.) and typographical features (Atkins et al., 1992: 9). In XML, metatextual information related to a chunk of text is encoded as *mark-up* (tags). As for the *underlying structural features*, information of this kind points to the functioning of the chunk to which it relates. The fact that they are essentially grounded on formatting and lay-out in Word documents accounts for the particularities of their encoding:

- one piece of metatextual information can be encoded as a bundle of formatting/ lay-out features;

- the same formatting and lay-out properties can be used for encoding different metatextual properties

- retrieving some types of metatextual information may involve not only consideration of the item in question, but also of other text (unmarked text as well as text bearing the same property);

- up to a certain degree, the same information can be encoded in different ways; as a result, we find a widespread variation in the encoding of features not only between texts, but even within individual texts (polymorphism and inconsistencies).

This points to a fundamental difference between Word and XML encoding: a Word document consists of text and layout, whereas an XML document is encoded through text and metatext (elements, attributes). The structure of an XML document is thus explicit; in the case of a Word document, not only does the structure need to be inferred, but the structural dimensions themselves have to be added by the reader (through experience, knowledge of encoding conventions, particular interest etc.). The status of any chunk within an XML document is fully defined through occurrence within its parent element (e.g., the ‘title’ is the text contained by a title element); in Word encoding, it has to be inferred from formatting and layout information (e.g. the title is what appears on one [more] separate line[s], usually in a salient font, at the beginning of the document etc.). Inferring functions from Word formatting properties is possible because these properties are not void of meaning: bold, italic, underlining, larger font are understood as highlighting, chunks may be separated out by line breaks and white space etc. The knowledge upon which our interpretations draw has been conveyed through a long tradition of printed documents. XML tags, on the contrary, do not by themselves convey any text-external meaning; an element is defined distributionally in terms of where it may occur and what it may contain – any further ‘meaning’ is beyond the scope (and the intention) of the XML formalism.

Styles in Word are an attempt to introduce explicit coding of conceptual functions, but, representing a short-cut for a bundle of formatting features (font, paragraph style), they are themselves features of the typographic surface: even though ‘heading’ styles are available, it can hardly be argued that they are what ‘defines’ a heading in a Word document. Styles are mandatory, non-defining properties; they can be omitted or even applied ‘wrongly’ (which is frequently the case where they are attributed by Word automatically). Neither of this is true in XML encoding.

Consequently, XML format automatically produced by Word does not (and cannot) reflect the document's conceptual structure, but formatting and layout only. Retrieving a Word document's structural and functional properties by its very nature involves reading and interpreting the document surface by human readers, referring to their knowledge of standards, norms and traditions of encoding.

For the information encoded, this has important consequences: whereas an XML tag, by definition, has a very precise denotation, Word formatting and lay-out information invites interpretation on multiple semiotic levels – on a literal level to begin with (e.g. bold character formatting meaning 'bold'), but further levels of interpretation can be activated by specific contextual configurations (e.g. bold characters on a separate line may come to mean 'section title'). Such *inferred information*, attributing a meaning to features of formatting and lay-out, is functional in nature and thus opposed to the strictly explicit formalism that we find in XML encoding.

4.2. The task of annotation

In principle, of course, this opposition could be dismissed as immaterial, and one may question its relevance for the process of encoding and re-coding at all. Since formatting and lay-out information of a Word document is objective, it could as such be formalised easily. However, one may hold that such representation of document information would hardly be adequate. It seems a plausible assumption that this is neither how writers conceive their document, nor how the 'competent' reader will (or: is meant to) process it, and for this reason the analyst may choose to identify chunks of the text in the same way as they are perceived by the users, viz. as carrying functional information.

However, as it is clear that such functional encoding can be based only on a *meaningful reading, i.e. interpretation* of the document and its features in question, two limitations have to be imposed. First, interpretation should be limited to cases and phenomena that are reasonably ‘objective’, i.e. about which competent readers largely agree. Though borderline cases and uncertainties will inevitably occur, the functional judgement should be reliable in its principle and thus, *de facto* present a theory-neutral stance. Second, it should be limited to cases where it is necessary, i.e. where a ‘literal’ (formalistic) reading of the features would be a gross and evident misrepresentation of the text (this implies the necessity to tag as equivalent features those which share the same function, but on different ‘formal’ grounds). The latter is a judgement of *relevance* that has to be made in view of the purpose and intended uses of the corpus under construction.

The task of re-coding Word as XML, as is typical for most annotation tasks, thus consists in *explicating information* that is supposed to be implicitly present in the raw text (McEnery & Wilson, 1996: 24). Understanding by a *competent human reader* is indispensable for this. In the BAWE corpus, the following functions are annotated[2]:

- *Document title* – for the sake of convenience, we consider other elements before the beginning of running text or the first section heading, if there is not a special label for them, as ‘title parts’ (such as: author name, date, title and code of the module for which the assignment was written etc.).

- *Table of contents*

- *Sections and subsections* – keeping track of sections and subsections, signalled through section headings, gives access to the document's explicit hierarchy of chunks of text.

- *Lists* – can, in their prototypical form of an enumeration of items, be considered as not being part of running text (both from a syntactic point of view and from the perspective of thematic development) and should therefore be distinguished from 'normal' text. On the other hand, paragraphs of running text may carry the signals of list items (numbering, bullet points etc.). Although the exact boundary is somewhat arbitrary and not always easy to determine, their resemblance in terms of formatting must not be given too much weight, and the latter should not be considered as 'lists'.

- *Block quotes* – may be in English or another language (whereas our corpus texts are in English only); they are usually rather long passages, for which the author is not to be held responsible. Since they can be identified easily, it was decided that they should be marked as separate from normal running text.

- *Formulae/ formal expressions* – include what is habitually referred to as *formulae* (i.e. mathematical, physical, chemical etc.), but also chunks of text encoded in some non-natural language formalism (e.g. computer code, phonetic transcription etc.).

- *Abstract or summary* (if it is labelled as such)

- *Bibliography/ references section* – is normally considered as back matter (i.e. not part of running text) and therefore needs special marking up.

- *Appendix* – the content of which, as a rule, is unpredictable; it may consist of non-linguistic material (illustrations, graphs, tables etc.), but even where it contains

> language, it cannot be assumed by default that appendix texts are written by the author of the assignment. It therefore seemed reasonable to exclude appendices from linguistic analysis of student writing.

The markup strategies relating to items carrying these functions differ: parts of the document title, lists and block quotes are fully wrapped between opening and closing tags. Features based on headings (sections, abstract, bibliography, appendix) are annotated by placing opening and closing tags around the heading only, which simplifies the task; the chunks of text appearing between these headings will be attributed to the section to which they belong later, in a step of automatic transformation. Tables, figures and table of contents are deleted and tags placed around their captions; finally, some items are completely removed and replaced with (empty) tags: tables and figures without captions and formal expressions.

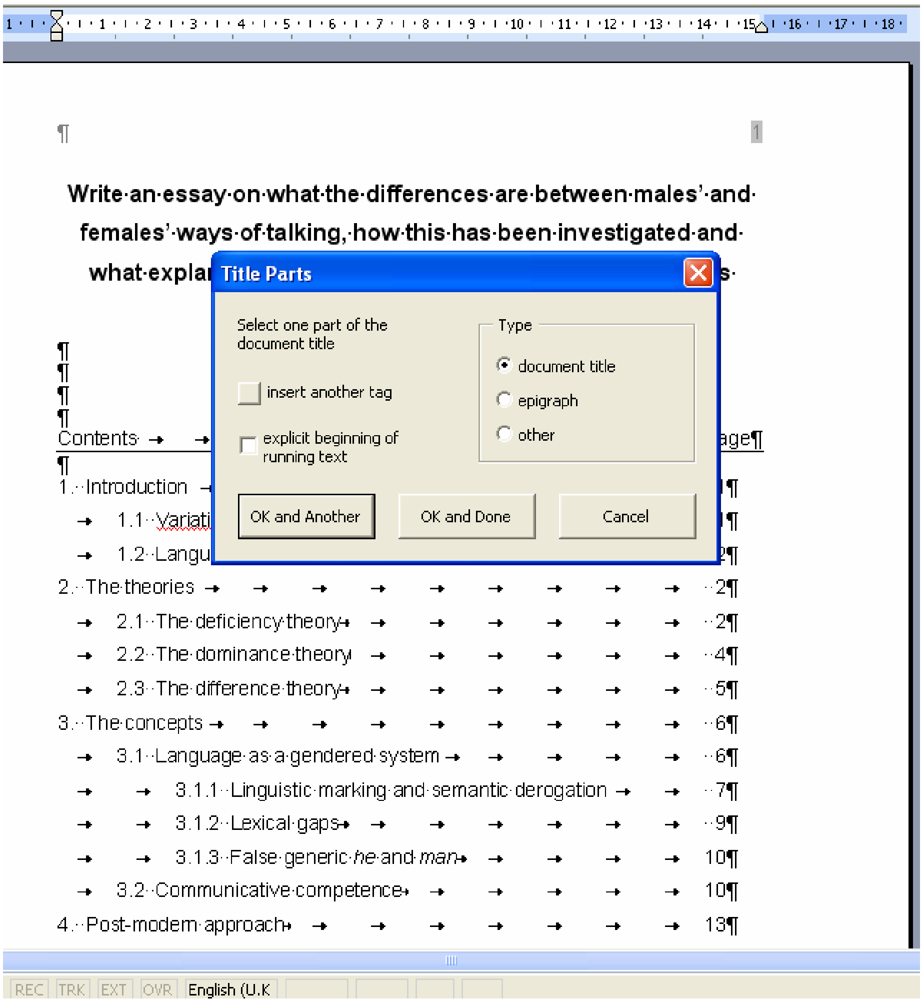
The question arises as to how far these features can be detected and annotated automatically, considering that Word provides built-in possibilities for realising many of them through specific lay-out devices: e.g., title styles may be applied to section headings, lists may be marked by applying automatic bulleting or enumeration styles, a function is available for inserting a table of contents etc. Even though it would be possible to detect occurrences of such features automatically, there is no stable relation between these features and the intended functional values, to the effect that the task of annotation cannot be automated: heading styles need not be applied by the student writer and may, in fact, be applied wrongly, enumerations and listing symbols can simply be typed, a table of contents can be written by hand, and formatting inconsistencies inevitably will occur. Intervention of a human tagger is therefore indispensable for recognising the functions in question.

Apart from these lay-out based functions, annotation of linguistic properties (lemmatisation, part of speech and, possibly, other linguistic dimensions), which is often regarded as prototypical for corpus annotation as such (Leech, 1997: 2 & 2005: 17), will be added to the text in the course of further processing; they will not be of concern for our discussion here.

4.3. Organising the task of annotation: an interface for tagging Word documents

In concluding this section, some consideration has to be given to practical issues involved in the task of explicitly encoding functional information (cf. issues of interactive higher-level annotation discussed in Garside & Rayson, 1997: 179-187). Difficulties arising are of two types: since human interpretation takes up such a central part, reliability and consistency of annotation becomes an issue. Careful documentation, communication between human taggers and occasional second tagging are strategies applied to enhance the reliability of annotation by revealing systematic inconsistencies between taggers. Another source for errors is poor concentration and fatigue; since systematically checking annotated files and correcting errors is beyond the possibilities of the BAWE project, it was found advisable to limit the potential for making errors by reducing, as much as possible, the complexity of the annotation task. An interface for interactive manual annotation was therefore developed in the form of a series of Word macros, written in Visual Basic and making use of *graphical user interface* possibilities (termed ‘user forms’).

This interface has been set up to guide the tagger through the annotation process step by step: at the beginning of the tagging process, the tagger indicates the functions they would like to annotate (e.g. document title, sections, bibliography, lists etc.). Following this, the features chosen are annotated one by one; it is recommended that all the occurrences of the same feature be tagged together, although it is also possible to temporarily ‘switch’ to annotating another feature. For each feature, a customised tagging box becomes available providing further options (e.g. ‘main’ vs. ‘epigraph’ vs. ‘other’ for document title parts, the level of embedding for section headings, ‘ordered’ vs. ‘bulleted’ vs. ‘simple’ for lists etc.). The tags are inserted by clicking an ‘OK’ button.



1

Write an essay on what the differences are between males’ and females’ ways of talking, how this has been investigated and what explai s

Contents Page

1. Introduction
1.1 Variati
1.2 Langu 2
2. The theories 2
2.1 The deficiency theory 2
2.2 The dominance theory 4
2.3 The difference theory 5
3. The concepts 6
3.1 Language as a gendered system 6
3.1.1 Linguistic marking and semantic derogation 7
3.1.2 Lexical gaps 9
3.1.3 False generic *he* and *man* 10
3.2 Communicative competence 10
4. Post-modern approach 13

Title Parts

Select one part of the document title

□ insert another tag

□ explicit beginning of running text

Type
● document title
○ epigraph
○ other

Figure 1: Graphical interface for tagging parts of the document title. The major options available are: ‘document title’, ‘epigraph’ or ‘other’ title part (e.g. author [anonymised], module for which the assignment was written etc.)

This interface has been designed to facilitate the human tagger's task in various respects:

- Operating within Word, the human tagger still has the original formatting available during the tagging process. Interpreting formatted text involves considerably less effort than interpreting unformatted text.

- Tags can be selected from options, thus avoiding any typing. The options appear as checkboxes, radio buttons, drop down lists or labelled keys.

- By organising the tagging process in two layers, i.e. first choosing between functions available and then annotating these functions, the tagging interface, changing throughout the process, is always focused on the function being annotated. The tagger only chooses between *relevant* options for this function.

- Thus, the tagging interface is tailored to the requirements of the BAWE corpus[3]: both layers of annotation, functions and specific options describing their realisation, are designed to direct and limit the annotator's choice to the functional categories outlined above.

The tags created are inserted as particular codes of character sequences within the text of the Word document. It should be noted that since Word does not allow to make a distinction between text and meta-text, there is no possibility to insert genuine XML tags at this point.

Any sequence of characters would be subject to transformation in the process of recoding the Word document as XML file; in particular, the tag delimiters '<' and '>' would not be recognised as metacharacters. It was thus decided to insert 'pseudo-tags' at this stage.

Only after the document has been converted to XML format (using the function 'save as XML', available since MS Word 2003), these 'pseudo-tags' are transformed to genuine XML tags by a Perl script. Likewise, a cascade of Perl scripts is used to achieve final transformations: normalisation of hyphens, dashes and 'smart quotes', transformation of Microsoft XML input to a TEI-conformant XML tree, importation of contextual information ('metadata') from external spreadsheets, mark-up of sentence boundaries following simple heuristics based on strong punctuation signs and numbering to paragraphs and sentences. In a final step, the resulting document is checked for validity against the TEI DTD by a parser.

5. CONCLUDING REMARKS

In this article, we outlined the overall approach to sampling adopted in the BAWE corpus of student writing and discussed issues related to its encoding and mark-up. Approximately 3,500 undergraduate and postgraduate student assignments are to be collected in a variety of disciplines across the four broad disciplinary groupings *Arts & Humanities, Life Sciences, Physical Sciences* and *Social Sciences.*

Assignments are submitted by students as documents in Microsoft Word format and need to be transformed into XML encoded corpus files. We pointed out that preprocessing the Word document before conversion to XML format is an essential step in order not to lose relevant information.

In particular, attention was drawn to the two fundamentally different ways in which Word and XML formats encode document information: functions related to chunks of text, implicitly signalled in Word documents through formatting and lay-out, have to be rendered explicit in XML. It was argued that interactive annotation of such features in the Word document is an indispensable step before changing the document encoding to the plain-text format XML. The interface developed for interactive annotation was briefly presented.

A full exploration of the data in the BAWE corpus will only be conducted after the completion of the corpus and in the final stages of the project. The investigation will be centred on characteristics of proficient student writing and how these characteristics can be categorised into genres and sub-genres (see also http://www.warwick.ac.uk/go/BAWE/overview). An example of functional genre-related studies is Gardner et al.'s (2005) use of a subset of the pilot corpus for a preliminary exploration of ideational meaning in student writing. The results of a multidimensional analysis of the data, to be carried out by Douglas Biber and his colleagues at Northern Arizona University (see e.g. Biber, 1988), will be available as a further source of information.

At the moment of writing (May 2006) the three-year long BAWE project has just entered its second year. The first year of the project was mainly devoted to the preparation stage of the collection process and to the development and testing of software that will facilitate the text processing.

Although planning and testing were the two most prominent issues on the agenda, some effort was also put in to the assignment collection itself, and approximately 900 student assignments were collected during the course of the first year. It is our hope that the full quota of assignments will be ready for inclusion in the BAWE corpus some time during 2007.

REFERENCES

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

Biber, D. 1988. Variation across speech and writing. Cambridge: Cambridge University Press.

Biber, D., Conrad, S. and Reppen, R. 1998. Corpus linguistics. Investigating language structure in use. Cambridge: Cambridge University Press.

Burnard, L. 2000. Reference guide for the British National Corpus (world edition). Edited by Lou Burnard. Published for the British National Corpus Consortium by the Humanities Computing Unit at Oxford University Computing Services. October 2000. Available online: http://www.natcorp.ox.ac.uk/docs/userManual/ [accessed 26/04/2006].

Burnard, L. 2005. ‘Metadata for corpus work’ in M. Wynne (ed.), Developing linguistic corpora: a guide to good practice, pp 30-46. Oxford: Oxbow Books. Available online from http://ahds.ac.uk/linguistic-corpora/ [accessed 2006-03-28].

Gardner, S., with Heuboeck, A. and Nesi, H. 2005. ‘Ideational meaning in a corpus of student academic writing’ [paper presented at The 17th European Systemic-Functional Linguistics Conference & Workshop, 1st-4th August, 2005, King’s College London, UK].

Garside, R. and Rayson, P. 1997. ‘Higher-level annotation tools’ in R. Garside, G. Leech & A. McEnery (eds.), Corpus annotation. Linguistic information from computer text corpora, pp 179-193. London: Longman.

Leech, G. 1997. ‘Introducing corpus annotation’ in R. Garside, G. Leech & A. McEnery (eds.), Corpus annotation. Linguistic information from computer text corpora, pp 1-18. London: Longman.

Leech, G. 2005. ‘Adding linguistic annotation’ in M. Wynne (ed.), Developing linguistic corpora: a guide to good practice, pp 17-29. Oxford: Oxbow Books. Available online from http://ahds.ac.uk/linguistic-corpora/ [accessed 2006-04-24].

McEnery, T. and Wilson, A. 1996. Corpus linguistics. Edinburgh: Edinburgh University Press (Edinburgh textbooks in empirical linguistics).

Nesi, H., Sharpling, G. and Ganobcsik-William, L. 2004. ‘Student papers across the curriculum: Designing and developing a corpus of British student writing’, Computers and Composition 21 (4), pp 439-450.

Nesi, H., Gardner, S., Forsyth, R., Hindle, D., Wickens, P., Ebeling, S., Leedham, M., Thompson, P. and Heuboeck, A. 2005. ‘Towards the compilation of a corpus of assessed student writing: An account of work in progress’ in P. Danielsson & M. Wagenmakers (eds.) Corpus Linguistics Vol. 1, no. 1. Available online from http://www.corpus.bham.ac.uk/PCLC/

Nesi, H., and Gardner, S. 2006. ‘Variation in disciplinary culture: University tutors’ views on assessed writing tasks’, in R. Kiely, P. Rea-Dickins, H. Woodfield & G. Clibbon (eds.), Language, Culture and Identity in Applied Linguistics. British Studies in Applied Linguistics Vol. 21. London: Equinox Publishing, pp 99-117.

Sperberg-McQueen, C. M. and Burnard, L. (eds.) 2004. TEI P4 – Guidelines for Electronic Text Encoding and Interchange, XML-compatible edition. Available online from http://www.tei-c.org/P4X/

APPENDIX

The subset of TEI used in the BAWE corpus (for a definition of the TEI elements see http://www.tei-c.org/P4X/REFTAG.html):

```
<back>, <body>, <cell>, <distributor>, <div1>, <div2>, <div3>, <div4>, <div5>, <div6>, <div7>, <docTitle>, <encodingDesc>, <extent>, <figure>, <fileDesc>, <formula>, <front>, <head>, <hi>, <item>, <list>, <name>, <note>, <notesStmt>, <p>, <particDesc>, <person>, <profileDesc>, <publicationStmt>, <quote>, <row>, <s>, <sourceDesc>, <table>, <TEI.2>, <teiHeader>, <text>, <title>, <titlePage>, <titlePart>, <titleStmt>
```

Example of a BAWE text fragment encoded in TEI:

```
<div1 type="section">

<head rend="bold italic">Test data:</head>

<p n="p20.25">

<s n="s1.1;p20.25">In this following printout, we can see several tests in several cases. </s>

</p>
```

<p n="p21.25">

<s n="s1.1;p21.25">

<formula notation="" id="BAWE_3052a-form.005"/> </s>

</p>

<p n="p22.25">

<s rend="italic" n="s1.1;p22.25">In this example several command line are entered until the user has entered 'exit'. </s>

</p>

<list type="ordered">

<item rend="italic">'hello' needn't argument and printouts a simple message</item>

<item rend="italic">'helloarg' is a simple program which printouts all the arguments</item>

<item rend="italic">'blabla' doesn't exist => child error => parent informs user</item>

<item rend="italic">Some tests follow using the program 'ls' (/bin/ls)</item>

</list>

<p n="p23.25">

<s n="s1.1;p23.25">

<formula notation="" id="BAWE_3052a-form.006"/> </s>

</p>

```
</div1>
```

[1] The project *An investigation of genres of assessed writing in British Higher Education*, which includes development of the *British Academic Written English* corpus, is funded by the Economic and Social Research Council (project number RES-000-23-0800).

[2] For functional information annotated in the – equally TEI conformant – BNC see Burnard, 2000.

[3] The options and labels used are 'hard coded' in the macros, which restricts the potential of their reusability to documents which are structurally similar to the ones contained in the BAWE corpus; adapting the interface to other needs would imply changing the source code. With this *caveat*, the source code can be made available to other users on an as-is basis (please contact the authors).