

<b>Institution:</b> <b>University of Glasgow</b>
<b>Unit of Assessment:</b> <b>29 – English Language and Literature</b>
<b>Title of case study:</b> <b>Public and educational use of language research: the online Scottish Corpora</b>
<b>1. Summary of the impact</b>

Researchers at the University of Glasgow have created the first freely accessible online database of written and spoken texts in Scottish English and Scots. Together, the Scottish Corpus of Text and Speech (SCOTS) and the Corpus of Modern Scottish Writing (CMSW), both developed at Glasgow, provide over 10 million words of text from a range of sources, complemented by audio and video recordings and digitised manuscripts and documents. They have succeeded in raising interest in and awareness of Scottish English and Scots among the general public: 40% of SCOTS's resources were contributed by the public, and the website achieved 165,000 page views per month at launch. The database is also widely used by commercial lexicographers and professionals in secondary education. It is an 'essential data source' for Scottish Language Dictionaries, 'in day-to-day use' by the Oxford English Dictionary, and from 2006-2013 has been deployed by school examination boards across the UK (Highers, A-Levels, Cambridge International, and Oxford, Cambridge and RSA exams).

<b>2. Underpinning research</b> (indicative maximum 500 words)
--

Between 2002 and 2011 John Corbett and Christian Kay led a team which sought to address the dearth of Scottish linguistic resources by assembling an extensive database of Scottish English and Scots. The research to create this database was expanded and processed into two linguistic corpora – that is, searchable electronic banks of texts designed for linguistic analysis:

- The Scottish Corpus of Texts and Speech (SCOTS) brings together 1300 text documents dating from 1945 to the present. SCOTS contains more than 4.6 million words, including 1 million words of speech which have been made available in audio/audio-visual recordings and orthographic transcriptions. The Glasgow team has also created corpus resources such as accompanying metadata, integrated analysis and visualisation tools for Scots and Scottish English.
- The Corpus of Modern Scottish Writing (CMSW) collects texts from earlier periods of history (broadly 1700 to 1945), adding a further 356 documents with page images and totalling 5.4 million words.

Alongside the Helsinki Corpus of Older Scots (1450-1700), the work carried out by the Glasgow team in creating these two corpora makes it possible to analyse the Scots language across a period of five and a half centuries. Taken together, both corpora fill a major gap in the research materials available to scholars of the English Language. Other British corpora, such as the British National Corpus and the Bank of English, contain some Scottish material but have not collected it comprehensively, whilst the International Corpus of English does not include Scotland in its collection of major varieties of World English.

So SCOTS, the first of the two Glasgow corpora, creates a much-needed resource facilitating research into the relationship between current-day standard and non-standard languages, as exemplified in Scotland by the continuum of language varieties stretching from Scottish English to Broad Scots. Before this, information was lacking on how extensively and in what contexts Broad Scots is still used, and on the range of linguistic features characterising Scottish English. Other benefits of SCOTS include its potential to provide a powerful tool for sociolinguistic research into Scots and Scottish English.

Subsequently, the CMSW project opened up research into these questions for earlier periods of history (1700-1945). The period covered by CMSW covers a major part of the history of modern Scotland and Scots, beginning with the last stages of the standardisation of written English and the

## Impact case study (REF3b)

onset of the Vernacular Revival in literary Scots that produced writers like Robert Burns. As a whole, language use in Scotland in the modern period, conventionally dated from 1700, can be described as a continuum with Standard English at one end, and social and regional varieties of Broad Scots at the other. Out of the interaction between Broad Scots and written Standard English the varieties of today's Scottish English are said to emerge. The Scottish Corpora projects provide evidence for empirical analysis of these varieties, and open up this evidence base for wider use.

The research was carried out between 2002 and 2011 by teams led by Corbett, now Honorary Senior Research Fellow at the University of Glasgow. Key researchers also included: Kay; Jeremy Smith; Jane Stuart-Smith; Fiona Douglas (until 2004); Wendy Anderson (from 2004); Jennifer Bann (from 2009); David Beavan (until 2011); Jean Anderson; and PhD student Dorian Grieve. Work on the corpora, therefore, also contributed to the development of new generations of language scholars.

In addition to creating the resource, members of the project teams for both SCOTS and CMSW used the two corpora to carry out linguistic analysis of Scottish language varieties. Douglas, J. Anderson, Kay, W. Anderson, and Beavan have published on the problems involved in creating and designing corpora of this kind, including procedures adopted in building corpora of language varieties which are not fully documented. Corbett, W. Anderson, and Kay have published linguistic findings based on analysis of the Scottish corpora – for example on intensifying adverbs, evaluative terms, and distinctive features of the grammar of Scots – while W. Anderson and Corbett jointly authored an introductory book demonstrating how freely available corpora such as SCOTS can be used for the study of English at different linguistic levels, including vocabulary, grammar, discourse and pronunciation. The online nature of the corpus, they argue, means that researchers, students and the public can replicate such analysis for themselves. A history of Scots literary orthography is ongoing and a monograph on this topic by Bann and Corbett is scheduled for publication in 2014.

<b>3. References to the research</b> (indicative maximum of six references)
---

- J Corbett and C Kay, *Understanding Grammar in Scotland Today* (Glasgow: ASLS, 2009.)
- W Anderson and J Corbett, *Exploring English with Online Corpora* (Basingstoke: Palgrave Macmillan, 2009). ISBN 9780230551404 [REF 2]
- J Anderson, D Beavan and C Kay, 'SCOTS: Scottish Corpus of Texts and Speech', in *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*, ed. by J. Beal, K. Corrigan and H. Moisl (Basingstoke: Palgrave Macmillan, 2007), pp. 17-34. ISBN-10: 1403943664 / 13: 978-1403943668 [available from HEI]
- F Douglas and J Corbett, "'Huv a wee seat, hen": Evaluative terms in Scots', in *The Power of Words: Essays in Lexicography, Lexicology and Semantics*, G Caie, C Hough and I Wotherspoon, eds (Amsterdam and New York: Rodopi, 2006), pp. 35-56. ISBN-10: 9042021217 / 13: 978-9042021211 [available from HEI]
- W Anderson, W. (2006) 'Absolutely, Totally, Filled to the Brim with the Famous Grouse: Intensifying Adverbs in SCOTS', *English Today* 22.3 (2006), pp. 10-16 [PDF link]
- W Anderson and D Beavan, 'Internet Delivery of Time-Synchronised Multimedia: The SCOTS Corpus', *Proceedings from the Corpus Linguistics Conference Series* 1.1 (Birmingham, July 2005). [PDF link]
- F Douglas, 'The Scottish Corpus of Texts and Speech: Problems of Corpus Design', *Literary and Linguistic Computing* 18.1(2003), pp.23-37. doi: [10.1093/llic/18.1.23](https://doi.org/10.1093/llic/18.1.23)

### Key research grants

- EPSRC 2002-2004 for Scottish Corpus of Texts & Speech. PI: C. Kay; CIs: J. Corbett, J. Anderson; RAs: F. Douglas; D. Beavan. Grant value: £160k.
- AHRB resource enhancement 2004-2007 for Scottish Corpus of Texts & Speech. PI: J. Corbett; CIs: J. Stuart-Smith, C. Kay; RA: W. Anderson; Computing officer: D. Beavan; Technician: F. Edmonds. Grant value: £310k.
- AHRC 2007-08 for Corpus of Modern Scottish Writing. PI: J. Corbett; CI: J. Smith; RAs: W. Anderson, J. Bann (2009-11); Computing officer: D. Beavan. Grant value: £430k.

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

The accessibility of the two corpora resources maximised the impact of the Glasgow team's research beyond academia. SCOTS **engaged public interest** from the outset, introducing linguistics to a wide audience through its web portal and articles in mainstream media. Other key impacts of the corpora have principally been in the fields of **technology, commercial lexicography** and **education**.

##### Public engagement

At the time of launch the SCOTS website averaged 165,000 page views per month, and it has achieved an average 36,000 page views per month as at July 2013. The resource itself drew heavily on material supplied by the public to build up its stock of linguistic resources, and so found innovative ways to appeal for non-specialist assistance, including radio and press pieces. The team also published longer pieces in two Scottish interest magazines, *ScotLit* and *Leopard*, and an article about the project appeared in *The Big Issue* in March 2009. A large proportion (c.40%) of the corpus material was obtained in response to these activities, with the longer articles being cited with particular frequency as the reason for engaging with the project.

The applications of SCOTS have been appreciated particularly in the field of English Language Teaching outside academia: between 2005 and 2012 Corbett gave a series of presentations on topics such as the use of the corpus in ESOL (English for speakers of other languages) in North and South America (Pittsburgh, Brasilia, Rio, Santiago), the UK (Glasgow, London, Swansea) and Middle East (Istanbul), to combined audiences of over 500 specialists. In a less specialised context, in June 2010 and 2011 Bann gave public presentations at Glasgow's West End Festival to audiences of c. 90 people.

##### Technological development

The Corpora projects have led to technological developments in and beyond the field of corpus and computational linguistics. Beavan developed synchronised audio transcription resources which draw on corpus data to produce data visualisations, providing accessible and concise overviews of language not readily apparent through close reading and ideal for encouraging non-specialists to explore linguistic data. These tools were adopted for use with other corpora – such as the British National Corpus – and informed the development of other text analysis tools, such as Voyant Tools. Beavan's '[Collocate Cloud](#)' tool won the international award for the 'Best New Idea for Improving a Current Web-Based Tool' at the 2008 TADA Research Evaluation eXchange.

##### Commercial lexicography

SCOTS is heavily used in lexicography, forming a major data source for Scottish Language Dictionaries and for the revisions of the *Oxford English Dictionary*. The Director of Scottish Language Dictionaries writes:

The [SCOTS Corpus](#) has become an essential data source for Scottish Language Dictionaries' editors [...] The content and presentation of this Corpus have unparalleled advantages which render it indispensable for certain focused kinds of search.

An Etymologist at the *OED* confirms:

Both corpora (SCOTS and CMSW) are in day-to-day use by us in our revision of OED [...] [T]hey are invaluable sources of quotation and form evidence which goes straight into our revised OED3 entries, allowing us to deepen and extend our coverage of Scots. [...] Without recourse to these corpora OED3 would certainly be the poorer.

##### Education

The SCOTS resource has influenced educational content. Between 2006 and 2013 selected texts drawn from the corpus have been used by examination boards for Higher and A-Level exams. These are the Scottish Qualification Authority (2006, 2007, 2008, 2009); University of Cambridge

**Impact case study (REF3b)**

International Examinations (2012); and Oxford, Cambridge and RSA Examinations (2011, 2012, 2013), who commented:

OCR would like to thank the SCOTS Project for allowing us to make use of its material in our examination and support materials. The resource is of value to us both in terms of its range and ease of access but, most importantly, its accuracy and attention to detail.

Given its free accessibility, the SCOTS corpus has considerable value for use in schools, promoting knowledge about how language works, reasons for the co-existence of different kinds of language at one time, and how language changes over years, decades and centuries. Anderson, Corbett and Kay have demonstrated the uses of the corpus at Continuing Professional Development events for teachers, including sessions on 'The Scots Language', 'Computing in English Studies' and 'Grammar for Teachers', attended by 172 teachers. They also led a monthly Scottish Literature Reading Group based on the corpus (2009-10), attended by 35 teachers. In 2010, Corbett delivered a lecture to 100 teachers at the *ASLS Schools Conference*, and distributed copies of Corbett and Kay's *Understanding Grammar in Scotland Today*, which draws its data from the SCOTS corpus.

*Understanding Grammar* was described in the English Excellence Group Report as 'A very important resource with excellent exemplification for this area', and the book was used as the basis for Christine Robinson's *Modern Scots Grammar* (Edinburgh: Luath Press, 2012), a grammar book for younger students. Anderson and Corbett also published *Exploring English with Online Corpora* (Palgrave, 2009), which is widely used in University and ELT teaching, and was described in *Literary and Linguistic Computing* as 'an excellent introduction to corpus research for readers without prior knowledge of linguistics' which 'successfully put the [SCOTS] corpus in the spotlight'. The book has sold c.400 copies in Europe and over 700 copies overseas, and a second edition is currently under consideration.

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

- SCOTS can be found here: <http://www.scottishcorpus.ac.uk>
- CMSW can be found here: <http://www.scottishcorpus.ac.uk/cmsw>
- A [review of Anderson and Corbett's \*Exploring English with Online Corpora\*](#), with comments on the SCOTS Corpus
- [The English Excellence Report on the SCOTS Corpus](#)
- Oxford Cambridge and RSA Examinations, OCR Assessment Services to corroborate use of resource in examinations (contact details provided by HEI)
- Research Manager, UCL Centre for Digital Humanities to corroborate website statistics (contact details provided by HEI)
- Director of Scottish Language Dictionaries to corroborate use of SCOTS in lexicography (contact details provided by HEI)
- Etymologist at Oxford English Dictionary to corroborate use of SCOTS and CMSW in revisions of OED (contact details provided by HEI)