Institution: University of Brighton

Unit of Assessment: B11 Computer Science and Informatics

Title of case study: Using Computational Lexicography for Dictionary Production with the Sketch Engine **ICS [2]**

1. Summary of the impact

The University of Brighton (UoB) has developed a new corpus-evidence-based approach to lexicography along with supporting tools and training resources. This approach has resulted in the development of a computational lexicography tool, the Sketch Engine, commercialised by Lexical Computing Ltd. The Sketch Engine has been adopted by four of the UK's five major dictionary publishers, national language institutes in nine European countries and over 100 universities, to support commercial dictionary production, language technology products and to enable language teaching. It has also been used to substantiate arguments in a pervasive debate about language use in the art world.

2. Underpinning research

The 1990s saw a dramatic increase in the availability of text in digital form and, with it, new challenges and opportunities for lexicographic research. The arrival of text corpora (collections of texts) containing billions of words allowed researchers to explore the detailed behaviour of words, based on hard evidence from many text sources, on a scale that had previously been impossible. This research thread developed into a key research theme at the UoB, and ultimately a spin-off company, with substantial influence in both academic and commercial approaches to the emerging field of computational lexicography.

This research began with the appointment of EVANS in 1993, then a SERC Advanced Fellow [reference 3.7], exploring the relationship between structure and processing in languages. In 1995, EVANS was awarded SERC funding [3.8] to develop ways in which large-scale language resources could be exploited to produce better computational models of lexical information. EVANS recruited KILGARRIFF as a research fellow, and the project developed a novel approach to lexicography, based on statistical analysis of the empirical behaviour of individual words in large online text corpora. Previous practice relied on lexicographers' intuitions, experience and manually collected examples of use. By contrast, this new approach offered a more rigorous account of word usage, based on large-scale linguistic evidence, with better coverage and direct access to supporting evidence. This work attracted the interest of dictionary publishers, notably Macmillan, which funded consultancy work to introduce these ideas into its dictionaries [3.1].

One aspect of lexicography that was thrown into sharp focus by this approach was the question of distinguishing word senses. Corpus analysis research showed that traditional sense distinctions made by lexicographers were approximate and incomplete at best, calling into question whether it was possible to make sense distinctions in a principled way at all. KILGARRIFF addressed this issue in a particularly influential paper [3.2] that has been reprinted in several collections of readings on the lexicon. The thrust of his argument is that word senses do not exist in an objective manner; the best that can be done is to cluster word occurrences with similar meanings, and how you do this depends on the task at hand.

From 1998, KILGARRIFF led an international research effort into corpus-based study of word senses through the SENSEVAL initiative [3.4], supported by EPSRC funding [3.9,3.11]. SENSEVAL supported the comparative evaluation of computational word sense disambiguation systems developed by internationally leading research teams. This was achieved by developing data resources [3.3], shared tasks and evaluation protocols. It ran evaluation campaigns in 1998, 2001 and 2004 and, after a name change to SEMEVAL, in 2007, 2010, 2012 and 2013.

Meanwhile, this emerging notion of *corpus-based computational lexicography* was further developed in the follow-on funding from EPSRC [3.10] (employing KILGARRIFF and TUGWELL). A pioneering outcome was the notion of *word sense profiles*, now generally known as *word sketches*. These serve as the foundation for corpus-based lexicography research and development and are at the heart of software tools to support the lexicographer in creating, analysing and exploring word usage [3.6].

This underpinning research led to the formation of Lexical Computing Ltd, which was the key vehicle for taking the developed technology, the Sketch Engine, to market. The company was formed in 2003 by KILGARRIFF, at a time when the commercial dictionary publishing sector was beginning to engage with the opportunities and challenges of digital publishing.

Key researchers:

Roger Evans: Senior Research Fellow (Oct 1993–Sept 1994), Principal Research Fellow (Oct 1994–July 1997), Reader (Aug 1997–to date).

Adam Kilgarriff: Senior Research Fellow (Feb 1995–Aug 2002), Senior Lecturer (Sept 2002–Oct 2004).

David Tugwell: Research Fellow (Sept 1999–Sept 2002), Hourly Paid Lecturer (Oct 2002–Feb 2005).

3. References to the research

The three outputs marked with a # best indicate the quality of the research.

[3.1] # KILGARRIFF, A. (1997) Putting frequencies in the dictionary, *International Journal of Lexicography* 10(2), pp.135–155. [Quality validation: this paper was published in a major journal in the field and has been cited over 130 times (Google Scholar)]. 10.1093/ijl/10.2.135.

[3.2] # KILGARRIFF, A. (1997) I don't believe in word senses, *Computers and the Humanities* 31, pp. 91–113. [Quality validation: this publication was part of UoB's RAE2001 submission, and has been particularly influential, cited over 320 times (Google Scholar). 10.1023/A:1000583911091. It has been reprinted in three collections since its original publication in 1997:

- FONTENELLE ed. *Practical lexicography: a reader*: Oxford University Press.
- NERLICH, TODD, HERMAN and CLARKE eds. *Polysemy: flexible patterns of meaning in language and mind.* Walter de Gruyter, pp. 361–392.
- PUSTEJOVSKY, J. and WILKS, Y. (eds.) *Readings in the lexicon: interdisciplinary perspectives*. Cambridge, MA: MIT Press, (in press).]

[3.3] KILGARRIFF, A. (1998) Gold standard datasets for evaluating word sense disambiguation programs, *Computer Speech and Language,* 12(4), pp. 453–472. [Quality validation: published in a major journal in the field.] 10.1006/csla.1998.0108.

[3.4] KILGARRIFF, (2000) A. Rosenzweig, English framework and results, *Computers and the Humanities* 34 (1-2), pp. 15–48. [Quality validation: published in a major journal in the field. Cited nearly 200 times (Google Scholar).] 10.1023/A:1002693207386.

[3.5] KOELING R., KILGARRIFF A., TUGWELL D., and EVANS, R. (2003) An evaluation of a lexicographer's workbench: building lexicons for machine translation, *Proceedings of the 7th International EAMT workshop as part of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, pp. 6–16. [Quality validation: refereed workshop paper at principal European conference in the field.] 10.1007/3-540-36456-0_23.

[3.6] # KILGARRIFF, A., RYCHLY, P., SMRZ, P. and TUGWELL D. (2004) The Sketch Engine, *Proceedings of Euralex*. Lorient, France, pp. 105–116. [Quality validation: refereed conference paper at major European conference in lexicography. Subsequently reprinted: *Lexicology: critical concepts in linguistics* (Hanks, ed. Routledge, 2007). The associated technical report has been cited almost 500 times (Google Scholar).]

Key research grants:

[3.7] EVANS, SERC Advanced Fellowship, SERC [B/ITF/187, 1988–1994, total funding: £86,000].

[3.8] EVANS with KILGARRIFF, Structural enhancement of automatically-acquired lexicons (SEAL), EPSRC [GR/K18931, 1995—1998, total funding: £127,103].

[3.9] EVANS with KILGARRIFF, A Manually Sense-tagged Gold Standard Corpus (SENSEVAL), EPSRC [GR/M03481, 1998–1999, total funding: £10,255]. EPSRC post-project assessment: significant contribution (management excellent).

[3.10] EVANS with KILGARRIFF, A semi-automatic lexicographer's workbench for writing word sense profiles (WASPS), EPSRC [GR/M54971, 1999–2002, total funding: £287,207]. EPSRC post-project assessment: outstanding.

[3.11] EVANS with KILGARRIFF, Manual Tagging for SENSEVAL (MATS), EPSRC [GR/R02337, 2001—2002, total funding: £15,341]. EPSRC post-project assessment: outstanding.

4. Details of the impact

The Sketch Engine has attracted considerable attention, letting users access information on between 30 million and 15 billion words for a wide range of languages (61 languages are currently covered). Lexical Computing Ltd, formed in 2003, has expanded throughout the last 10 years, with key organisations employing the Sketch Engine throughout the impact period. Staff are now employed in the UK and the Czech Republic, along with freelancers in a number of other countries; half of the company's business is overseas.

Commercial lexicography: The Sketch Engine is part of everyday use for lexicography worldwide and is currently used by commercial dictionary producers, including Cambridge University Press, Collins, Macmillan and Oxford University Press (OUP), along with Cornelsen Verlag, Le Robert, and Shogakukan (source 5.1). At OUP, for example, the use of the Sketch Engine has served to revolutionise their in-house research into word behaviour, building a detailed statistical profile of a word in seconds (5.2).

Macmillan started implementing the Sketch Engine during 2007, resulting in embedded use throughout the impact period. The word sketch approach has enabled Macmillan to suspend print and focus on online dictionaries; *Macmillan Dictionary Online*, launched in 2009, has seen 'explosive growth' and from 2012 has fully replaced the print version (5.3). Word sketches are used to find patterns in the use of words, phrases and grammatical configurations, allowing lexicographers to build up a picture of the most important facts about words, which then form the basis of how that word is described in the dictionary. The building of dictionary content comes directly from this analysis, resulting in content that is more accurate, comprehensive and detailed than was possible with previous methods.

Dictionary producers such as Macmillan can now work efficiently and reflectively, to manage evolving data use caused by the growth in digital information. They have moved away from intuitive and introspective methods for building information to this more comprehensive, evidence-based approach. In addition to this, Macmillan has acknowledged that the unique feature of its dictionaries, the labelling of red and black words, is only possible through the comprehensive analysis made possible by the Sketch Engine. Red and black words distinguish between high-frequency core vocabulary and the less common words needed mainly for reference. The *Macmillan English Dictionary* is the only advanced learner's dictionary to highlight effectively the most important 7,500 words an advanced learner needs to be able to use so that they can become as fluent in English as native speakers (5.4).

Professional training, language teaching and learning: An annual one-week, intensive commercial facing training course, LEXICOM, has been delivered by KILGARRIFF and colleagues during the impact period. LEXICOM has been delivered at venues all over the world attracting participants from the publishing industry (managers and editors), other commercial and not-for-profit companies (engineers and linguists), universities and government agencies (terminologists and translators). This widespread exposure has led to the application of the Sketch Engine in national language institutes in nine European countries and over 100 universities, including Reading, Leicester, Portsmouth, Warwick and Birmingham. Internationally it is being used in key language annotation modules by Brandeis University, which adopted the Sketch Engine in 2012, and for foreign language teaching at the Institute for Applied Slovene Studies in Slovenia (5.5, 5.6). The Sketch Engine has been highlighted as a key learning resource on independent forums, such as 'Methodologies and Approaches to ELT' [5.7]. It is promoted as a tool that can operate reflexively and works particularly well for small, short-term projects such as translating and preparing topic-based teaching material. Lang-8, a language-exchange social network, also highlights Sketch Engine as an effective interface for English writing and language learning (5.8).

Spreading the application: The Chief Executive of brand naming company, *Operative Words*, describes the Sketch Engine as the 'most powerful naming tool available', instrumental in information gathering to enable branding techniques and new creative directions for their portfolio (5.9). In 2012, a Guardian article highlighted how the Sketch Engine was used to analyse thousands of exhibition announcements to discover the specific characteristics of 'Art Speak' and its effect on the public (5.10). The analysis was originally published in an American art journal, Triple Canopy, in July 2010 and this journal article has since become a widely circulated piece of online cultural criticism, sparking further debates on other forums, including Wordpress, Tumblr, Google+, Ikono, Artblog and Artsia (the Society of International Artists). In addition, the BBC has joined forces with OUP to explore children's writing. The Oxford Children's Corpus, in 2011, was adapted to include a component on children's writing, primarily with data from the BBC Radio 2 '500 Words' short story writing competition. Lexical Computing Ltd is working with OUP to analyse the language that children use. The 74,000 entries, received in 2012 now form a large part of the Children's Writing component of the Oxford Children's Corpus (5.11).

In 2011, the Sketch Engine was used to undertake 'A Corpus Linguistic Analysis of Ecosystems Vocabulary in the Public Sphere' commissioned by the UK National Ecosystem Assessment (5.12).

5. Sources to corroborate the impact

5.1 Evidence of the commercial impact of this work can be found on the Sketch Engine website. Available at: www.sketchengine.co.uk. [Accessed: 12 November 2013].

5.2 Oxford Dictionaries, 'Using the Corpus.' Available at: http://oxforddictionaries.com/words/using-the-corpus [Accessed: 12 November 2013]. Embedded use of the Sketch Engine in Oxford dictionaries.

5.3 The Wire, 'All for the love of a good reference book'. 13 November 2012. Available at: http://www.theatlanticwire.com/entertainment/2012/11/all-love-good-referencebook/58950/?goback=%2Egde_4293299_member_186426184 [Accessed: 12 November 2013]. Evidence that Macmillan has phased out printed dictionaries.

5.4 Macmillan Dictionaries, 'From Corpus to Dictionary'. Available at: http://www.macmillandictionaries.com/features/from-corpus-to-dictionary/ [Accessed: 12 November 2013. Evidences the use of word sketches in dictionary development.

5.5 'Language Annotation for Machine Learning'. Available at: https://sites.google.com/site/brandeisnlaml/course-news/sketchengine [Accessed: 12 November 2013]. Use of the Sketch Engine in a key language module at Brandeis University is evidenced by this website.

5.6 'Aston Corpus Summer School 2011, Corpora in Lexicography'. Available at: http://acorn.aston.ac.uk/SummerSchool2011/006-iztok-kosem2-PART%20ONE%20Corpora-Lexicography-AstonSummerSchool2011.pdf [Accessed: 12 November 2013]. Evidence of the use of the Sketch Engine in foreign language teaching in Slovenia.

5.7 'Methodologies and Approaches in ELT'. Available at: https://sites.google.com/site/eltmethodologies/approaches/data-driven-learning/corpus-resources-for-teaching/sketch-engine. [Accessed: 12 November 2013]. The Sketch Engine is being used for guidance to effective methodologies for learning.

5.8 'Sketch Engine – English corpora available online'. Available at:http://lang-8.com/odon/journals/901660 [Accessed: 12 November 2013]. This provides evidence of use on an online language learners' forum.

5.9 'How to create names using the world's most powerful naming tool.' Available at: http://operativewords.blogspot.co.uk/2011/08/how-to-create-names-using-worlds-most.html [Accessed: 12 November 2013]. Branding company's use of Sketch Engine.

5.10 The *Guardian*, 'A user's guide to artspeak', 27 January 2013. Available at: http://www.guardian.co.uk/artanddesign/2013/jan/27/users-guide-international-art-english [Accessed: 12 November 2013]. Evidences the international art English debate.

5.11 '*Oxford Children's Corpus: a Corpus of Children's Writing, Reading, and Education'*. Report available on request. This report confirms the use of Sketch Engine in the formation of the Children's Writing element of the Corpus.

5.12 UK National Ecosystem Assessment report with evidence of the use of Sketch Engine on page 41. Report available on request.