Institution: Birmingham City University

Unit of Assessment: 29 – English Language and Literature

Title of case study:
Academic, educational and commercial benefits of effective textual search and annotation

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

Based in the School of English, the Research and Development Unit for English Studies (RDUES) conducts research in the field of corpus linguistics and develops innovative software tools to allow a wide range of external audiences to locate, annotate and use electronic data more effectively. This case study details work carried out by the RDUES team (Matt Gee, Andrew Kehoe, Antoinette Renouf) in building large-scale corpora of web texts, from which examples of language use have been extracted, analysed, and presented in a form suitable for teaching and research across and beyond HE, including collaboration with commercial partners.

2. Underpinning research (indicative maximum 500 words)

Over the past two decades, RDUES has built an international reputation by taking a novel empirical and linguistically-informed approach to the testing of hypotheses about the nature and regularities of word patterns in text, with applications in the automatic identification of textual topic, word meaning, and semantic equivalence. RDUES' greatest success in recent years has been the **WebCorp** suite of online linguistic search tools (http://www.webcorp.org.uk). Released as a prototype in 2000 and developed through an EPSRC project (2000-03), **WebCorp Live** was designed to test the hypothesis that the web could complement traditional text corpora by providing evidence of rare, new and changing language use. This was achieved through the development of software which adds layers of refinement to conventional web search engines such as Google to produce linguistic 'concordances' showing examples of words or phrases in context. In order to keep pace with technological change and the metamorphic nature of the web, the WebCorp tools have required on-going revision and redevelopment, funded internally since 2004. The latest version, released in 2011, provides support for languages beyond Western Europe (including Chinese and Japanese) and offers improved performance. During the REF period, the team has also developed an entirely new tool, the **WebCorp Linguist's Search Engine (WebCorpLSE)**. This was designed to bypass the commercial search engines on which WebCorp Live relies as gate-keepers to the web by creating a large scale web search engine for language study. WebCorpLSE has crawled the web, downloading and processing texts to build a 10 billion word, linguistically-tagged web corpus, including sub-corpora for specific research purposes: the *Anglo-Norman Correspondence Corpus* and *Birmingham Blog Corpus*, as well as literary, news, and general web corpora.

The **Repulsion** project (2006-07) took as its starting point the established linguistic notion of collocation: the strength of association between pairs of words, or how frequently they appear as close neighbours. This project was novel in focussing on the inverse: dispreference between word pairs. In the process, the team developed new techniques for the measurement and visualisation of both lexical preference (collocation) and dispreference (repulsion).

Research on WebCorpLSE and Repulsion has found a new audience beyond the academic community through an AHRC **Knowledge Transfer Fellowship**. This award and a partnership with Stratford Grammar School allowed the development of a new interface and set of search functions tailored to the requirements of A-Level English Language students. A series of 'master classes', with group activities and interactive quizzes, introduced students to concepts, analytical techniques and software tools developed during the course of the research projects detailed above, distilled into a form suitable for the new audience.

The WebCorpLSE literary sub-corpora and technological knowledge underpinning the system have also been used in the development of **eMargin**: a web-based system for the collaborative annotation of texts (http://emargin.bcu.ac.uk). This work began as an attempt to bridge the gap between two distinct approaches to textual analysis: the top-down, quantitative approach of corpus linguistics and the fine-grained, introspective approach of literary close-reading. As the eMargin software has developed, it has found new audiences beyond English and beyond academia (see section 4).

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

Peer-reviewed research outputs relating to WebCorp and Repulsion:

2004-13: WebCorp Live and WebCorpLSE software systems and user interfaces: http://www.webcorp.org.uk, http://www.webcorp.org.uk/lse, http://www.webcorp.org.uk/mc.

2006: A. Kehoe 'Diachronic Linguistic Analysis on the Web with WebCorp' in A. Renouf & A. Kehoe (eds.) *The Changing Face of Corpus Linguistics*, Amsterdam: Rodopi: http://bit.ly/webcorp (returned to RAE2008)

2007: A. Kehoe, A. & M. Gee 'New corpora from the web: making web text more "text-like"' in *Studies in Variation, Contacts and Change in English Volume 2: Towards Multimedia in Corpus Studies,* University of Helsinki: http://www.helsinki.fi/varieng/journal/volumes/02/kehoe_gee (returned to RAE2008)

2007: A. Renouf & J. Banerjee 'Lexical Repulsion between sense-related pairs' in *International Journal of Corpus Linguistics* 12:3, 415-443, DOI: 10.1075/ijcl.12.3.05ren (returned to RAE2008)

2009: A. Kehoe & M. Gee 'Weaving Web data into a diachronic corpus patchwork', in A. Renouf & A. Kehoe (eds.) *Corpus Linguistics: Refinements and Reassessments*, Amsterdam: Rodopi: http://bit.ly/corpuspatch (listed in REF2)

End of award peer assessments for EPSRC projects GR/R16884/01 (WebCorp) and EP/E001300/1 (WebCorpLSE) projects: both graded as 'outstanding' on completion.

Key Grants:

2006-07	**Repulsion: The investigation of an organising force in text** PI: A. Renouf EPSRC: £125,855 (EP/D502551/1)
2006-08	**WebCorp Linguist's Search Engine** PI: A. Renouf EPSRC: £124,954 (EP/E001300/1); HEFCE-SRIF: £90,000
2009-11	**Introducing A-level English Language students to empirical text study using the WebCorp Linguist's Search Engine** PI: A. Renouf AHRC Knowledge Transfer Fellowship: £75,136 (AH/H01716X/1)
2011-13	**eMargin – an online collaborative textual annotation resource** PI: A. Kehoe JISC Learning & Teaching Innovation Grant: £44,336 PI: A. Kehoe JISC Embedding Benefits Grant: £15,000

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

Academic Beneficiaries:

The impact of the WebCorp tools extends significantly beyond Birmingham City University and beyond the UK, with over 15,000 searches per month throughout the REF period from users in 170 countries (with particular growth in China – see Fig. 1). The tools allow users to carry out quantitative analyses of the kind previously impossible on the web. As a result, WebCorp has been used by researchers internationally as a source of data and analyses for monographs, chapters, and articles in peer-reviewed journals, on topics ranging from Historical Linguistics to Legal Discourse and from C19 Fiction to Climate Change [1]. To give a specific example, the *Anglo-Norman Correspondence Corpus,* a unique resource built by the RDUES team and searchable through WebCorpLSE, was used as a data source for an article on historical pragmatics in *Lingua*.

In terms of teaching, usage records and user feedback show that hands-on sessions using WebCorp Live are included in university linguistics syllabi at Toronto, Paris, Washington, Stanford, Oxford, Cambridge, and the Open University, amongst other institutions. The software is included in the meta-search tool at Proz.com, the world's largest community of translators (http://www.proz.com/wts), and is used heavily in translation work as a result. The WebCorp tools were and continue to be developed iteratively in response to feedback from users, some examples of which include:

> *This search tool is a very good idea. As a Dutch teacher of English I use it to check, for instance, the idiomaticity of my students' work.* (University of Nijmegen; October 2009)

> *I am a Linguistics graduate student at Udayana University, Bali doing research on*

> *polysemy in Indonesian. I have found the WebCorp tools beneficial for my research […] since no Indonesian corpus like the BNC is available* (Udayana University; August 2010)
>
> *We are working on a research project dealing with gender issues, discourse analysis and corpus linguistics. Thank you so much […] and congratulations for your fantastic work.* (GENTEXT, University of Valencia; April 2011)
>
> *I'm currently using WebCorp as the search tool in a project funded by a Brazilian research agency called Fapesp. Because the collocations [being studied] are new in Brazilian Portuguese, WebCorp seemed perfect for our project.* (Universidade Federal de São Carlos, Brazil; June 2011) [2]

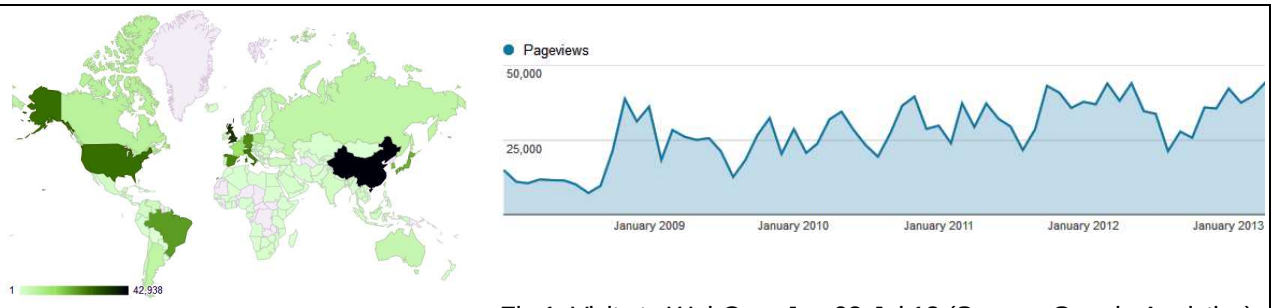

Fig 1: Visits to WebCorp Jan 08-Jul 13 (Source: Google Analytics)

The open-source eMargin annotation tool has addressed the limitations of the traditional approach to close reading and opened up new possibilities for collaborative study. The academic benefits of eMargin are being felt increasingly across the HE sector, with more than 800 registered users from over 100 institutions worldwide. The wide range of academic uses are evident in the names of the groups created by eMargin users, with examples including 'University of Leicester – MA Editing and Textual Criticism', 'Lancaster University – LING 450 Stylistics', 'Central Methodist University – EN216 Imaginative Writing', 'VU University Amsterdam – Metaphor in Language', 'University of Leicester – American Studies Centre', 'University of Huddersfield – Music, Humanities and Media', and 'University of Edinburgh – Divinity – Jesus & the Gospels'. Several institutions have taken maximum advantage of eMargin by integrating the tool with their own Virtual Learning Environments (Moodle, Blackboard, etc.) using its IMS-LTI connectivity [3]. Researchers at Lancaster University used eMargin to annotate interview transcripts in an ESRC-funded project investigating the role of metaphor in the experience of end-of-life care in the UK (http://ucrel.lancs.ac.uk/melc/). The impact of this project, to which eMargin has made a substantial contribution, will be felt fully beyond academia in the next period.

The RDUES team worked in collaboration with staff from the English, Student Development, and Course Design departments at the University of Leicester to test eMargin in a classroom environment. 96% of their English students found eMargin 'easy' to use and 92% agreed that 'reading others' comments helped me formulate my own ideas' [4]. The Leicester team ran a workshop using eMargin at the *24th Annual Lilly-West Conference on College and University Teaching* in California (March 2012), and invited Kehoe and Gee to speak at an HEA workshop in Leicester on social annotation (July 2012). Kehoe was also invited to present eMargin to postgraduate research students at a cross-disciplinary AHRC 'Hidden Collections' training event (University of Nottingham, November 2012). One participant, a PhD student in early cinema at the University of Glasgow, wrote afterwards

> I think eMargin is a remarkable tool, especially for teaching, as it makes the most of well-established practices but puts them in a context where collaboration becomes second nature. [5]

The Programme Manager for e-learning at JISC summarised the experiences of eMargin users across HE when he wrote

> It's perfect for critiques of papers (especially policy papers in my world), deep analysis of research publications, giving feedback on written work, looking at the structure of poetry or prose… the possibilities are endless. […] What really shines through with eMargin is that it meets a clear need, and it is designed around the practices and expectations of learners and educators. [6]

Non-academic Educational Impact:

The AHRC KT Fellowship project brought direct benefits to students and teachers at the partner school. The project had a major impact on the teaching of English Language at A-level through the introduction of concepts, analytical techniques and software tools not previously found in pre-university study. For example, work on Repulsion has informed the development of an online quiz in which students had to decide which of the words 'road' and 'street' was most appropriate in a given context. This reveals that, whilst we may assume 'road' and 'street' to be synonymous, they actually behave in subtly different ways in text, collocating with ('toll', 'rage') and ('cred', 'robbery') respectively. The Head of English at the school explained how

> this very powerful tool […] proved very useful for highlighting key areas of language change that are essential to the A2 course which students pursue; and the activities you devised and the resource itself helped develop student understanding to a very high degree. Several students went on to incorporate elements of recent lexical language change into their A2 coursework investigations; and we continue to have access to the resource to develop further our research of language change as and when we need. It was also highly useful to the students to work first-hand with experts in their field. [7]

Teachers also found Continuing Professional Development benefits from the work through increased awareness of the latest developments in linguistics. This is especially important for teachers of English Language, many of whom have not studied the subject as a significant element of their own first degree. The RDUES team sought to maximise future impact of the work by providing a workshop for 30 trainee teachers in the School of Education at the University of Birmingham in 2011. An AHRC peer-reviewer called RDUES' KT Fellowship bid 'tremendously exciting' and 'in many ways the model of a KT project'.

Commercial Impact:

RDUES' expertise, data resources and WebCorpLSE technology have been exploited in commercial work with Grey London, one of the largest communications agencies in the UK. Kehoe and Gee were employed as consultants on the language used by young people in social media, as part of a product launch by the sportswear brand Puma and Procter & Gamble, manufacturer of Puma's Sync fragrance range (May 2013). The 'Puma Dance Dictionary' campaign centred on a website (http://www.pumadancedictionary.com/) translating user messages into videos of dance moves, shareable via social media. This was accompanied by television advertisements across Europe. The Digital Producer leading the campaign described how the researchers provided

> an insight in to the language of social media that we would not have been able to gain through other means. […] The fact that the linguistic output […] provided was categorized allowed us to reduce costs by identifying words with a similar meaning. […] The consultancy […] allowed us to reflect the actual language used by our target audience more closely. [8]

The resulting campaign was well received by the target market and in the trade press, with *Marketing* magazine calling it 'innovative' and 'a great interactive experience'. [9]

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

1. For outputs by other researchers using WebCorp see http://www.webcorp.org.uk/publications *Lingua* article (P. Larrivée, 120:9, 2240-2258, DOI: 10.1016/j.lingua.2010.03.001)
2. Extracts from WebCorp feedback log, submitted via http://www.webcorp.org.uk/feedback
3. eMargin user database available on request.
4. Further student comments at http://repository.jisc.ac.uk/5388/1/eMargin_final_report.pdf
5. Blog, María A. Vélez-Serna: http://earlycinema.gla.ac.uk/hidden-collections-corpora-workshop
6. Blog, David Kernohan, JISC: http://elearning.jiscinvolve.org/wp/2013/05/10/collaborating-on-textual-analysis-with-emargin
7. Testimonial letter from Head of English at Stratford Grammar School, available on request.
8. Testimonial letter from Digital Producer at Grey London, available on request.
9. *Marketing* magazine (10/05/13): http://www.marketingmagazine.co.uk/article/1181766/Words-become-dance-moves-Pumas-latest-viral