BAWE: an introduction to a new resource

Hilary Nesi
Coventry University – UK

The British Academic Written English (BAWE) corpus was developed with ESRC funding as part of the project entitled ‘An investigation of genres of assessed writing in British Higher Education’ (2004-2007). The project aimed to identify the characteristics of proficient student writing, and to compare these across disciplines and levels of study. The corpus consists of just under 3000 student assignments of a good standard (6,506,995 words), at all levels from first year undergraduate to taught masters degree, and in many disciplines. Information about discipline and level is provided in the header for each assignment file, alongside other types of contextual information which did not influence collection policy such as gender, year of birth, native speaker status, and years of UK secondary education. We believe that BAWE is currently the only complete corpus of its kind in the public domain. It offers opportunities to investigate student writing which has been judged to conform to departmental requirements, but which differs markedly from expert and near-expert academic writing in terms of its communicative intent.

Keywords: academic, assignment, essay, EAP, genre

Background to the project

The project ‘An investigation of genres of assessed writing in British Higher Education’ grew out of a concern that too little was known about the types of writing students produced in British universities, and a concern that inappropriate genre models were used for academic writing courses.

The research article is as popular a genre for analysis today (e.g. Ozturk, 2007; Bruce, 2008) as in the 1980s (e.g. Swales 1983, 1984). The discourse of doctoral theses has also been investigated fairly thoroughly (e.g. Thompson, 2005; Charles, 2006). This focus on published articles and theses is understandable, since they represent the standard many academic writers aspire to, and they are readily available in the public domain. Nevertheless they do not represent the bulk of what is written in academic contexts, i.e. the texts produced by students on taught degree programmes, for assessment, generally with the intention of demonstrating academic knowledge and skills as opposed to presenting research findings.

Of course the university assignment is not an entirely neglected genre, and there have been a number of excellent studies of small collections of student writing, usually within jusr one or two disciplines and with reference to one particular discourse feature (see, for example, Woodward-Kron, 2002; North, 2005). Before the development of the BAWE corpus, however, no fully documented collection existed which might enable large scale comparisons of assignments across disciplines and levels of study. Two such corpora are under development in the United States (the Michigan Corpus of Upper-level Student Papers (MICUSP), and the ‘Viking’ corpus at Portland State University), but at the time of writing both of these contain less than a million words.

Our initial attempt to create a small corpus of student assignments was not entirely successful, and provided some insight into why such a corpus did not yet exist. Our pilot project ran from May 2001 to November 2002, during which time we collected 499 assignments from 70 student writers. The contributors, however, tended to come from a limited range of disciplines (largely from the humanities, with very few from the hard sciences) and there was a disproportionate number of assignments from the first year of study (44%) (see Nesi, Sharpling and Ganobcsik-Williams, 2004). The project did not adopt any particular collection policy, and simply accepted any assignment offered by any willing student. This helps to explain why the hard sciences and the later years of study were not well represented, as fewer scientists were interested in contributing, they produced less written work, and there was diminishing availability of assignments in the upper levels (students could contribute work written in preceding years, but could not contribute work that had not yet been assessed). It was evident that it would be necessary to devise a more systematic approach to data collection to fulfil the aims of the main project, which received funding from the ESRC in 2004..

For this project we proposed to integrate ethnographic, multidimensional and functional linguistic approaches to text description, each of which suggested a different method of sampling (as discussed in Gardner, forthcoming). Ethnographic aspects of the study favoured cluster sampling and the targetting of specific university discourse communities, but random sampling seemed an appropriately objective way of collecting data for computational analysis, and purposive sampling, involving the targetting of specific text types, promised to provide the richest array of data for genre analysis.

Our final collection policy involved stratified sampling, a compromise which took into account these conflicting approaches to corpus analysis, together with the practical constraints on policy implementation. We did conduct interviews with staff and students (see Nesi and Gardner, 2006; Gardner and Powell, 2006), but we rejected the idea of sampling selected clusters of contributors because we did not have the resources (or the persuasive power) to guarantee contributions from sufficient numbers of individuals within specified departmental communities. We considered random sampling, but even if it had been possible to identify a random sample of potential student contributors, our experience with the pilot corpus had taught us that it would be impossible to force contributions from them. We abandoned more purposive sampling, although we wanted to gather several instances of each assignment type we encountered, because it soon became clear that it would be impossible to create a multi-million word corpus if we set restrictions on the genre of contributions, as well as on their grade, discipline and year of study.

Corpus holdings

We used a 4-by-4 matrix to guide data collection. This combined four years of study with four broad disciplinary groupings, and we intended to fill each of the 16 cells with a roughly equal quantity of assignments, rejecting all but a few contributions which were superfluous to these requirements (we retained an ‘other’ category, to round up numbers). The following table represents our ideal corpus structure in more detail, and our plan to collect 3,500 assignments across 28 disciplinary fields.

Disciplinary Group	**Subject**	**Per Year** (1, 2, final, and Masters level)	**Total**
Arts & Humanities	Applied Linguistics/Applied English Language Studies	32	128
	Classics	32	128
	Comparative American Studies	32	128
	English Studies	32	128
	History	32	128
	Philosophy	32	128
	(Archaeology)	16	64

Disciplinary Group	Subject	Per Year (1, 2, final, and Masters level)	Total
Life Sciences	Agriculture	32	128
	Biological Sciences/ Biochemistry	32	128
	Food Science and Technology	32	128
	Health and Social Care	32	128
	Plant Biosciences	32	128
	Psychology	32	128
	(Medical Science)	16:48	64
Physical Sciences	Architecture	32	128
	Chemistry	32	128
	Computer Science	32	128
	Cybernetics & Electronic Engineering	32	128
	Engineering	64	256
	Physics	32	128
	(Mathematics)	16	128
Social Sciences	Anthropology	32	128
	Business	32	128
	Economics	32	128
	Hospitality, Leisure and Tourism Management,	32	128
	Law	32	128
	Sociology	32	128
	(Publishing)	16	64
Other	Other	43	172
Total			**3500**

Table One: the plan for BAWE corpus collection.

Our matrix was not designed to represent proportionally the quantity of writing produced in each discipline and at each level, or to ensure perfect representation of all the genres produced in the target disciplines. Students usually write more in their final year(s), and some disciplines are understood to be more discursive than others (as indicated in British university rules concerning PhD thesis length – usually a maximum of 80,000 words in the Humanities and Social Sciences, but only 50,000 words in the Sciences). Also we knew we could not collect assignments for every module in every discipline, and that module tutors were liable at any time to introduce new tasks with different generic expectations. We realized we might miss some unusual genres, especially if only a few students selected a particular writing task, or if they received low grades (we only accepted assignments graded 60% or above). Nevertheless steps were taken to encourage variety in the corpus in terms of both assignment type and authorship, by prompting contributors to submit additional work belonging to a different genre, if possible, whilst preventing individuals from contributing more than three assignments from any single module.

Assignments were collected at Oxford Brookes, Reading and Warwick, and, in the final year of the project, Coventry University (to make up numbers in disciplines which still lacked sufficient contributions). Most cells of our matrix were not quite filled, as can be seen from Table Two.

Disciplinary Grouping		Yr 1	Yr 2	Yr 3	Masters	Total
Arts and Humanities	students	101	83	61	23	268
	assignments	239	228	160	78	705
	texts	254	232	160	82	728
	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	186	203	92	246	727
	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	201	156	159	121	637
	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	215	205	165	210	804[3]
	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		**856**	**796**	**576**	**659**	**2896**[3]
Total words		**1,440,185**	**1,781,686**	**1,558,715**	**1,704,015**	**6,506,995**[4]

[1] Includes 3 students of unknown level.

[2] Includes 9 assignments of unknown level.

[3] Includes 9 texts of unknown level.

[4] Includes 22,394 words of unknown level

Table Two: numbers of students, assignments, texts and words by grouping and year.

The number of texts recorded in the table exceeds the number of assignments, because some assignments turned out to consist of more than one independent text, submitted together to receive a single grade.

Table Three provides a more complete picture of the disciplines represented in the corpus. In this table ‘discipline’ is not synonymous with ‘department’, because some assignments in the same field came from more than one university, and departments with slightly different names have been conflated (*Computer Science* and *Computing*, for example). We recognize that ‘discipline’ is a difficult concept to define, however, and that ‘variation in epistemology and discourse occurs not only across disciplines, but also within disciplines’ (Nesi and Gardner, 2006: 101).

Disciplinary Grouping	**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
	Total	**180**	**193**	**82**	**228**	**683**

Disciplinary Grouping	Discipline	1	2	3	4	Total
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
	Total	**181**	**149**	**155**	**111**	**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.

Table Three: number of assignments by discipline and year

The corpus was encoded according to the guidelines of TEI P4 (Sperberg-McQueen and Burnard, 2004), but since the TEI standard was devised for a wide range of texts, a special DTD containing only a subset of all TEI elements and attributes was created for

BAWE (see Heuboeck, Holmes and Nesi, 2008). Information of the following types was encoded:

- header information
- document structure and hierarchy
- types of front and back matter
- functional features within running text
- character formatting
- anonymized personal information (related to student, university or third parties)

The header provides information about the discipline and level of each assignment, alongside other types of contextual information which did not influence collection policy. For example although we have recorded the gender and the first language of each contributor, gender proportions vary from cell to cell, and the proportion of non-native speakers is much greater in some disciplines, and at Masters level. In the British university context a contributor's choice of first language sometimes reflects affiliation rather than proficiency, so in view of this we also recorded the number of years of UK secondary education each contributor had received. Header information concerning first language, secondary education, and assignment grade (merit or distinction, corresponding to first or upper second class degree level) can thus be used to filter assignments according to individual requirements; some researchers want a sub-corpus of native speaker assignments at distinction level, for example, presumably because they view this as being in greatest conformity with the norms of the British academic discourse community.

Findings

The following broad 'genre families' were identified in the corpus:

Case Study: A description of a particular case with recommendations or suggestions for future action, written to gain an understanding of professional practice (e.g. in business, medicine, or engineering).

Critique: A text including a descriptive account, explanation, and evaluation, often involving tests, written to to demonstrate understanding of the object of study and to demonstrate the ability to evaluate and / or assess the significance of the object of study.

Design Specification: A text typically including an expression of purpose, an account of component selection, and a proposal; and possibly including an account of the development and testing of the design.

Empathy writing: A letter, newspaper article or similar non-academic genre, written to demonstrate understanding and appreciation of the relevance of academic ideas by translating them into a non-academic register, for a non-specialist readership.

Essay: A discussion, exposition, factorial, challenge or commentary, written to develop the ability to construct a coherent argument and develop critical thinking skills.

Exercise: Data analysis or a series of responses to questions, written to provide practice in key skills and to consolidate knowledge of key concepts.

Explanation: A descriptive account and explanation, written to demonstrate understanding of the object of study and the ability to describe and/or assess its significance.

Literature Survey: A summary including varying degrees of critical evaluation, written to demonstrate familiarity with the literature relevant to the focus of study.

Methodology Recount: A description of procedures undertaken by the writer, possibly including Introduction, Methods, Results, and Discussion sections, written to develop familiarity with disciplinary procedures and methods, and additionally to record experimental findings.

Narrative Recount: A fictional or factual recount of events, written to develop awareness of motives and/or the behaviour of organisations or individuals (including oneself).

Problem question: A text presenting relevant arguments or possible solution(s) to a problem, written to practise the application of specific methods in response to simulated professional scenarios.

Proposal: A text including an expression of purpose, a detailed plan, and persuasive argumentation, written to demonstrate the ability to make a case for future action.

Research Report: A text typically including a Literature Review, Methods, Findings, and Discussion, or several 'chapters' relating to the same theme, written to demonstrate the ability to undertake a complete piece of research, including research design, and to appreciate its significance in the field.

One obvious conclusion that can be drawn from this categorisation scheme is that university students write for a range of purposes, not all of them identical to the purposes of academics. Some assignments are generically similar to texts produced in the professions, but only the Research Report bears much generic resemblance to the thesis or research article.

The distribution of the genre families in the corpus is presented in Table Four. The essay is the best represented category, although in the Physical and Life Sciences it is outnumbered by submissions belonging to other genre families (Methodology Recounts, Design Specifications, and Critiques). Also, some genre families are rare or totally absent from some disciplinary groupings, particularly the Arts and Humanities.

	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 Survey	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

Table Four: Distribution of genre families by disciplinary group

Multidimensional analysis revealed the corpus to be carefully written and information-rich, but there were also significant differences among genre families, as can be seen from Table Five. The entirely negative scores on the ‘involved’ and ‘narrative’ dimensions indicate a high informational focus and a low level of narration, whilst the entirely positive scores for ‘explicit’ and ‘abstract’ qualities indicate lexically dense text containing passives, past participial clauses, and other features typical of academic prose. Mixed scores on the ‘persuasive’ dimension, however, indicate variation in the degree of argumentation (Proposals being the most persuasive, and Literature Surveys the least). Student writing simply does not need to ‘create a research space’ in the manner of research article introductions, because the centrality of the topic is not usually in question, and the tutor is duty-bound to read the text.

	Involved	**Narrative**	**Explicit**	**Abstract**	**Persuasive**
Essay	-14.327	-2.4788	6.234	5.920	-1.8345
Methodology Recount	-15.856	-3.6533	4.506	7.304	-2.5011
Critique	-14.833	-3.0714	5.988	6.381	-1.6127
Explanation	-15.411	-3.5878	5.042	5.848	-2.2744
Case Study	-16.402	-2.8617	5.772	4.450	-0.4519
Exercise	-12.098	-3.8543	4.628	5.678	-1.3301
Design Specification	-13.090	-4.0223	4.079	6.750	0.6702
Proposal	-16.421	-3.7855	6.326	4.793	1.2799
Narrative Recount	-4.818	-1.1128	3.814	3.957	-0.7439
Research Report	-16.186	-3.1156	5.524	7.198	-2.4064
Problem Question	-11.950	-2.7730	5.222	6.429	1.6295
Literature Survey	-17.907	-2.6214	6.311	5.047	-3.4343
Empathy Writing	-11.500	-2.7369	4.533	4.472	0.7713

Table Five: Multiple Range Test Scores for Genre Families

Multidimensional analysis also revealed significant differences between the four disciplinary groupings in terms of their information load, and significant differences between first and final year undergraduate assignments on all but the ‘persuasive’ dimension.

Conclusion

Clearly the BAWE corpus is a very rich resource, offering a currently unique opportunity to investigate thousands of academic texts which have been judged to conform to departmental requirements (on the evidence of the grade awarded), but which differ markedly from professional academic writing in terms of their communicative intent. Several close analyses of the corpus are planned or in press, and proposals for further investigations will be welcomed by the research team.

Acknowledgements

The project *An Investigation of Genres of Assessed Writing in British Higher Education,* including the development of the British Academic Written English corpus, was funded by the Economic and Social Research Council (project number RES-000-23-0800), under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for English Language Teacher Education, Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes). Other members of the project team were Sian Alsop, Dawn Hindle, Maria Leedham, Signe Ebeling, Alois Heuboeck, Jasper Holmes, Richard Forsyth and Laura Powell. The multidimensional analysis was conducted by Doug Biber and his team at Northern Arizona University.

References

Bruce, I. 2008. “Cognitive genre structures in Methods sections of research articles: A corpus study” *Journal of English for Academic Purposes* 7/1: 38-54

Charles, M. 2006. “Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines” *English for Specific Purposes* 25/3: 310-331

Gardner, S. Forthcoming. “Integrating ethnographic, multidimensional, corpus linguistic and systemic functional approaches to genre description: an illustration through university history and engineering assignments”. *Proceedings of the 19th European Systemic Functional Linguistics Conference and Workshop*, Universität des Saarlandes, Saarbrücken, July 2007.

Gardner, S. and Powell, L. 2006. ‘An investigation of genres of assessed writing in British Higher Education’. Paper presented at the annual seminar *Research, Scholarship and Practice in the area of Academic Literacies*, University of Westminster, 30 June 2006. http://www.coventry.ac.uk/researchnet/external/content/1/c4/33/84/v1193312407/user/genresbhe_handout.pdf [Access date 27/05/2008].

Heuboeck, A., Holmes, J. and Nesi, H. 2008 *The BAWE Corpus Manual*. http://www.coventry.ac.uk/researchnet/external/content/1/c4/51/60/v1212053950/user/BAWE.pdf [Access date 27/05/2008].

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 and G. Clibbon (eds.), *Language, Culture and Identity in Applied Linguistics*, British Studies in Applied Linguistics Vol. 21. London: Equinox Publishing, 99-117.

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” *Proceedings from the Corpus Linguistics Conference* Series 1/1. www.corpus.bham.ac.uk/PCLC/NesiStudentWriting.doc [Access date 27/05/2008].

Nesi, H, Sharpling, G. and **Ganobcsik-Williams, L.** 2004. “The design, development and purpose of a corpus of British student writing” *Computers and Composition* 21/4: 439-450.

North, S. 2005. “Different values, different skills? A comparison of essay writing by students from arts and science backgrounds” *Studies in Higher Education* 30/5: 517-533

Ozturk, I. 2007. “The textual organisation of research article introductions in applied linguistics: Variability within a single discipline” *English for Specific Purposes* 26/1: 25-38.

Sperberg-McQueen, C. M. and Burnard, L. (eds.). 2004. *TEI P4 – Guidelines for Electronic Text Encoding and Interchange, XML-compatible edition.* http://www.tei-c.org/P4X/ [Access date 27/05/2008].

Swales, J. M. 1983. “Developing materials for writing scholarly introductions”. In *Case Studies in ELT*, R. R. Jordan (ed.). London: Collins ELT.

Swales, J. M. 1984. “Research into the structure of introductions to journal articles and its application to the teaching of academic writing”. In *Common Ground: shared interests in ESP and communication studies*, R. Williams, J. Swales & J. Kirkman (eds.) Oxford: Pergamon Press.

Thompson, P. 2005. **“**Points of focus and position: Intertextual reference in PhD theses” *Journal of English for Academic Purposes* 4/4: 307-323

Woodward-Kron, R. 2002. **“**Critical analysis versus description? Examining the relationship in successful student writing” *Journal of English for Academic Purposes* 1/2: 121-143

The author

Hilary Nesi joined Coventry University as Professor in English Language in October 2007, having worked for twenty years in the Centre for English Language Teacher Education at the Unversity of Warwick. She led the project to create the BASE corpus of British Academic Spoken English (2001–2005) and also the ESRC funded project 'An Investigation of Genres of Assessed Writing in British Higher Education' (2004-2007), which involved the creation of the BAWE corpus. She is chief academic consultant for *the Essential Academic Skills in English (EASE) series of multimedia interactive self-access materials on CD-ROM, based around video clips of authentic academic discourse drawn from the BASE corpus, and she is currently involved in several projects relating to the use of English as a medium of instruction in international higher education.*