Introducing BAWE: A New Lexicographical Resource

Hilary Nesi
Coventry University

This paper reports on the compilation of the British Academic Written English (BAWE) corpus, a collection of almost 3,000 proficient student assignments produced at three representative universities in the UK. BAWE was designed to fill a gap in current corpus resources by complementing other writing collections which represent expertly written academic text-such as the TOEFL 2000 Spoken and Written Academic Language Corpus, or non-expert and non-discipline specific student writing-such as the Louvain Corpus of Native English Essays, and the Cambridge Syndicate Examination corpus. Prior to the development of BAWE the few small corpora of writing produced by university students within their disciplines had either been compiled for individual scholarly purposes, or were in the form of inadequately documented and unannotated "essay banks" for student use. The BAWE corpus, in contrast, is a large, formally compiled collection of assignments at four levels of study, from first year undergraduate to masters level, accompanied by detailed contextual information. Thirteen broad macrogenres have been identified in the corpus, including the essay, of course, and writing generically similar to the published research article, but also including other types of writing, neglected in the literature, which reflect the purpose of university level study. The full corpus will be freely available to researchers from January 2008, and it is foreseen that it will provide a-currently unique-resource for designers of dictionaries for advanced learners, particularly those learners studying at university level in the medium of English.

Introduction

The BAWE corpus contains almost 3,000 good-quality examples of university student writing produced in 33 disciplines, across four years of study (first year undergraduate to masters level). Developed with ESRC funding1. following on from the success of a pilot version (Nesi, Sharpling & Ganobcsik-Williams 2004), it aims to fill a gap in corpus provision and aid identification of linguistic features typical of university student writing.

Prior corpora of academic writing

Dictionaries for advanced learners of English have been corpus-based for twenty years now, and there is a growing tendency for lexicographers to refer to smaller, more specialized resources to complement large general mixed reference corpora, and to provide deeper insights into language use in specific domains. Notes on academic English usage are a particularly important feature of advanced learners' dictionaries, because English is so widely used as an academic lingua franca, and because of the growing popularity of English-medium degree programmes (in EFL as well as ESL and L1 contexts).

Examples of professionally edited and expertly written academic text are easy to obtain and are already present in all the major proprietary corpora, although they are not always distinguished from other types of generally "informative" published texts. Such texts may represent the type of writing students are likely to encounter in their studies, but they do not represent the type of writing they are likely to produce. Until now, however, lexicographers have not had the chance to access good quality student writing, representative of the

[1] As part of a project entitled *An investigation of genres of assessed writing in British Higher Education* (RES-000-23-0800, 2004-2007).

assignments set for degree programmes in English-medium universities. The nearest available approximations to a resource of this type have been corpora designed for the study of language acquisition processes (learner corpora).

The first and probably the best-known learner corpus is the *International Corpus of Learner English* (ICLE), which contains essays on general topics produced by advanced learners of English studying in a non-English-medium environment. Insights from ICLE have been incorporated into the *Macmillan English Dictionary for Advanced Learners* (MED) (second edition)**,** in various types of MED usage note, and in an extended central section on academic writing (De Cock et al. 2007).

Two other larger and more heterogeneous learner corpora are the *Longman Learners' Corpus* (LLC), to which English language teachers around the world submit their students' writing, and the *Cambridge Learner Corpus* (CLC), which makes use of the thousands of exam scripts written by students taking Cambridge ESOL English examinations. Both these resources have been used extensively by Longman and Cambridge University Press when developing learners' dictionaries and related materials. LLC findings inform the Usage Notes in the *Longman Active Study Dictionary*, for example, and CLC findings are used in the 'Common Learner Error Notes' in the *Cambridge Advanced Learner's Dictionary.*

Learner corpora are extremely useful for identifying non-native speakers' overuse and underuse of lexical and grammatical items, especially when matched with 'control' corpora such as the Louvain Corpus of Native English Essays (LOCNESS), or informative texts from the BNC (Rundell & Granger 2007). Learner corpus research, however, focuses on language acquisition processes rather than analysis of academic discourse, and learner corpus holdings cannot be said to represent the range of genres university students will be expected to produce in English-medium educational contexts. Most learner corpus contributions were originally written for English language examinations, or as homework for English language classes. They therefore tend to take the form of argumentative essays on personal or general topics which do not require any preparation on the part of the writer. In contrast to language learning tasks, writing for academic purposes usually requires advance preparation, extensive referencing to extra-textual sources or data, and accommodation to the norms of a disciplinary discourse community.

The BAWE corpus

The BAWE corpus contains good quality assignments (receiving grades equivalent to an upper second or first class honours degree) in a wide range of disciplines (see Table One).

Arts & Humanities	Applied Linguistics/ Applied English Language Studies, Archaeology Classics, Comparative American Studies, English, History, Philosophy
Life Sciences	Agriculture, Biochemistry, Food Science and Technology, Health and Social Care, Medical Science, Plant Biosciences, Psychology
Physical Sciences	Architecture, Chemistry, Computer Science, Engineering, Mathematics, Physics
Social Sciences	Anthropology, Business, Economics, Hospitality, Leisure and Tourism Management, Law, Politics, Publishing, Sociology

Table 1. Target disciplines

Examination scripts were not collected. These are generally handwritten, and would have been expensive to convert to electronic format. Perhaps most importantly, even proficient examination answers rarely conform to departmental writing norms. This is because many examination candidates do not take time to attend in any detail to linguistic and organizational features, and tend to concentrate on information content.

Contributors completed a submission form detailing their gender, first language, department, degree course, assignment module, assignment title, grade, tutor and assignment year (1st, 2nd, 3rd/4th or taught postgraduate level). The first language of contributors, though recorded, was

not a factor in the corpus design, however. British universities are multicultural, multilingual environments, and in their departments students are assessed on merit without regard for their language background. All contributors were considered acceptably proficient users of English, given that their assignments had been awarded high grades.

Sampling decisions, explained more fully in Nesi et al., (2005) and Alsop & Nesi (under review), were taken in order to achieve roughly equal amounts of assignments at each level and in each of four disciplinary groupings: Arts and Humanities, Life Sciences, Social Sciences and Physical Sciences. The final corpus contains 2,761 assignments from 627 contributors, totalling 2,896 independent texts and 6,514,776 words. Each text in the corpus has been assigned to a "genre family" representing texts with a similar social purpose and rhetorical structure (see Table Two).

Genre family	Frequency	Range*	Examples
Essay	1225	24	Commentary, discussion, exposition
Methodology Recount	347	15	field report, forensic report, lab report
Critique	319	24	academic paper review, film review, financial report evaluation
Explanation	198	15	methodology review, disease overview, system overview
Case Study	194	12	organisation analysis, patient case notes, tourism report
Exercise	114	15	Calculations, data analysis, stats exercise
Design Specification	92	7	building design, product design, website design
Proposal	76	15	building proposal, marketing plan, research proposal
Narrative Recount	72	14	Biography, reflective recount, urban ethnography
Research Report	61	17	research paper, topic-based dissertation
Problem Question	40	7	law problem question, logistics simulation, medical problem
Literature Survey	35	11	annotated bibliography, anthology, summary
Empathy Writing	32	11	information leaflet, job application, newspaper article

Table 2. Genre families in the BAWE corpus.
*Across the 24 departments where 50 or more assignments have been collected

Essays are the most frequent type of genre in the corpus, as we might expect, but all the genre families are found within all the broad disciplinary groupings, except for Case Studies and Problem Questions, which were absent from Arts and Humanities disciplines. Rhetorical features vary greatly across the genre families, and structures and styles appropriate to one genre are not necessarily appropriate to another. Multidimensional analysis has been carried out on the corpus to examine variation across levels, disciplines and genres. There are significant differences in terms of progression across the four levels, and there are also significant differences between the four disciplinary groups in terms of their information load, with Arts and Humanities being most involved and Life Sciences being most informational. Genres were found to differ in terms of informational focus, density, persuasive features and levels of abstraction.

Practical applications

The BAWE corpus is currently a unique resource, although similar corpora are in preparation at the University of Michigan (the *Michigan Corpus of Upper-level Student Papers* (MICUSP)) and Portland State University (the *Viking Corpus of Student Academic Writing*, Conrad & Albers 2008).

When discussing the benefits of learner corpora for learner dictionary design, Rundell & Granger (2007) mention the importance of identifying *frequent* and *well-dispersed* language events, occurring in many different source texts: "This is what dictionary-writers are interested in describing, what teachers want to teach, and what students need to learn." The BAWE corpus has been carefully designed to enable the identification of features that are both common and wide-ranging, and it includes many types of assignment that students in English medium universities are likely to be required to produce, but which previous corpus and lexicographical studies have ignored.

References

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

Cambridge Advanced Learner's Dictionary. 3rd ed. Cambridge: Cambridge University Press, 2005.

Conrad, S.; Albers, S. (2008) "A new corpus of student academic writing". Paper presented at the *American Association for Corpus Linguistics Conference*, Brigham Young University, Utah.

De Cock, S. et al. (2007). "Improve your writing skills". In Rundell, M. (ed.). *Macmillan English Dictionary for Advanced Learners*. 2nd ed. Oxford: Macmillan Education. IW1-IW50.

Longman Active Study Dictionary. 4th ed. Harlow: Pearson Longman, 2007

Macmillan English Dictionary for Advanced Learners. 2nd ed. Oxford: Macmillan Education, 2007

Nesi, H. et al. (2005). *Towards the compilation of a corpus of assessed student writing: an account of work in progress*. Proceedings from the *Corpus Linguistics Conference Series.* Volume 1, issue 1. *www.corpus.bham.ac.uk/PCLC*.

Nesi, H.; Sharpling, G.; Ganobcsik-Williams, L. (2004). **"**The design, development and purpose of a corpus of British student writing". *Computers and Composition* 21 (4). 439-450).

Rundell, M.; Granger, S. (2007). "From corpora to confidence". *English Teaching Professional* 50. 15-18.