The BAWE Corpus Manual

FOR THE PROJECT ENTITLED

'An Investigation of Genres of Assessed Writing in British Higher Education'.

funded by the ESRC [project number RES-000-23-0800**]**

Alois Heuboeck: a.heuboeck@reading.ac.uk
Jasper Holmes: jasper.holmes@coventry.ac.uk
Hilary Nesi: h.nesi@coventry.ac.uk

Table of contents

1. Introduction

The research project *An Investigation of Genres of Assessed Writing in British Higher Education*, undertaken in the period 2004-2007 at the universities of Warwick, Reading and Oxford Brookes, was funded by the ESRC (project number RES-000-23-0800).

The British Academic Written English (BAWE) corpus, resulting from this project, is available in three formats: a collection of XML files (containing full markup), a TXT version containing only a minimal number of tags and a PDF version that represents the original documents. The main part of this document describes the document structure and tags used for markup of the XML version of the BAWE corpus. In this version, one XML file corresponds to one assignment. The features of the TXT version will be presented in section 12.

Online versions of the CORPUS are available through the BAWE corpus search interface at Coventry University (http://www.coventry.ac.uk/bawe) and through the Sketch Engine corpus query interface (http://www.sketchengine.co.uk/). Some of the markup has been modified for the version on the Sketch Engine, and there is a separate document describing query options made available through markup in that version.

Corpus files are available from the Oxford Text Archive (http://ota.ahds.ac.uk/).

2. Collection

The collection process and decisions made during collection are described in detail in Alsop and Nesi (under review); a summary description of the most significant points is given here.

Student assignments were collected from three universities: Oxford Brookes, Reading and Warwick. Assignments were collected from 35 disciplines (see Table 2), in 4 broad disciplinary groupings (DGs: see Table 1), and from students in each of three undergraduate years and those on masters courses.

Table 1 shows the numbers of students, assignments, texts and words collected in each year in each DG. The difference in the numbers of assignments and texts is

due to the fact that some assignments (**compound assignments**, see 6.3) consist of more than one text.

Table 1. Numbers of students, assignments, texts and words by disciplinary grouping and year

disciplinary group		Yr 1	Yr 2	Yr 3	Masters	Total
Arts and Humanities	students	101	83	61	23	268
	assignments	239	228	160	78	705
	texts	259	231	161	83	734
	words	468,353	583,617	427,942	234,206	1,714,118
Life Sciences	students	74	71	42	46	233
	assignments	180	193	113	197	683
	texts	191	208	119	203	721
	words	299,370	408,070	263,668	441,283	1,412,391
Physical Sciences	students	73	60	56	36	225
	assignments	181	149	156	110	596
	texts	186	156	169	129	640
	words	300,989	314,331	426,431	339,605	1,381,356
Social Sciences	students	85	88	75	62	313[1]
	assignments	207	197	162	202	777[2]
	texts	218	202	169	204	802
	words	371,473	475,668	440,674	688,921	1,999,130[4]
Total students		**333**	**302**	**234**	**167**	**1039[1]**
Total assignments		**807**	**767**	**591**	**6587**	**2761[2]**
Total texts		**854**	**797**	**618**	**619**	**2897[3]**
Total words		**1,440,185**	**1,781,686**	**1,558,715**	**1,704,015**	**6,506,995[4]**

[1] Includes 3 of unknown level.

[2] Includes 9 of unknown level.

[3] Includes 9 of unknown level.

[4] Includes 22,394 in texts of unknown level.

Table 2 shows the number of assignments by discipline and year.

Table 2. Number of assignments by discipline and year

disciplinary group	discipline	1	2	3	4	Total
Arts and Humanities	Archaeology	23	21	15	17	76
	Classics	33	27	15	7	82
	Comparative American Studies	29	26	13	6	74
	English	35	35	28	8	106
	History	30	32	31	3	96
	Linguistics	27	31	24	33	115
	Other	19	22	9	0	50
	Philosophy	43	34	25	4	106

	Total	239	228	160	78	705
Life Sciences	Agriculture	35	35	30	34	134
	Biological Sciences	52	50	26	41	169
	Food Sciences	26	36	32	30	124
	Health	35	33	12	1	81
	Medicine	0	0	0	80	80
	Psychology	32	39	13	11	95
	Total	180	193	113	197	683
Physical Sciences	Architecture	2	4	2	1	9
	Chemistry	23	24	29	13	89
	Computer Science	34	13	30	10	87
	Cybernetics & Electronics	4	4	13	7	28
	Engineering	59	71	54	54	238
	Mathematics	8	5	12	8	33
	Meteorology	6	9	0	14	29
	Other	0	1	0	0	1
	Physics	37	14	14	3	68
	Planning	8	4	2	0	14
	Total	181	149	156	110	596
Social Sciences	Anthropology	14	12	6	17	49
	Business	32	33	31	50	146
	Economics	30	30	23	13	96
	HLTM	14	21	29	29	93
	Law	37	37	31	28	134*
	Other	0	2	3	4	9
	Politics	37	33	15	25	110
	Publishing	11	4	0	15	30
	Sociology	32	25	24	21	110†
	Total	207	197	162	202	777‡
Total		807	767	591	587	2761‡

* Includes 1 of unknown year.

† Includes 8 of unknown year.

‡ Includes 9 of unknown year.

Students were paid £3 for each assignment submitted (towards the end of the collection period students in some underrepresented disciplines were rewarded at £5 and even £10 per assignment) and were asked to sign disclaimer forms assigning copyright to the respective university (disclaimer forms are given in Appendix 1). Contextual information about the student and the assignment were also collected at this stage. Students were asked to supply all the contextual information to be contained in the XML file headers (see below).

After all assignments were collected, a number of analyses were carried out, including a Multi-Dimensional Analysis (Conrad and Biber 2001) of the registers of

the texts, to be described in future publications, and a genre analysis described in the following section.

3. Genre and genre family

All assignments in the corpus were scrutinised for generic properties, and a large number of genres were identified. These were collected into 13 **genre families** (GFs), classes of genres sharing functional and structural properties. The full set of genres and GFs is given in Appendix 2. A list of the GFs, and their distribution by DG, is shown in Table 3.[1] The distribution of GFs by discipline is shown in the following tables.

Table 3. Distribution of GFs by DG

	Arts and Humanities	Life Sciences	Physical Sciences	Social Sciences	Total
case study	0	91	37	66	194
critique	48	84	76	114	322
design specification	1	2	87	3	93
empathy writing	4	19	9	3	35
essay	602	127	65	444	1238
exercise	14	33	49	18	114
explanation	9	117	65	23	214
literature review	7	14	4	10	35
methodology recount	18	158	170	16	362
narrative recount	10	25	21	19	75
problem question	0	2	6	32	40
proposal	2	26	19	29	76
research report	9	22	16	14	61
Total	724	720	624	791	2859

[1] These are **texts**, rather than **assignments**: during genre analysis, genres were assigned to the parts of compound texts. The total in this table unfortunately differs from that in Table 1. There is clearly a mismatch between the set of assignments coded as compound by taggers and those analysed as compound during genre analysis.

Table 4. Distribution of GFs by discipline, Arts and Humanities disciplines

AH	Arch	Classics	CAS	English	History	Ling	Other	Phil	Total
case study	0	0	0	0	0	0	0	0	0
critique	15	2	2	1	1	21	0	6	48
design specification	1	0	0	0	0	0	0	0	1
empathy writing	0	0	0	4	0	0	0	0	4
essay	49	78	71	89	94	75	48	98	602
exercise	1	0	0	7	1	5	0	0	14
explanation	2	0	0	0	0	7	0	0	9
literature review	0	1	0	5	0	0	0	1	7
methodology recount	7	0	0	0	0	9	2	0	18
narrative recount	1	0	0	4	0	5	0	0	10
problem question	0	0	0	0	0	0	0	0	0
proposal	0	1	0	1	0	0	0	0	2
research report	1	1	1	0	0	5	0	1	9
Total	77	83	74	111	96	127	50	106	724

Table 5. Distribution of GFs by discipline, Life Sciences disciplines

LS	Agriculture	BioSci	FoodSci	Health	Med	Psych	Total
case study	12	0	2	8	69	0	91
critique	37	20	9	9	1	8	84
design specification	1	0	0	0	0	1	2
empathy writing	2	0	13	1	1	2	19
essay	27	11	7	15	10	57	127
exercise	7	7	18	0	0	1	33
explanation	30	63	7	13	1	3	117
literature review	4	3	4	3	0	0	14
methodology recount	7	58	82	1	0	10	158

narrative recount	1	2	0	20	2	0	25
problem question	0	0	1	1	0	0	2
proposal	6	3	3	13	0	1	26
research report	1	8	0	1	0	12	22
Total	135	175	146	85	84	95	720

Table 6. Distribution of GFs by discipline, Physical Sciences disciplines

PS	Arch	Chem	CS	C&E	Eng	Math	Met	Phys	Plan	Total
case study	0	2	2	0	33	0	0	0	0	37
critique	1	11	8	3	31	6	4	10	2	76
design specification	0	0	41	13	28	0	3	2	0	87
empathy writing	1	0	2	0	1	4	0	1	0	9
essay	4	6	9	2	16	4	0	12	12	65
exercise	1	6	8	4	10	15	4	1	0	49
explanation	0	10	16	0	16	3	5	15	0	65
literature review	0	0	1	2	1	0	0	0	0	4
methodology recount	0	51	3	1	83	0	14	18	0	170
narrative recount	1	2	2	1	12	1	0	1	0	21[1]
problem question	0	0	0	0	6	0	0	0	0	6
proposal	1	0	6	0	10	1	0	1	0	19
research report	0	1	2	2	4	0	0	7	0	16
Total	9	89	100	28	251	34	30	68	14	624[2]

[1] Includes 1 from another discipline

[2] Includes 1 from another discipline

Table 7. Distribution of GFs by discipline, Social Sciences disciplines

SS	Anth	Bus	Econ	HLTM	Law	Pol	Publ	Soc	Total
case study	0	31	1	27	1	0	5	1	66

critique	8	29	17	11	25	11	2	11	114
design specification	0	2	0	0	0	0	0	0	3
empathy writing	0	0	0	1	0	0	2	0	3[1]
essay	27	49	55	29	85	97	4	91	444
exercise	0	12	5	0	0	0	1	0	18
explanation	3	5	0	5	0	0	10	0	23
literature review	4	0	0	2	1	2	1	0	10
methodology recount	2	2	10	1	0	0	0	1	16[2]
narrative recount	2	4	0	5	0	0	5	2	19
problem question	0	9	2	2	19	0	0	0	32
proposal	3	2	0	12	3	0	7	2	29[3]
research report	0	1	7	3	0	0	0	2	14
Total	49	146	97	98	134	110	38	110	791[4]

[1] Includes 7 from other disciplines.

[2] Includes 1 from another discipline.

[3] Includes 1 from another discipline.

[4] Includes 9 from other disciplines.

4. Markup

Markup is used in the BAWE corpus for encoding information of the following types:
document structure and hierarchy
header information
types of front and back matter
functional features within running text
features of highlighting (character formatting)
anonymized personal information (related to student, university or third parties)

The markup of the BAWE corpus follows the guidelines of TEI P4 (Sperberg-McQueen and Burnard 2004). Since the TEI standard has been devised for a wide range of texts, a special DTD containing only a subset of all TEI elements and

attributes has been created for BAWE using the online tool TEI PizzaChef (http://www.tei-c.org.uk/pizza.html). The resulting DTD is contained in the file tei_bawe.dtd, which is distributed with the corpus files.

The following sections will discuss:

character encoding used for BAWE corpus files

the tags used for annotation (customization of TEI P4)

their attributes and (possible) values

criteria for identifying the phenomenon in question

Notation:

Throughout this document, elements (tags) used in the BAWE corpus are signalled through their formatting as **`elementname`**, attributes of elements as *`attributename="value"`*. A parent-child relation is represented with a vertical line as: **`parentElement|childElement`**; alternative values for an attribute are also separated by a vertical line: *`attribute="value1|value2"`*.

5. Character encoding

The BAWE corpus is encoded in UTF-8 Unicode. An ASCII version is also available upon request; in the ASCII version, non-ASCII characters are encoded using an empty **`seg`** element:

- as **`seg`** *`n="entityName"`* for characters defined as entity in the DTD;
- as **`seg`** *`n="#x[hex]"`* for all other characters – where [hex] represents the hexadecimal number of the character in the Unicode character set (cf. http://www.unicode.org/Public/UNIDATA/UnicodeData.txt).

5.1. Normalization of characters

Some functionally equivalent characters may appear in different forms and have been normalized:

- smart single quotes are replaced by straight single quotes;
- smart double quotes are replaced by straight double quotes;
- em-dash is replaced by a simple dash preceded and followed by a space character;

- en-dash is replaced by a simple dash preceded and followed by a space character;
- horizontal ellipsis character (…) is replaced by three full stop characters;
- non-breaking space characters are replaced by simple space characters.

6. Structure of BAWE documents

The BAWE corpus consists of a number of XML files, each of them containing *one assignment*. Each file is TEI-conformant, but there are no explicit links between them. This document will describe the structure of individual BAWE corpus files.

- The **root of the document** is **`TEI.2`**; this element contains one assignment. **`TEI.2`** has the attribute *`id`*, whose value is the specific ID number identifying the assignment in the BAWE corpus. The ID value consists of four digits, identifying the author of the assignment, and one lower case letter, though which different assignments provided by the same author are distinguished. Together with the extension ".xml", the ID value constitutes the name of the corpus file. – Due to a restriction imposed by XML, the value of the attribute **`TEI.2`** *`id`* is preceded by an underscore.
- TEI.2 contains two elements: **`teiHeader`** and **`text`**. **`text`** contains the text as it was submitted (including any amendments made during the manual tagging process); **`teiHeader`** contains meta-information, related either to the assignment or its author.

6.1. The document header

An example of a header is provided in appendix 1.

a)Contextual information

The contextual information is related either to the assignment or to the author.

Assignment-related contextual information is found in **`teiHeader|fileDesc|sourceDesc`**, each piece of contextual information appearing in a **`p`**. The type of contextual information is identified in **`p`** *`n`*:

Table 8. Assignment-related contextual information

p *n=*	**possible values (content of p)**
level	level of studies: '1', '2', '3', '4' (1st-4th year UG), 'PG', Phase I, Phase II (two stages of postgraduate medical course)
date	date of writing: standardized as yyyy-mm
module title	not standardized, string as given by student (or retrieved from university's module directory)
module code	not standardized, string
genre family	'case study', 'critique', 'design specification', 'empathy writing', 'essay', 'exercise', 'explanation', 'literature survey', 'methodology recount', 'narrative recount', 'problem question', 'proposal', 'research report' - assigned by research team (see 3 above)
discipline	(AH): 'Linguistics', 'Archaeology', 'Classics', 'Comparative American Studies', 'History', 'Philosophy', (SS): 'Business', 'Economics', 'Law', ' Politics', ' Sociology', (LS): 'Agriculture', 'Biological Sciences', 'Food Sciences', 'Medicine', 'Psychology', (PS): 'Chemistry', 'Computer Science', 'Cybernetics & Electronic Engineering', 'Engineering', 'Physics', 'Mathematics', 'Meteorology', and 'other'
disciplinary group	'AH' (Arts and Humanities), 'SS' (Social Sciences), 'LS' (Life Sciences and Medicine), 'PS' (Physical Sciences) – attributed to the disciplines as indicated above
grade	'M' (merit, 60-69%), 'D' (distinction, 70-100%)
number of authors	n - representing number of students involved in authoring assignment

Author-related information is found in **`teiHeader|profileDesc|particDesc|person`**. Different pieces of information are stored in **p** elements, distinguished by the value of their *n* attribute:

Table 9. Author-related contextual information

n=	**possible values (content of p)**
gender	'm', 'f'
year of birth	four digits
first language	as indicated by student
education	assignment authors were asked for "secondary education (from 11 years but not including university)"; possible values are: 'UKa' (all in the UK) 'OSa' (all overseas) 'UK'+digit (UK1, UK2...) (n years (but not all) in the UK)

n=	**possible values (content of p)**
course	i.e. programme of study: non-standardized string
student ID	four digits assigned by the project team, whose purpose is to make it possible to determine whether two assignments have the same author

b)Information derived from the document

- document title: apart from the document text, the title appears in the header in **`titleStmt|title`** – which contains either the full or an abbreviated assignment title (as a string); this title may not be exactly identical with the title appearing in the assignment itself (which is contained in the text). Assignments that have no title may be given one here.
- presence and number of occurrence of specific features (elements): like assignment-related contextual information, this information is stored in **`teiHeader|fileDesc|sourceDesc|p`**. The type of information is specified in an attribute *n* of **p**; values are given as text content of **p**. Types of information (and their corresponding values of *n*) are:

Table 10. Textual information

p *n=*	**information type and possible values**
number of words	words in **body**, not counting notes; integer
number of s-units	s-units (see appendix 3) in **body**, not counting notes; integer
number of p	p (see below 3.2.b) in **body**, not counting notes; integer
number of tables	tables in **body**; integer
number of figures	figures in **body**; integer
number of block quotes	block quotes in **body**; integer
number of formulae	formulae in **body**; integer
number of lists	lists in **body**; integer
number of paragraphs formatted like lists	paragraphs in **body** with a bulleting or ordering character; integer
abstract present	possible values: 'abstract present', 'no abstract'
average words per s-unit	rounded to one decimal

p *n=*	**information type and possible values**
average s-units per p	rounded to one decimal
macrotype of assignment	possible values: 'compound assignment consisting of [nb] parts (see notesStmt for details)' [nb being the number of **div1** *type="text"* elements in the body], 'complex but non-compound assignment (see notesStmt for details)' 'simple assignment'

c)Tagger's notes concerning the document

Tagger's notes are meant to give clarification on individual assignments, or to help the user looking for a particular feature (e.g. types of **compound assignment**: see below 6.3). They appear in **teiHeader|fileDesc|notesStmt**, which contains any number of **note**; **note** always has the attribute: *resp="British Academic Written English (BAWE) corpus project"*. Notes are created by the tagger during the process of annotation and contain, as string, any information considered worth communicating to the user of the document; in particular, the following information (if applicable) appears as tagger's notes:

- content of deleted page header and footer
- word count or grade deleted
- appendix content
- candidate compound assignment and which decision was taken (see below 6.3)
- elements located on title page

Apart from the last two, notes are typed in manually, thus their text is not standardized. If no note has been made, **notesStmt** contains the empty **note** element (requirement of the DTD).

d)Other header content

BAWE-invariant:

- **publicationStmt|distributor**, always contains the string: 'British Academic Written English (BAWE) corpus'

- **publicationStmt|availability**, always contains six **p** specifying the conditions of availability of the BAWE corpus:

  p: 'The British Academic Written English (BAWE) corpus was developed at the Universities of Warwick, Reading and Oxford Brookes, under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC. Subject to the rights of the these institutions in the BAWE corpus, and pursuant to the ESRC agreement, the BAWE corpus is available to researchers for research purposes PROVIDED THAT the following conditions are met:'

  p: '1. The corpus files are not to be distributed in either their original form or in modified form.'

  p: '2. The texts are used for research purposes only; they should not be reproduced in teaching materials.'

  p: '3. The texts are not reproduced in full for a wider audience/readership, although researchers are free to quote short passages of text (up to 200 running words from any given text).'

  p: '4. The BAWE corpus developers (contact: Hilary Nesi) are informed of all projects, dissertations, theses, presentations or publications arising from analysis of the corpus.'

  p: '5. Researchers acknowledge their use of the corpus using the following form of words: "The data in this study come from the British Academic

Written English (BAWE) corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).'"

- **`teiHeader|encodingDesc|p`**, in all BAWE documents, always contains the string 'TEI P4 (documented in: BAWE-documentation.Markup.pdf)'
- an empty element **`fileDesc|extent`**: this is a requirement of the TEI DTD (from which the DTD used for BAWE is derived)

6.2. The text of the document

The text of the assignment is contained in **`TEI.2|text`**, which is further sub-divided by three elements: **`front`**, **`body`**, **`back`**.

The distinction between them is based on the notion of **running text**: **`body`** contains the running text, starting with the title of an abstract, the first section title or the first sentence; everything in the text occurring before the start of the running text is contained in **`front`**; everything occurring after the end of the running text in **`back`**.

Various features are annotated in each of these parts, discussed in the following sections. As for its hierarchical organization, the body is further structured on three levels:

a)Sections

Sections of the body are marked up using the **`div1`** element; subsections of **`div1`** by **`div2`**, subsections of this by **`div3`** etc. until **`div7`** (the maximum level of embedding allowed by TEI). If an assignment has no sections at all, the whole running text will be contained in one **`div1`** (the only child of **`body`**). Except for special types of sections (part of a compound assignment, table of contents, abstract, front or back

matter, bibliography, appendix etc. – discussed in their relevant sections), **div1** has the attribute *type="section"*.

A section or subsection (of any level) consists of a **head** (containing the section title) and any number of paragraphs (**p**). If a (sub-) section begins without a section title, there is no **head** element.

b)Paragraphs

The text of a section is further structured into paragraphs, marked up as **p**. Features marked up as direct child elements of their sections (i.e. not contained by a **p**) are:

- section heading
- figures and pictures
- tables
- lists
- block quotes
- notes

c)S-units

A third level of text structure is introduced by a sentence splitter script, operating on the basis of a simple algorithm (for details see appendix 3). The resulting units are called **s-units** (following TEI terminology), rather than sentences, to indicate the possibility of errors. They are marked up using the tag **s**.

d)Numbering

To facilitate orientation within the text, both **p** (within the body) and **s** are numbered using an *n* attribute:

- **p** in running text are numbered throughout the body following the format: **p** *n="p13.17"*, identifying the 13th paragraph of 17 paragraphs contained by the body (running text).
- **p** in **note** elements are not considered to be part of the running text and therefore not counted in the total number of paragraphs; their value for *n*

identifies the number of the current paragraph and the total number of paragraphs in the current note. The format is: **p** *n="pn3.3"* (i.e. the third of three paragraphs in the current note).

- **s** are numbered throughout a paragraph (of a note or running text) following the format: **s** *n="s4.7;p3.17"* identifying the 4th s-unit of 7 contained by the current paragraph, which is the 3rd of 17 paragraphs contained by the body. The numbering format for **s** in notes is: **s** *n="s1.2;pn1.1"*.

*6.3. Macro-types of assignments (**compound**, **complex** or **simple**)*

A number of assignments are atypical in the sense that it may seem inappropriate to treat them as single coherent texts. With respect to genre analysis, one may want to analyze their constitutive parts rather than the assignment itself – e.g. a collection of (short) essays. On the other hand, the assignment may well be set as one task, but consist of a number of sub-tasks, resulting in a similar re-iterative structure – e.g. an annotated bibliography.

The need to identify problematic cases arises from the projected use of the corpus: providing a base of data that should allow the investigation of text types (genres). The BAWE corpus has attempted to find a balance between a purely pragmatic approach (based on the **assignment** – as a unit of evaluation), and a text linguistic approach (**textuality** of units). As a result, one assignment is represented as one BAWE corpus file; but these problematic cases are identified and classified.

Hoey's (2001) concept of **text colonies** has proven useful for identifying assignments consisting of rather autonomous textual units; however, it does not allow a distinction between assignments that provide a rather strong framework of coherence (e.g. annotated bibliography) and assignments appearing as a loose association of rather independent texts (e.g. collection of essays). The fact of an assignment being a text colony was therefore insufficient for making a final decision.

In the BAWE corpus, three macro-types of assignments are distinguished:

(1) **simple** assignments: which do not have the properties of colonies

(2) **compound** assignments: set as one assignment, consisting of several, independent and rather long texts/tasks

(3) **non-compound** (but non-simple) assignments: also have the properties of colonies, but their parts seem not independent enough to be considered separate units alongside full assignments

The distinction between compound and non-compound (but not simple) types considers the (pragmatic) criterion of plausible independence of the parts. Overall thematic coherence and length of the individual parts are parameters taken into account.

At the level of **section markup**, both regular sections and the parts of compound assignments are represented by a **`div1`** element. A distinction is made through the value of an attribute *`type`*:

`div1` *`type="text"`*: designates the beginning of a new part of a compound assignment;

`div1` *`type="section"`*: is used for sections/parts of non-compound assignments (simple or not).

Where the parts of a compound assignment themselves have sections, these are represented by a **`div1`** element with *`type="section"`*:

(The *`type`* attribute is also used for identifying sections of a 'special' type – abstract, table of contents, front or back matter, bibliography, appendix – cf. the following sections and appendix 2.)

The macro-type of an assignment is stated in the **header** element **`teiHeader|fileDesc|sourceDesc|p`** *`n="macrotype of assignment"`* (cf. b)); the text content of the element is either 'compound assignment consisting of [nb] parts (see notesStmt for details)' [nb being the number of **`div1`** *`type="text"`* elements in the body]; 'complex but non-compound assignment (see notesStmt for details)'; or 'simple assignment'.

In addition, the identification of a potentially compound assignment, leading to a decision between (2) and (3), is documented in a **tagger's note** (cf. c)) stating their (non-)compoundness and the reason for deciding so (created automatically).

(1) 'Compound' assignments are identified through the note:

'Evaluated as candidate compound assignment. Assigned to [rule]: compound.' – where [rule] is replaced with the rule applied.

The rules for compound assignments are:

S1: collections (other than S1a-f)
S1a: collection of essays/discussions
S1b: collection of reports
S1c: collection of research proposals
S1d: collection of critical reviews
S1e: collection of assignments structured by questions
S1f: collection of long tasks/questions

A1: text + independent QA

A4: text + major reflection task (other than A4a)
A4a: creative writing + reflection

A5: Mixed (other than A5a)
A5a: essay and literary criticism

The cases listed under S1 refer to a 'symmetric' assignment structure, consisting of a number of texts of the same type. Rules A1, A4 and A5 refer to 'asymmetric' types of compound assignments: in A1, some type of text is followed by a questions and answer section; in A4 a piece of writing is followed by a (major) reflection task; A5 contains cases of compound assignments not covered by other rules (look for tagger's notes for specifications).

(2) 'Non-compound' non-simple assignments are identified through the note:

'Evaluated as candidate compound assignment. Assigned to [rule]: non-compound.' – where [rule] is replaced with the rule applied.

The rules for non-compound assignments are:

S2: diary
S3: assignment structured by questions
S4: set of tasks/questions/exercises
S5: annotated bibliography

A2: text + dependent QA

A3: text + minor reflection task (other than A3a-d)
A3a: lab report + reflection
A3b: essay + reflection
A3c: set of exercises + reflection
A3d: annotated bibliography + reflection

Assignments in S2-5 are 'symmetric' in the sense that they show an iterative structure, consisting of a number of parts of similar types; they either result in a coherent piece of writing or are too short for one to assume that they have been set as independent

tasks. S3 refers to cases where the assignment task is set as a number of questions, answered in turn (and, frequently, giving the questions as section headings); however, questions in these cases are not totally independent but meant to structure a piece of writing that finally appears as one coherent text. The parts of assignments of type S4 may be independent, although very short. The assignments listed in A3 consist of various types of texts followed by a (minor) reflection task.

By providing information on compoundness, it is our intention to allow the user of the corpus

a) to target instances of genres that appear as parts of compound assignments for specific analyses;

b) to exclude compound (and, where necessary, non-compound non-simple) assignments from certain analyses where textual cohesion and coherence of the whole is an essential assumption.

It is not our intention to provide a wide and representative material base for the study of compound texts (or colonies) in student writing; compoundness is therefore not a criterion available for selection in our online concordancing interfaces (the Sketch Engine, BAWE online corpus query interface at Coventry University – see references section for URLs).

7. Textual features occurring in the front

7.1. Title page

A document has a title page when its first page displays some (not necessarily all) of its front matter, with running text being excluded from it by some lay-out arrangement (white space, manual page break or other). (Running text then starts either with the first main section heading or, if there is no further division of the document, with the first sentence.)

If there is a title page, the document header contains a note stating 'The document has a title page which contains: [features in title page]'. The element **`front|titlePage`** (cf. next section) does *not* reflect the use of a title page as layout device, but is used for the encoding of all document titles, due to restrictions imposed by the TEI standard.

7.2. Document title

The document title contains everything that would intuitively be counted as a title (such as the two-fold structure [1. question or quote] plus [2. instruction for assignment]), thus including subtitles (but excluding other elements associated with the title, such as course/module title, school/department, student ID etc.). The title of the document can therefore comprise more than one adjacent paragraph.

Quotations given as part of the assignment task are part of the title and must not be confounded with mottos or leading quotations, which are inserted by the author and thus characterise the text individually ('epigraphs', cf. below).

The title of a document is encoded as **`front|titlePage|docTitle|titlePart`**. As noted above, **`docTitle`** always is wrapped in **`titlePage`** – regardless of the actual occurrence of a title page in the assignment.

7.3. Other elements associated with the title

Elements of the front frequently appearing around the document title are references to: student name and number, date, module title and code, module leader, assignment type, university, department, course of studies and the like.

All these elements are encoded as **`front|titlePage|titlePart`** – whereby adjacent elements may occur in the same **`titlePart`** element.

7.4. Epigraph

A motto or leading quotation ('epigraph'), associated with the document title, is encoded as **`front|titlePage|epigraph|p`**. Epigraphs characterize the individual assignment and should not be confused with quotations given in the assignment title (e.g. for the student to comment on).

7.5. Table of contents

A table of contents is a list of section titles; the list may be exhaustive or (especially in the case of longer and more complex documents) cover only a certain number of

levels of section titles. It may or may not indicate page numbers. Further, if the table of contents is written manually, there may be inconsistencies with the actual section titles.

A table of contents is marked up as **`front|div1`** *`type="toc" n="[number of section levels represented in the ToC]"`*. Its title appears in a **`div1|head`** element.

7.6. Other text as front matter

Other text as front matter appears in an element **`front|div1`** *`type="front text"`*.

This may include a list of keywords, abbreviations or cases referred to, a glossary, acknowledgments, or some other remark by the student preceding the actual text of the assignment.

7.7. Additional elements in the front

Some elements may occur within the front, although they are not necessarily *front matter* in themselves. These include:

- figures
- pictures

They are marked up like the corresponding elements in the body of the document (for details refer to the discussion in section 9).

8. Textual features occurring in the back

8.1. Bibliography

A bibliography section is any list of source material given at the end of an assignment or task. Apart from a classical list of 'references', the bibliography section may include: a table of law cases, a filmography, sources of illustrations, figures, tables and the like. There may be more than one bibliography section in one assignment, and the bibliography may itself be divided into sub-sections.

The bibliography section is marked up as **back|div1** *type="bibliography"*; for further subdivisions **div1|div2|div3** etc. are used. As usual with section headings, the title of the bibliography appears in **div1|head**. It is possible that there is no section heading indicating the start of the bibliography, in which case there is no **head** element.

The text of the bibliography is contained in **p** elements; it is frequently, but not always the case that one **p** corresponds to one bibliographical entry. If there is further text in the bibliography section (e.g. *The following books were consulted during the writing of this report* or the like), it is not marked up in any particular way.

Annotated bibliographies are a special case of bibliographies; since they constitute tasks themselves, rather than being appended to a task, and since they contain text written by the student, they are marked up as normal sections of the running text, thus **body|div1**. Each bibliographical entry is then marked up as a sub-section of the annotated bibliography.

In terms of macro-types of texts, annotated bibliographies – as colonies of texts (Hoey 2001) – are considered to be non-compound non-simple texts (cf. section 6.3).

8.2. Appendix

The appendix is a section which is entitled *appendix*, *annex* or something very similar. It may contain different kinds of information, like tables, figures, illustrations, but also text. An assignment can have one or more appendices, and an appendix can be divided into sub-sections. Appendices are marked up using the element **back|div1** *type="appendix"*.

The content of appendix sections is deleted and only the section title is kept in the **head** element: besides the appendix section title, **div1** *type="appendix"* only contains an empty **p**. A note of the type of material contained by the appendix is made manually by the tagger in the header's **notesStmt**.

In some documents, the appendix may be missing or located in another file. We check for missing appendices by running a search on the strings *appendix*, *annex*,

#app and *#app#* (# representing a space character). If occurrences of any of these are found, the tagger has to decide whether these refer to a missing or external appendix. If so, the missing or external appendix is represented by an element **div1** *type="missing or external appendix"*, containing only an empty **p**.

8.3. Other text as back matter

Other text as back matter appears in an element **front|div1** *type="back text"*. This may include a glossary, abbreviations, acknowledgements or other text in a section other than bibliography or appendix.

9. Textual features occurring in the body

9.1. Abstract

In the BAWE corpus, an abstract is defined as a section before the main assignment text labelled *abstract*, *summary*, *executive summary* or the like. Abstracts are considered to be a special type of section within running text and tagged as **body|div1** *type="abstract"*. As with every **div1**, the title of the abstract is in a **head** element.

9.2. Figures

Figures are: graphs, images, drawings etc. and may either be inserted into running text or occur as front/back matter (e.g. on the title page, in the appendix). Apart from the figure itself, there may also be some text associated with it: a caption, source statement, or explanation.

Figures are marked up using the **figure** element, with the only content **figure|head** (the caption). This caption may span across several paragraphs in MS Word, although the paragraphs are not further marked up. Any explanatory or illustrative text attached to the figure is also marked up as **head**.

The **figure** element has an *id* attribute, which provides a symbolic link to the original file: the ID contains the file name and numbers the figure of a given type

throughout the document. An example of an ID value for a figure is: **figure** *id="BAWE_1004b-fig.003"* (i.e. the third figure in the document 1004b).

Figures are not embedded within **p**, but directly within the division element to which they belong (**div1**, **div2** etc.).

The **figure** element appears where the figure stood in the original document; if the figure was surrounded by a paragraph of running text, **figure** may either precede or follow the corresponding **p**, whichever seems more plausible.

9.3. Pictures

Pictures, including caption and accompanying text, are marked up like figures, using the **figure** element. They are distinguished from figures through the value of their *id* attribute, which contains 'pic' instead of 'fig'. An example is: **figure** *id="BAWE_1004b-pic.003"* (i.e. the third picture in the document 1004b). Figures and pictures are numbered separately.

Like figures, pictures are not embedded within **p**, but directly within the division element to which they belong (**div1**, **div2** etc.). If a picture is surrounded by a paragraph of running text, **figure** may either precede or follow the corresponding **p**.

9.4. Tables

Tables are spans of the text outlined as a row-column intersection. They may have been created using MS Word tables or some other means of text lay-out (e.g. tab stops, white space). Tables can contain numbers, words or other symbols. Like figures and pictures, there may be captions or explanatory text accompanying the table. Where a MS Word table is used merely for layout purposes (e.g. creating a list), it is not considered as 'table' in the BAWE corpus.

In the BAWE corpus, tables are marked up as **table**; if there is a caption and/or other accompanying text, it is contained in a **head** element.

In analogy with figures and pictures, the **table** elements of a document are numbered by their *id* attribute. The format used is: **table** *id="BAWE_6004b-tab.003"* (i.e. the third table of the document 6004b).

Apart from the (optional) **head**, **table** contains one **row** element, which contains one empty **cell**. This is due to a requirement of TEI.

Like figures and pictures, tables are not embedded within **p**, but directly within the division element to which they belong (**div1**, **div2** etc.).

9.5. Lists

a)'Genuine' lists

We consider as a list a sequence of items laid out in a way so that each item appears on a new line. In addition, further white space (indentation etc.) or symbols marking each list item may be used. In rare cases (i.e. where a list is created in a parallelism with other lists in the document), a list may consist of only one item.

Three types of lists are distinguished by their layout:

- **ordered lists**: items are marked by consecutive numbers or letters (e.g. *1), 2), 3)...*; *a, b, c...*; *i, ii, iii...* or the like)
- **bulleted lists**: all items are marked by the same symbol, e.g. bullet points, hyphens, arrows etc.
- **simple lists**: items are marked only by the use of white space and carriage returns, no special symbols are used

For the markup of lists, the element **list** is used, which appears directly in the section to which it belongs (**div1**, **div2** etc.). The type of the list is specified in the attribute *type="ordered|bulleted|simple"*. If the list has a caption (title), it is marked up as **list|head**.

As for the list items, an automatic method of identification is applied: each MS Word paragraph in the list is considered to constitute a list item. It is clear that this method leads to false divisions where an item contains multiple paragraphs. Items are marked up as **list|item**. The item directly contains the list text; neither **p** nor **s** elements are used within lists.

In the case of ordered or bulleted lists, the item will start with the ordering/bulleting character if it was inserted by hand; where ordering or bulleting is automatic (using list styles), these characters are not preserved in the BAWE file.

In a nested list of lists (i.e. one or more items of the outer list contain an inner list), only the outer list is marked up.

b)False lists (or list-like formatted paragraphs)

Typically, a list item would contain a rather small amount of text, consisting of only one word or phrase. As lists and their items would not be expected to form full sentences (although that may be the case), they can be seen as elements embedded in, but different from running text.

However, ordering characters or bullet points can also be used for presenting a sequence of quite large textual units; in these cases, it appears that the running text itself is presented as a sequence of units. In the BAWE corpus, such items are considered as paragraphs of running text carrying **list-like** formatting.

Like other formatting properties (cf. section 9.9), this information is marked up in the *rend* attribute of the paragraph. Thus, we get: **p** *rend="ordered|bulleted"*.

N.b. As the *rend* attribute is designed to convey all information on special formatting, its value for 'list-like' formatted paragraphs may not be *exactly* 'ordered' or 'bulleted', but may contain other properties as well (e.g. 'ordered bold', 'bulleted underlined italic' etc.).

The boundaries between genuine lists and running text presented in a list-like format are inherently fuzzy, and any quantitative threshold (e.g. number of words contained etc.) would therefore be arbitrary. For the purpose of BAWE markup, a list was only considered to be genuine if it did not primarily consist of syntactically complete sentences; if a major part of the list consisted of complete sentences, the paragraphs containing numbering or bulleting characters were marked up as list-like formatted.

9.6. Formulae

A **formula** in the BAWE corpus consists of a formulaic expression in the widest sense, which includes phenomena such as:

- algebraic expressions

- logical expressions
- chemical formulae
- computer code
- phonetic transcriptions
- example sentences (e.g. for linguistic analysis)

The defining property of formulaic expressions is that they are either expressions in a non-natural language code or natural language expressions as object of study, not as a means of communication.

However, formulaic expressions are only marked up as formulae if they are sufficiently salient in the text. By this we mean:

- any expression inserted with the MS formula editor
- any *complex* formulaic expression, i.e. one that cannot be represented as a simple sequence of characters (e.g. fraction, square root)
- formulaic expressions separated typographically from running text (carriage return)

Everything else, i.e. any linear sequence of characters within a paragraph of natural language, but also formulaic expressions starting a new line, but linked with natural language, e.g. to form a gloss-like expression (e.g. c^2*: the square of the speed of light*), is not marked up as formula, but treated as running text.

The element used is the empty element **`formula`**; the formula content is thus suppressed in the corpus. The **`formula`** element has two attributes *`notation=""`* (a mandatory attribute according to the TEI P4 standard, but irrelevant for BAWE; its value is always an empty string) and an *`id`* attribute whose value follows the format: *`id="BAWE_6214b-form.001"`* (meaning: the first formula in the document 6214b).

9.7. Block quotes

Quotations that appear in a (number of) separate paragraph(s) of their own (possibly also containing some reference to the source) are marked up as block quotes. This includes quotations from scientific as well as literary works (poems, drama…).

The element used is **quote**, which directly contains the text quoted. The **quote** element is neither wrapped in, nor does it contain **p** elements. An attribute *lang* specifies the language of the quote (*lang="[name of language]"*; *lang="English"* is used for modern standard English, *lang="English-non-std."* for other varieties of English).

For every value for *lang* within a document, there is also a **notesStmt|note** in the header defining it in the following format:

```
<note resp="British Academic Written English (BAWE) corpus
project">
Language used in quote:
<foreign id="Latin">Latin</foreign>
</note>
```

9.8. Notes

Foot- and endnotes of the document are only marked up if they have been created using the MS Word foot-/endnote function. In order not to interrupt the flow of running text, a two-part system for marking up notes has been used for BAWE:

1. at the place where the note is inserted into running text, i.e. in an **s** element, a **ref** indicates the presence of a note. The **ref** element is empty and identifies the note in a *target* attribute; the format is:
   a) for footnotes: **ref** *target="BAWE_6211a-ftnote.002"* (i.e. second footnote in the document 6211a)
   b) for endnotes: **ref** *target="BAWE_6002a-endnote.001"* (i.e. first endnote in the document 6002a)

2. All notes occurring in a paragraph of running text (**p**) are inserted immediately after **p**, in the order in which they occurred. Notes occurring within a **quote** or **list|item** are inserted immediately after **quote** or **list**.
   A **note** element is used for the markup of foot- or endnotes, specifying the type of note in the attribute *place="foot|end"* and identifying the note through the value of an *id* attribute, which is the same as the value of *target* in the corresponding **ref**. Thus, for the above examples we get:

a) **note** *place="foot" id="BAWE_6211a-ftnote.002"*

b) **note** *place="end" id="BAWE_6002a-endnote.001"*

The **note** element is thus directly contained by the section to which it belongs (element **div1**, **div2** etc.). The text within foot- or endnotes is divided into paragraphs (**p**), which are subdivided into s-units (**s**).

The numbering of **note|p** and **note|s** is analogous to that of **p** and **s** occurring within sections of running text:

- **p** and **s** numbers are contained in the attribute *n*
- **p** numbers are counted within the note (not continued from running text); the format is: **p** *n="pn2.3"*, i.e. the second out of three **p** in this note
- like in running text, **s** numbers are counted within the **p**, repeating the current **p** number: **s** *n="s1.4;pn2.3"* (first of four **s** in the second of three **p** in this note)

9.9. Front/back matter embedded in running text

Some parts of the assignment are logically front matter, but already preceded by running text, so that they cannot be tagged as **front** without changing the original sequence of elements. The same is true of elements that are logically back matter, but followed by more running text. Examples are: a list of keywords after an abstract or introduction; or a references section after the first part of a two-part assignment.

Such elements appear in a **div1** *type="front-back-matter"*. Their text is contained in **p** (without numbering and without further **s** elements).

It should be emphasized that **div1** *type="front-back-matter"* actually occurs within **body**; the value of the attribute is meant to allow the user to disregard these elements where necessary.

10. Highlighting

In the BAWE corpus, the following properties of special character formatting are marked up:

- bold

- italics
- underlined (with any style)
- colour (other than black)
- subscript
- superscript

(Although sub- and superscript are strictly speaking not formatting properties used for highlighting, they are included here.)
Information on special formatting is contained by a *rend* attribute; the following values are used:

rend=	feature
bold	**bold**
italic	*italics*
underlined	underlined, underlined etc.
colour	colour
sub	subscript
sup	superscript

If more than one feature of special character formatting is present, they are combined in the *rend* attribute and separated by a space, e.g. *rend="bold underlined"* (cf. appendix 2 for a complete list of actual values of the *rend* attribute).

Location of the *rend* attribute:

- if the entire text content of the element containing that text shares the same special formatting, *rend* is an attribute of that parent element (i.e. **s**, **p**, **head**, **item**, **quote** etc.);
- if the highlighting applies only to a part of the text content of the parent element, a **hi** element is created which has the attribute *rend* (with the appropriate value) and contains the highlighted text.

11. Other information encoded

11.1. Anonymization of personal data

All explicit information allowing the identification of the assignment author or other persons directly involved in any aspect of the production of the assignment has been removed from BAWE corpus documents. To indicate where such information occurred in the original assignment, the empty element **name** is used. The type of information removed is documented in the value of the attribute *type*. Possible values for types of personal information are:

Table 11. Anonymization codes

type=	information type encoded
student name	student name
student ID	student number
university	university
tutor name	tutor, supervisor or module leader
dedicatee	dedicatee
date	date (in medical patient reports)
other	other personal information removed

(Other personal information may include address, email address, names of other persons involved etc.)

11.2. URLs

URLs are identified as strings according to the following rules:

- They start with a prologue which can be any of these:

  http://, https://

  http://www., http://www2., https://www., https://www2.

  www., www2.
- Followed by one word character: i.e. alphanumeric or underscore (_)
- Followed by any sequence of either of these characters:

  word characters

  special characters: ***.,!?+-$*()'%/~#=&***
- Followed by an alphanumeric character

- Followed by an optional /

The whole string is replaced in BAWE documents with an empty **seg** element, which has two attributes: *type="URL"* and *n="[string value of URL]"*. Highlighting of URLs (e.g. underlining) is not marked up in any way.

12. The TXT version of the BAWE corpus

In addition to the XML version described in sections 1-11 above, the BAWE corpus is also available as a collection of TXT files, with one assignment corresponding to one TXT file. This version is intended for use with non-XML aware software. The features of the TXT version of BAWE are described in this section.

Generally, the TXT version contains some very simplified markup derived from the full markup of the XML version. The various features marked up in the XML files are treated in the following ways in the TXT version:

- **Contextual information** (**teiHeader**) is omitted in the TXT version.
- **Front and back matter** (i.e. the content of **front** and **back**) is omitted; the TXT version only contains text of the XML body. Likewise, front/back matter embedded within running text (cf. 9.9above) is omitted.
- **Sections** of running text are not marked up explicitly, but section headings are enclosed in a **heading** tag.
- **Paragraph** (**p**) tags are omitted.
- **Sentence** (**s**) tags are omitted.
- The **macro-type of assignments** (including types of sections; cf. 6.3above) is not indicated.
- **Abstracts** or summaries (at the start of running text) are enclosed in an **abstract** tag.
- **Figures** (including caption or other accompanying text): are replaced with an empty **figure** tag.
- **Pictures** (including caption or other accompanying text): are replaced with an empty **picture** tag.

- **Tables** (including caption or other accompanying text): are replaced with an empty **`table`** tag.
- **Lists**: are enclosed in a **`list`** tag.
- **Block quotes**: are enclosed in a **`quote`** tag.
- **Footnotes**[2]: are enclosed in a **`fnote`** tag.
- **Endnotes2**: are enclosed in an **`enote`** tag.
- **Highlighting information**: is omitted.
- **Anonymization tags** (representing student name and number, university etc.; cf. 11.1above): are omitted.
- **URLs**: are replaced with the string 'URL'.
- **Formulae**: are replaced with the string 'FORMULA'.
- **Encoding**: like the standard XML version, the TXT version is encoded in UTF-8 Unicode.

13. The PDF version of the BAWE corpus

The PDF version of the corpus consists of PDFs created from the untagged, anonymised Word files. It is intended to be useful for researchers who are interested in layout properties of student assignments.

[2] Like in the XML version, notes do not interrupt running text, but appear after the paragraph in which they were inserted in the original assignment.

14. References

Alsop, S. and H. Nesi under review. Issues in the development of the British Academic Written English (BAWE) corpus. Submitted to *Corpora*.

Conrad, S. and D. Biber. 2001. *Variation in English: multi-dimensional studies*. Harlow, Essex: Pearson.

Hoey, M. 2001. *Textual interaction: an introduction to written discourse analysis.* London: Routledge

Library of Congress. 2006. *Codes for the representation of names of languages. Part 2: alpha-3 code* (ISO 639-2). Available online from http://www.loc.gov/standards/iso639-2/

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

14.1. Websites

BAWE corpus. Project homepage at the University of Warwick.
http://www2.warwick.ac.uk/fac/soc/celte/research/bawe/

BAWE corpus. Documents and online query interface at Coventry University.
http://www.coventry.ac.uk/bawe

The Sketch Engine. Corpus query interface.
http://www.sketchengine.co.uk/

TEI PizzaChef.
http://www.tei-c.org.uk/pizza.html

Unicode. Character set.
http://www.unicode.org/Public/UNIDATA/UnicodeData.txt

15. Project team

Prof Hilary Nesi, English Language Unit, Coventry University (principal investigator)
Dr Sheena Gardner, School of Education, University of Birmingham
Dr Paul Thompson, Department of Applied Linguistics, University of Reading
Dr Paul Wickens, Westminster Institute of Education, Oxford Brookes University
Dr Richard Forsyth, CELTE, University of Warwick
Mr Alois Heuboeck, Department of Applied Linguistics, University of Reading
Dr Jasper Holmes, CELTE, University of Warwick
Ms Sian Alsop, CELTE, University of Warwick
Dr Signe O. Ebeling, Westminster Institute of Education, Oxford Brookes University
Ms Dawn Hindle, CELTE, University of Warwick
Ms Maria Leedham, Westminster Institute of Education, Oxford Brookes University

Appendix 1copyright waiver forms

Oxford Brookes:

The Corpus of British Academic Written English (BAWE)

The International Centre for English Language Studies (Westminster Institute of Education) has received ESRC funding to develop an electronically stored corpus of proficient student writing. Data for the corpus will be collected at Warwick University, Reading University, and Oxford Brookes University. The full corpus (containing about 4000 examples of anonymised student work) will eventually be accessible to researchers via the Oxford Text Archive and the ESRC Archive. Other interested parties such as writing tutors and students may be granted limited access to parts of the corpus, for example via an online concordancer.

The corpus will contain a wide variety of proficient assessed writing produced at postgraduate and undergraduate level, drawn from all faculties within the University.

Corpus holdings will be categorised according to faculty, department, course module and year of study.

The primary purpose of the corpus will be to enable us to identify the characteristics of student writing, and compare the writing produced in different disciplines and at different levels of study. Corpus samples may also eventually be used in teaching materials.

In order to compile and assemble this corpus, it is the intention of ICELS to advertise for model assignments, through posters and fliers and through departmental contacts. Contributing students will be given £3.00 for each assignment they submit to the corpus, and will be required to make an assignment of their Intellectual Property rights in their submitted assignments.

The Corpus of British Academic Written English (BAWE)

a)Submission form

Surname: ..... Forename: .....

Male/Female Date of birth: ..... 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. 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: .....

..... *("the 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):

In consideration of the sum of £3.00 paid by Oxford Brookes University ("the University") to me the receipt of which I hereby acknowledge I hereby *assign* to the University the Assignment and my intellectual property rights in the same together with waiving my moral rights. I warrant that the Assignment is my own work and agree that I shall advise the University of the quotation or inclusion in the Assignment of any textual or illustrative material taken from other sources and provide the University with full details of the original source of that material.

I acknowledge that the Assignment may be submitted to the JISC Plagiarism Advisory Service (a facility which carries out electronic comparison of students' work against electronic sources, including work submitted by students at other institutions).

..... (Signature)

..... (Date)

Thank you for contributing to the BAWE corpus.

Reading:

BRITISH ACADEMIC WRITTEN ENGLISH CORPUS
Agreement

"Work" means the assignment identified overleaf, of which the student signing below certifies to be the sole author.

IT IS HEREBY AGREED as follows:

1. In consideration of the sum of £3 paid by The University of Reading to the Student (the receipt whereof is hereby acknowledged) the Student hereby assigns to the University all copyright in the Work as defined above. This means that The University will be the owner of all the rights in the Work worldwide.
2. The Student shall advise the University of the quotation or inclusion in the Work of any textual or illustrative material taken from other sources and provide the University with full details of the original source of that material.
3. **The University shall have the right to obtain protection in any form deemed appropriate by the University throughout the world in respect of the Work.**
4. The Student hereby asserts to the University his/her moral right to be identified as the author of the Work in accordance with sections 77 and 78 of the Copyright Designs and Patents Act 1988 or any subsequent enactment thereof. For the avoidance of doubt, whilst this is generally the case, the British Academic Written English Corpus will hold the Work anonymously.
5. The University grants a non-exclusive royalty-free licence on the Work to the Student for his/her personal use.

I have read and understood the information about copyright and I agree to assign copyright of my assignment to The University of Reading. I certify that I am the sole author of this work. I also permit my work to be submitted to the JISC Plagiarism Advisory Service (a facility which carries out electronic comparison of students' work against electronic sources, including work submitted by students at other institutions).
I acknowledge receipt of £3 in payment for submission of this assignment.

..... Date.....

(Signature)

Signed on behalf of The University of Reading:

…….. Date……..

Dr. Paul Thompson

Warwick:

BAWE Corpus of British Academic Written English Submission Form

Thank you for contributing to the BAWE corpus. Please read the information about copyright overleaf, complete the information about yourself and the assignment you have submitted, and sign the form at the bottom.

AUTHOR INFORMATION

Family Name	
Given Name	
Date of Birth	
Male / Female	
Preferred E-mail Address	
First Language	
Secondary School Education (please pick one): All in UK / All overseas / Some in UK, some overseas Please state years in UK	

ASSIGNMENT INFORMATION

Essay ID	
Year of study when assignment written: Undergraduate: 1st / 2nd / 3rd / 3rd (with intercalated year) Postgraduate: 4th	

Home department	
Course of Study	
Brief title of assignment	
Year and month assignment written	
Module title	
Module Tutor	
Module code	
Grade	
Number of Authors	
Type of assignment (please pick one): Case-Study / Essay / Exercise / Notes / Presentation / Report / Review / Other (please specify)	

I have read and understood the information about copyright and I agree to assign copyright of my assignment to the university. I certify that I am the sole author of this work, or in case of co-authorship, that I have obtained permission to submit from my co-authors. I also permit my work to be submitted to the JISC Plagiarism Advisory Service (a facility which carries out electronic comparison of students' work against electronic sources, including work submitted by students at other institutions).

..... (Signature)

..... (Date)

Copyright Information

"Work" means all essays, problem answers, reports, questionnaires, statistical reports, summaries, whether recorded electronically or otherwise

IT IS HEREBY AGREED as follows:

In consideration of the sum of £3 paid by the University to the Student (the receipt whereof is hereby acknowledged) the Student hereby assigns to the University;

a. the benefit of the Work the right title and interest therein and all rights powers liberties and immunities arising or accrued therefrom free from all encumbrances to the intent that any rights granted pursuant to the Work shall be in the name of the University and shall vest in the University absolutely including but not limited to the entire copyright in the Work throughout the world for the full legal term of copyright and all renewals and extensions thereof and;

b. the right to apply for prosecute and obtain protection in any form deemed appropriate by the University throughout the world in respect of the Works.

The Student shall advise the University of the quotation or inclusion in the Work of any textual or illustrative material taken from other sources and provide the University with full details of the original source of that material.

The Student hereby asserts to the University his/her moral right to be identified as the author of the Work in accordance with sections 77 and 78 of the Copyright Designs and Patents Act 1988 or any subsequent enactment thereof.

==========

BAWE Project:
The Corpus of British Academic Written English (BAWE)

The Centre for English Language Teacher Education has received ESRC funding to develop an electronically stored corpus of proficient student writing. Data for the corpus will be collected at Warwick University, Reading University, and Oxford Brookes University. The full corpus (containing over 3000 examples of student work) will eventually be accessible to researchers via the Oxford Text Archive and the ESRC Archive. Each text will be made anonymous before being deposited in these archives. Other interested parties such as writing tutors and students may be granted limited access to parts of the corpus.

The primary purpose of the corpus will be to enable us to identify the characteristics of student writing, and compare the writing produced in different disciplines and at different levels of study. Corpus samples may also eventually be used in teaching materials. For example, CELTE intends to use BAWE data in a forthcoming CD-ROM (*EASE: Developing Academic Writing Skills*).

The corpus will contain a wide variety of proficient assessed writing produced at postgraduate and undergraduate level, drawn from all faculties within the University. Corpus holdings will be categorised according to faculty, department, course module and year of study. Contributing students will be given £3.00 for each assignment accepted into the corpus, and will be required to sign a copyright disclaimer granting permission for the assignment to be used for the purposes outlined above.

Appendix 2genre and genre family

Genre Families	Social purpose/ Components/ Genre network	Genres (examples from each family)

Case Study	to gain an understanding of professional practice through the analysis of a single exemplar description of a particular case, often multifaceted, with recommendations or suggestions for future action typically corresponds to professional genres (e.g. in business, medicine, and engineering)	business start-up company report (starts with executive summary) investigation report organisation analysis patient case notes patient report single issue tourism report
Critique	to demonstrate understanding of and the ability to evaluate and / or assess the significance of the object of study includes descriptive account, explanation, and evaluation; often involves tests may correspond to part of a research paper, professional design specification or expert evaluation	academic paper review approach evaluation business environment analysis business / organisation evaluation financial report evaluation interpretation of results legislation evaluation (legal) case report policy evaluation product/ building evaluation programme evaluation project evaluation review of a book/ film/ play/ website system evaluation teaching evaluation
Design Specification	to demonstrate the ability to design a product or procedure that could be manufactured or implemented typically includes purpose, component selection, and proposal; may include development and testing of design may correspond to a professional design specification, or to part of a proposal or research report.	application design building design database design game design label design product design system design website design

Empathy writing	to demonstrate understanding and appreciation of the relevance of academic ideas by translating them into a non-academic register, to communicate to a non-specialist readership may be formatted as a letter, newspaper article or similar non-academic genre may correspond to professional writing	expert information for journalist expert advice to industry expert advice to lay person information leaflet job application letter (e.g. reflective letter to a friend; business correspondence) newspaper article
Essay	to develop the ability to construct a coherent argument and develop critical thinking skills may be discussion (issue, pros/cons, final position); exposition (thesis, evidence, restate thesis); factorial (outcome, conditioning factors); challenge (opposition to existing theory); comparison (series of comparative points or arguments); or commentary (series of comments on a text) may correspond to a published academic/specialist paper	challenge commentary comparison discussion exposition factorial
Exercise	to provide practice in key skills (e.g. the ability to interrogate a database, perform complex calculations, or explain technical terms or procedures), and to consolidate knowledge of key concepts data analysis or a series of responses to questions may correspond to part of report or research paper	calculations data analysis mixed (e.g. calculations + essays) short answers stats exercise

Explanation	to demonstrate understanding of the object of study; and the ability to describe and/or assess its significance includes descriptive account, explanation may correspond to a published explanation, or to part of a research paper or professional design specification	business overview concept /job/ legislation overview instrument overview methodology overview organism / disease overview product development overview site/ environment overview species / breed overview substance / phenomenon overview system/ process overview
Literature Survey	to demonstrate familiarity with literature relevant to the focus of study includes summary of literature relevant to the focus of study and varying degrees of critical evaluation may correspond to a published paper or anthology, or to part of a research paper	annotated bibliography anthology literature review notes taken from multiple sources summary book chapter summary series of articles
Methodology Recount	to become familiar with disciplinary procedures and methods, and additionally to record experimental findings describes procedures undertaken by writer may include Introduction, Methods, Results, and Discussion sections, or these functions may be realised iteratively may correspond to a section within a research report or research paper	computer analysis data analysis report experimental report field report forensic report lab report materials selection report (program)development report

Narrative Recount	to develop awareness of motives and/or behaviour in individuals (including self) or organisations fictional or factual recount of events, with optional comments may correspond to published literature, a professional proposal or a report, or to part of a research paper	accident report account of literature search account of website search biography character outline creative writing: short story plot synopsis reflective recount report on disease outbreak urban ethnography
Problem question	to practice applying specific methods in response to simulated professional problems problem (may not be stated in assignment), application of relevant arguments or presentation of possible solution(s) in response to scenario problems or situations may resemble or be based on real legal, engineering, accounting or other professional cases	law problem question logistics simulation medical problem
Proposal	to demonstrate ability to make a case for future action includes purpose, detailed plan, persuasive argumentation may correspond to professional or academic proposals	book proposal building proposal business plan catering plan legislation reform marketing plan policy proposal procedural plan research proposal

Research Report	to demonstrate ability to undertake a complete piece of research including research design, and an appreciation of its significance in the field may include Literature Review, Methods, Findings, Discussion; or may include several 'chapters' relating to the same theme may correspond to a published experimental research paper or topic-based research paper	research paper topic-based dissertation

Appendix 3example of a document header

```
<teiHeader>
<fileDesc>
     <titleStmt>
          <title>
               Nutrition in Health &amp; Disease (Week 5 Q&amp;A)
          </title>
     </titleStmt>
     <extent/>
     <publicationStmt>
          <distributor>British Academic Written English (BAWE)
          corpus</distributor>
          insert availability statement
     </publicationStmt>
     <notesStmt>
          <note resp="British Academic Written English (BAWE) corpus
          project">
               Page header contains: student name; date.
               Page footer contains: page number.
          </note>
          <note resp="British Academic Written English (BAWE) corpus
          project">
               Word count deleted at the end of the document.
          </note>
          <note resp="British Academic Written English (BAWE) corpus
          project">
               Evaluated as candidate compound assignment. Assigned to
               S3: assignment structured by questions: non-compound.
          </note>
     </notesStmt>
     <sourceDesc>
          <p n="level">4</p>
          <p n="date">2005-10</p>
          <p n="module title">Human Nutrition in Health &amp;
          Disease</p>
          <p n="module code">FB3N2</p>
          <p n="assignment type">exercise</p>
          <p n="discipline">Food Sciences</p>
          <p n="disciplinary group">LS</p>
          <p n="grade">M</p>
```

```
          <p n="number of authors">1</p>
          <p n="number of words">775</p>
          <p n="number of s-units">34</p>
          <p n="number of p">10</p>
          <p n="number of tables">0</p>
          <p n="number of figures">1</p>
          <p n="number of block quotes">0</p>
          <p n="number of formulae">1</p>
          <p n="number of lists">0</p>
          <p n="number of paragraphs formatted like lists">0</p>
          <p n="abstract present">no abstract</p>
          <p n="average words per s-unit">22.8</p>
          <p n="average s-units per p">3.4</p>
          <p n="macrotype of assignment">complex but non-compound
          assignment (see notesStmt for details)</p>
     </sourceDesc>
</fileDesc>
<encodingDesc>
     <p>TEI P4 (documented in: BAWE-documentation.Markup.pdf)</p>
</encodingDesc>
<profileDesc>
     <particDesc>
     <person>
          <p n="gender">m</p>
          <p n="year of birth">1982</p>
          <p n="first language">eng</p>
          <p n="education">UKa</p>
          <p n="course">Food Science</p>
          <p n="student ID">6085</p>
     </person>
     </particDesc>
</profileDesc>
</teiHeader>
```

Appendix 4tagset used in the BAWE corpus

This section provides an empirical list of elements (tags), their possible parent and child elements, attributes they may have and their values used in the BAWE corpus. They are grouped together in three lists: tags occurring within front, tags occurring within body and tags occurring within back.

Tags used in the header are not included in this enumeration; an example of a header is provided in appendix 1.

Tags occurring within front

element	may occur within	may contain	attributes	values
div1	front	head, p	n	1, 2, 3, 4

element	may occur within	may contain	attributes	values
			type	front-back-matter, front-text, toc
docTitle	titlePage	figure, titlePart	(no attributes)	
figure	docTitle, front	head	id	[figure ID nb]
front	text	div1, figure, titlePage	(no attributes)	
head	div1, figure	hi, seg	rend	bold, bold italic, underlined, underlined bold
hi	head, hi, p, quote, titlePart	hi, name, ref	rend	bold, bold italic, italic, sup, underlined, underlined bold
name	hi, p, titlePart	(no child elements)	type	other, student ID, student name, tutor name, university
note	titlePart	p	id	[note ID nb]
			place	foot
p	div1, note	hi, name	rend	bold, italic, underlined bold
quote	titlePart	hi	(no attributes)	
ref	hi	(no child elements)	target	[note ID nb]
seg	head, titlePart	(no child elements)	n	[hex value or entity name of special character]
text		front	(no attributes)	
titlePage	front	docTitle	(no attributes)	
titlePart	docTitle	hi, name, note, quote, seg	rend	bold, bold italic, italic, underlined, underlined bold, underlined bold italic, underlined italic
			type	main, quote, secondary

Tags occurring within body

element	may occur within	may contain	attributes	values
body	text	div1	(no attributes)	
cell	row	(no child elements)	(no attributes)	
div1	body	div2, figure, head, list, note, p, quote, table	n	1, 2, 3
			type	abstract, front-back-matter, section, text
div2	div1	div3, figure, head, list, note, p, quote, table	type	front-back-matter
div3	div2	div4, figure, head, list, p, quote, table	(no attributes)	
div4	div3	div5, figure, head, list, p, table	(no attributes)	
div5	div4	head, p	(no attributes)	
figure	div1, div2, div3, div4	head	id	[figure ID nb]
formula	head, item, s	(no child elements)	id	[formula ID nb]
			notation	[empty]
head	div1, div2, div3, div4, div5, figure, list, table	formula, hi, list, note, ref, seg	rend	bold, bold italic, italic, sup underlined bold, underlined, underlined bold, underlined bold italic, underlined italic
hi	head, hi, item, p, quote, s	hi, name, seg	rend	bold, bold italic, italic, sub, sub italic, sup, sup bold, sup bold italic, sup italic, underlined, underlined bold, underlined bold italic, underlined italic

element	may occur within	may contain	attributes	values
body	text	div1	(no attributes)	
item	list	formula, hi, name, quote, ref, seg, table	rend	bold, italic, underlined, underlined bold, underlined italic
list	div1, div2, div3, div4, head, note	head, item	type	bulleted, ordered, simple
name	hi, item, s	(no child elements)	type	other, student name, tutor name, university
note	div1, div2, head	list, p	id	[note ID nb]
			place	end, foot
p	div1, div2, div3, div4, div5, note	hi, ref, s	n	[p nb]
			rend	bulleted, ordered
quote	div1, div2, div3, item	hi, ref, seg	lang	English-non-std., French, Greek, Latin
			rend	italic
ref	head, item, p, quote, s	(no child elements)	target	[note ID nb]
row	table	cell	(no attributes)	
s	p	formula, hi, name, ref, seg	n	[s nb]
			rend	bold, bold italic, italic, sup, sup italic, underlined, underlined bold, underlined bold italic, underlined italic
seg	head, hi, item, quote, s	(no child elements)	n	[hex value or entity name of special character]
table	div1, div2, div3, div4, item	head, row	id	[table ID nb]

Tags occurring within back

element	may occur within	may contain	attributes	values
back	text	div1	(no attributes)	
div1	back	div2, figure, head, p	type	appendix, missing or external appendix, bibliography, front-back-matter
div2	div1	figure, head, note, p	type	bibliography
figure	div1, div2	head	id	[figure ID nb]
formula	p	(no child elements)	id	[formula ID nb]
			notation	[empty]
head	div1, div2, figure	hi, note, ref	rend	bold, bold italic, italic, underlined, underlined bold, underlined bold italic, underlined italic
hi	head, hi, p	hi, seg	rend	bold, bold italic, italic, sub, sup, underlined, underlined bold, underlined italic
name	p	(no child elements)	type	other, student name, tutor name, university
note	div2, head	p	id	[note ID nb]
			place	foot
p	div1, div2, note	formula, hi, name, ref, seg	rend	bold, bold italic, italic, underlined, underlined bold, underlined italic
ref	head, p	(no child elements)	target	[note ID nb]
seg	hi, p	(no child elements)	n	[hex value or entity name of special character]
text		back	(no attributes)	

Appendix 5'sentence splitting' algorithm

Running text (within sections of the body) is structured in **p** and, further, **s** elements. **s** elements reflect an attempt to structure the text into sentence-like units. However, as this is done automatically, the **s** boundaries identified may not in all cases be identical with real sentence boundaries. The algorithm implemented is explicated in this section.

S-units are created by the following rules:

- The start of a paragraph of running text also marks the start of an **s** element.
- The end of a paragraph of running text also marks the end of an **s** element.
- A strong punctuation sign (dot, question mark and exclamation mark) within a paragraph of running text is marks an **s** boundary if it is followed by either:

  a space character + a capital letter or number; or
  a single or double quote + a space + a capital letter or number; or
  a space + a single or double quote + a capital letter or number; or
  a space + a dash; or
  a single or double quote + a space + a dash; or
  a space + a single or double quote + a dash; or
  a closing bracket + a space + a capital letter or number; or
  a space + an opening bracket + a capital letter or number; or
  a single or double quote + a space + an opening bracket + a capital letter or number; or
  a single or double quote + a space + a single or double quote + a capital letter or number; or
  a space + a **hi** element; or
  an optional space + the end of a **hi** element; or

a single or double quote + the end of a **hi** element;
or
a space + a single or double quote + a **hi** element;
or
an optional space + a **ref** element; or
a single or double quote + an optional space + a **ref** element; or

- However, no **s** boundary is marked in cases of the following abbreviations followed by a dot:

app.	Fr.	Mrs.	St.
Av.	frk.	Ms.	Sta.
dr.	Frk.	op. cit.	Sto.
Dr.	Hon.	p.	W.C.
Dra.	hr.	pp.	z. B.
e. Kr.	Hr.	Prof.	z.B.
et. al.	ibid.	Sr.	
f. Kr.	Ibid.	Sra.	
fr.	Mr.	Srta.	

- No s-boundary is introduced in the sequence: capital letter + dot + space + capital letter.

The class of capital letters is defined as:

A-Z Æ Ø Å É Ê Ë È Ä Á Â À Í Î Ï Ì Ö Ó Ô Ò Ü Ú Û Ù Ç

N.b. there are no s-units in **front** and **back**, nor in text contained either by **quote**, **list|item** or **head**.

The Perl script used for splitting the text into s-units was provided by Dr Jarle Ebeling and has been adapted for BAWE markup.