**Institution:** Lancaster University

**Unit of Assessment:** 29: English Literature and Language

**Title of case study:**

Corpus Research: Its Impact on Industry

**1. Summary of the impact** (indicative maximum 100 words)

UCREL (the University Research Centre for Computer Corpus Research on Language) has been pioneering advances in corpus linguistics for over 40 years, providing users with corpora (collections of written or spoken material) and the software to exploit them. Drawing together 8 researchers from the Department of Linguistics and English Language and 1 from the School of Computing and Communications at Lancaster University, it has enabled the UK English Language Teaching (ELT) industry to produce innovative materials which have helped the profitability and competitiveness of that industry, and assisted other, principally commercial, users to innovate in product design and development.

**2. Underpinning research** (indicative maximum 500 words)

The British National Corpus (BNC) Consortium worked from 1991-1994 to produce a 100-million word corpus of modern British English, for use in commercial, educational and academic research. The BNC contains 90% written and 10% spoken text, representing a wide cross-section of current British English. UCREL's work on the BNC was conducted by a team of 15-20 researchers led by Prof Geoffrey Leech (emeritus since 2002) and Roger Garside (senior lecturer, retired 2008) [R1]. The BNC consortium included Oxford University Press, Longman, Chambers Harrap, the British Library, and Oxford University Computing Services. It was published in 1994. A revised version was released worldwide in 2001. The *BNC XML Edition* (2007) is the version currently distributed by OUCS (http://www.natcorp.ox.ac.uk/).

A major contribution of UCREL to the BNC was the development of (a) many levels of corpus annotation – interpretative information (e.g. syntactic, semantic, pragmatic, anaphoric) in the form of searchable codes or tags; and (b) software which can (semi-)automatically add annotation to a corpus [R2]. This annotation allows users to take advantage of the levels of meaning in a corpus, as well as the levels of form. For example, "set" as a noun has rather different meanings from "set" as a verb, so limiting a search by part-of-speech makes it possible to tap into different groups of meanings. The BNC was grammatically annotated at Lancaster University [G1, G2] using CLAWS (*Constituent Likelihood Automatic Word-tagging System*), which UCREL has developed continuously since the early 1980s. CLAWS consistently achieves 96-97% accuracy across various types of English text.

The experience of creating and annotating the BNC fed into UCREL's central role in establishing cross-linguistic standards for corpora and multiple levels of corpus annotation, as seen in Leech's leading contribution (with researcher Andrew Wilson – now senior lecturer) to the European Commission's 1993-1999 EAGLES initiative [R3].

From inception to completion, UCREL's research has been shaped by engagement with non-academic users especially relating to education. Our research underpinning dictionaries, grammars [R4] and learning materials has involved collaboration with, for example, a range of publishers including Cambridge University Press, Oxford University Press and Pearson, and, more recently, we have also worked with companies delivering multilingual dictionary content on new media, notably mobile phones [G4], and with Trinity to help them develop their language testing business [G6].

From the late 1990s, UCREL researchers led by Tony McEnery (lecturer – now Professor) have also extended best-practices developed while working with the BNC to corpora of other languages.

Notable amongst these are (i) the EMILLE corpus, containing written and spoken data for fourteen South Asian languages [R5][G3] (developed by McEnery with Baker, lecturer now Professor, and Hardie, researcher now senior lecturer) and (ii) the Lancaster Corpus of Mandarin Chinese (developed by McEnery with Xiao, researcher now lecturer) extends the model of carefully-representative corpus design to Mandarin [R6]. The Lancaster Edition of the Callhome Chinese Corpus is one of very few openly-available corpora of spoken Chinese. More recently UCREL contributed expertise (McEnery and Hardie) to the creation and annotation of the Nepali National Corpus [G5].

**3. References to the research** (indicative maximum of six references)

All authors marked with an asterisk are currently, or were at the time of authoring the output, members of UCREL based in the Department of Linguistics and English Language at Lancaster University.

*Research outputs include (all available upon request)*:

[R1] Leech*, G., Garside, R. and Bryant*, M. (1994) "CLAWS4: the tagging of the British National Corpus", in *Proceedings of COLING 94 (the 15th International Conference on Computational Linguistics), Kyoto, 5-9 August, 1994*, pp.622-628.
[The COLING conference reviews full papers in blind peer review before accepting them. 134 citations. The collaborating author, Garside, was based at the Department of Computing, Lancaster University, UK.]

[R2] Garside, R., Leech*, G., and McEnery*, A. (eds.) (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.
[This book was subject to Longman's peer review and commissioning process. Additionally, all chapters in the book were peer reviewed prior to them being accepted for publication. 222 citations. The collaborating author, Garside, was based at the Department of Computing, Lancaster University, UK.]

[R3] Leech*, G. and Wilson*, A. (1999), 'Guidelines and standards for tagging', in H. van Halteren (ed.), *Syntactic Wordclass Tagging.* Dordrecht: Kluwer, pp.55-80.
[The output of the EAGLES project looking at part-of-speech annotation. The work in this chapter was subject to debate and review by the expert group who participated in EAGLES. This expert group was composed of leading corpus and computational linguists. 33 citations.]

[R4] Biber, D., Johansson, S., Leech*, G., Conrad, S. and Finegan, E. (1999) *Longman Grammar of Spoken and Written English* London:Longman.
[A major innovative grammar of the English language. The grammar has been widely and favourably reviewed. 3209 citations. The collaborating authors were based at: (i) the University of Northern Arizona, USA (Biber, Conrad), the University of Southern California, USA (Finegan) and (iii) the University of Oslo, Norway (Johansson).]

[R5] Baker*, P., Hardie*, A., McEnery*, A., Xiao*, R., Bontcheva, K., Cunningham, H., Gaizauskas, R., Hamza, O., Maynard, D., Tablan, V., Ursu, C., Jayaram, B.D. and Leisher, M. (2004) "Corpus linguistics and South Asian languages: corpus creation and tool development". *Literary and Linguistic Computing* 19(4): 509-524.
[The paper was subject to blind peer review prior to being accepted for publication. 20 citations. The collaborating authors were based at (i) the Central Institute of Indian Languages, Mysore, India (Jayaram); (iii) Computing Laboratory, New Mexico State University, USA (Leisher) and (ii) the Department of Computer Science, Sheffield University, UK (Bontcheva, Cunningham, Gaizauskas, Hamza, Maynard, Tablan, Ursu).]

[R6] Xiao*, R. and McEnery*, T. (2004) *Aspect in Mandarin Chinese*, Amsterdam: John Benjamins.
[This book was subject to John Benjamins' peer review and commissioning process, involving blind peer review of the final book manuscript. 88 citations.]

*Grants held as part of the underpinning research described include:*

| | Principal Investigator | Title [Funder] | Award £ | Period |
|---|---|---|---|---|
| [G1] | Leech | *Word-class tagging of the British National Corpus* [SERC and DTI] | 271,000 | 1991-1994 |
| [G2] | Leech | *British National Corpus: Tagging enhancement* [EPSRC] | 200,000 | 1994-1996 |
| [G3] | McEnery | *Enabling Minority Language Engineering* [EPSRC] | 280,000 | 2000-2003 |
| [G4] | McEnery | *Benedict – The New Intelligent Dictionary* [European Commission] | 300,000 | 2002-2005 |
| [G5] | McEnery | *Nepali Language Resources and Localization for Education and Communication* [European Union] | 27,000 | 2005-2008 |
| [G6] | McEnery | *Trinity Corpus Project* [Trinity College London and ESRC] | 577,037 | 2013-2018 |

All grants were awarded in open competition. Award was based upon selection via rigorous blind peer review. The grants listed are selected from over 40 funded projects run within UCREL during the period 1993 to 2013 (http://ucrel.lancs.ac.uk/projects.html).

**4. Details of the impact** (indicative maximum 750 words)

Corpora developed at Lancaster have had a notable impact on the English Language Teaching (ELT) industry. Most of this impact has been achieved via the BNC. The UK ELT industry has been estimated by the Department for Business, Innovation and Skills to be worth £1,996.2 million annually to the British economy while the ELT publishing industry adds another £304 million to that sum [C1]. The impact of our work on ELT is substantial and includes:

1. The Oxford Advanced Learner's Dictionary (OALD), a major basis for which is the BNC, has sold over 35 million copies worldwide and is the world's best-selling advanced learner's dictionary. It remains the best-selling title at Oxford University Press [C2]. OALD is now available as an app, website and CD Rom and is on its 8th edition with two million (hard) copies sold since the edition's 2010 release. The BNC is also used as a basis for OUP's web based teaching resources.
2. All Longman dictionaries are compiled using the Longman Corpus Network (which includes the BNC and the Lancaster/Longman corpus). This corpus network has been massively influential across the full range of Longman's ELT material including their learner dictionaries and Longman Language Activator series [C3]. Longman see the corpora as essential to their range, saying that they provide "the wealth of information for writing coursebooks and dictionaries that both accurately represent the English language and satisfy students' needs at every level." [C4] Longman-Pearson's ELT products are important to the company – in 2012 Longman-Pearson's ELT products led to its International Education division being both its source of its (i) highest growth in sales (10%) and (ii) greatest increase in operating profit (21%) [C5]. Between 1980 and 2011 Geoffrey Leech was the Vice Chair of Pearson's LINGLEX advisory group which advises Longman-Pearson on their ELT research and publication strategy.
3. The BNC was used by Cambridge University Press to update its Cambridge ESOL Business English, Key English and Preliminary English Test Vocabulary List. These lists are vital preparation for students taking the Cambridge English exams [C6], an important part of the UK ELT industry – 2010 saw entries for the Cambridge English exams grow to nearly 3.5 million, with more than 11,500 universities, employers and government departments worldwide recognising and using Cambridge English qualifications.

Fostering innovation in other, principally commercial, users, including:

1. The BNC is licensed to 1,581 institutions, including the following non-academic users: Toyota Motor Europe, Budweiser Budvar UK, Deutsche Bank, Canon Information Technology (Beijing) Co., BAE Systems and Ordnance Survey. Additionally 1,811 users access the BNC via

Lancaster's online interface, BNCweb. 192 of these users originate from non-academic bodies, including Microsoft, Sony, HSBC and the British Council.

2. In 2012 the Indian government launched a series of bilingual dictionaries which couple English with one of the following South Asian languages: Bengali, Hindi, Kannada, Malayalam, Oriya and Tamil. The 12,000 words and phrases covered by the dictionaries have been culled from the BNC. The South Asian language examples are furnished by corpora provided by the Central Institute of Indian Languages, who were helped in building the corpora by the Enabling Minority Language Engineering project at Lancaster. [G3][C7]
3. Kielikone Oy, a Finnish company, has customers in over 100 countries and is one of the world's leading providers of digital dictionary solutions. [C8] Kielikone uses the semantic taggers developed by Lancaster in almost all current services provided by them, including their MOT dictionary [C9]. This software has also been utilised in mobile applications of the dictionary [C10].
4. McEnery and Hardie have recently undertaken contracted work for RIM, manufacturers of the BlackBerry device, to solve specific language-related problems (further details cannot be disclosed due to a commercial confidentiality agreement).
5. The Nelralec project [G5] has had impact parallel to these commercial impacts producing a popular dictionary of Nepali (see http://www.nepalisabdakos.com/). Given that Nepal is a developing country, this activity has been third-sector, not commercial.
6. The Lancaster Edition of the Callhome Chinese Corpus has been distributed via the Linguistic Data Consortium to industrial research laboratories between 2009 and 2011 to enable the development of Natural Language Processing involving Chinese (e.g. machine translation). These users are based around the world and include: Autonomy Systems Ltd., UK; IBM, USA; NTT Communication Science Laboratories, Japan; and Loquendo SpA, Italy.

4. **Sources to corroborate the impact** (indicative maximum of 10 references)

The following give evidence in support of claims of commercial impact given in section 4.

[C1] http://www.bis.gov.uk/assets/biscore/higher-education/docs/e/11-980-estimating-value-of-education-exports.pdf (Department for Business, Innovation and Skills report which demonstrated the value of the UK ELT industry).

[C2] http://www.pearsonlongman.com/dictionaries/pdfs/corpus-lexicography.pdf (An article published by Longman detailing the importance of the Longman Corpus Network to their dictionary making activities).

[C3] http://www.longmanusahome.com/dictionaries/corpus.php (Material at the Longman USA website highlighting the importance of the BNC and the Longman Corpus Network to their publishing activities).

[C4] http://oald8.oxfordlearnersdictionaries.com/bnc.html (Material from the OUP website showing how important the BNC is to lexicography at OUP).

[C5] http://www.pearson.com/news/2012/july/pearson-2012-half-year-results.html (Gives Pearson's half yearly results for 2012 showing the importance of ELT publishing to the overall profitability of Pearson).

[C6] http://www.cambridgeesol.org/about/news/annual-review-2010.html (Support for claims regarding the volume of students taking the Cambridge English language tests).

[C7] http://www.ntm.org.in/languages/english/dictionaries.aspx (Support for the BNC as the basis for the English/South Asian language bilingual dictionaries).

[C8] ftp://ftp.cordis.europa.eu/pub/ist/docs/ic/benedict-ist-results_en.pdf (Article detailing Kielikone's use of the materials produced by the Benedict project).

[C9] http://www.kielikone.fi/en/Products/Dictionaries/ (The Kielikone catalogue).

[C10] http://www.windowsphone.com/en-US/apps/20e4dab7-089b-44d5-812a-3409dbd4c320 (Kielikone mobile phone application).

Contact who can corroborate our impact on industry and education is the General Secretary of the European Languages Resources Association and CEO of the Evaluations and Language Resources Distribution Agency.

Factual statement corroborating our impact on industry and beyond provided by a former Home Secretary in the UK government.