Life after the *Historical Thesaurus of the OED*

Christian Kay and Marc Alexander

The *Historical Thesaurus of the OED* (*HTOED*) was published by Oxford University Press (OUP) on 22 October 2009 (Kay et al. 2009). It consists of two substantial volumes, the first containing some 800,000 meanings arranged in semantic categories, the second an index. Publication was the culmination of 44 years of work by a team in the English Language department at the University of Glasgow, led initially by Professor M. L. Samuels, and from 1989 by Professor Christian Kay[1].

Final Days

Through accident rather than design, the very end stages of data entry in July 2008 overlapped with a conference at Glasgow involving a number of team members, fraying the nerves of all involved. So it was that midway through the conference the project director returned to the office to be informed that not only had the last slip been typed into the database, but by serendipity or careful planning this final entry was the word *thesaurus* itself. A week later, a CD containing all the data files was posted to OUP; the project team insisted on a formal handing-over ceremony on the department doorstep, with photographs of Christian Kay solemnly delivering the envelope into the hands of the University janitor entrusted with the task of conveying it to the mailroom. However, the nature of the data and the complexity of the project was such that, after this milestone, there followed a long series of further deadlines and other landmarks, rather than there being any one single point of completion.

[1] For a history of the project, see, *inter alia*, the front matter of *HTOED*, Kay (forthcoming), Wotherspoon (forthcoming), the Glasgow website at http://libra.englang.arts.gla.ac.uk/WebThesHTML/homepage.html, and for an earlier perspective, Collier and Kay's 1980 interim report in *Dictionaries* 3.

Dictionaries: Journal of the Dictionary Society of North America 31 (2010), 107–112

The data went on a long journey in the final year. Once the paper slips were categorized, they had been typed into dBASE files using an inhouse data entry program in batches of around 30. These files were combined into larger batches, converted to Microsoft Excel format, and then combined once more into 186 files roughly representing a large semantic domain each. At OUP, these were converted to plain text format, checked using Perl scripts, and then converted once more to three XML files to be provided to the typesetters, who then produced InDesign files for printing. These typeset pages were supplied as PDFs to Glasgow, where they were printed, proofed, scanned as TIFFs, and converted back to PDFs. The annotations on these were inserted into the XML and thus updated on the InDesign files. At each stage of this process, nicknamed the "dance of the files," the possibilities for disaster in the conversion process increased — in many ways, it was a blessing that those problems which did arise were usually relatively minor, although distinctly trying under tight publication deadlines.

There then began a short break while the volumes were typeset abroad, although any relief this provided was marred by a cautious member of the team calculating that the typesetter's claim of a "99.995% accuracy rate," when applied to *HTOED*'s 22.74 million pieces of data, would result in over 1,100 new errors being created. A welcome distraction from such speculations was provided by the arrival of OUP's designs for the *HTOED* wallchart, included with every copy to give an overview of its conceptual hierarchy. Proofreading then went quickly and with a minimum of hiccups, although it naturally revealed some blunders we were glad to remove before publication — one such was an unfortunate creature noted within the Life section as being "devoid of Brian," soon corrected to "devoid of brain."

The first advance copy arrived in Glasgow on 21 August 2009. A host of colleagues from across the entire campus, almost all of whom had lived with the presence of the Thesaurus project from their very first days at the University, found reasons to come to the department and finger its binding somewhat incredulously. It was at this point that we realised that we had actually succeeded despite all the obstacles along the way. As Philip Pullman wrote in his endorsement of the book: "… here is the information we had to spend hours hunting down through the thickets and coverts of the great *OED*, shot, stuffed, and mounted for us".

Publication

The calendar of publication and pre-publication events then took off. A pre-launch was held in London on 7 July 2009, and had the desired effect of stimulating press interest, resulting in numerous feature articles, radio interviews, and even an editorial in the *Times* of London. The press questions focused in the main on the sheer length of the project, the fire which almost destroyed the entire slip collection in 1978, people's favorite words, and the ages of the editors. The team began during the proofreading process to assemble a list of these "favorite" words to satisfy journalistic enquiries, although many of those collected by PhD student proofreaders turned out to be unpublishable in family newspapers.

The launch proper took place in Glasgow in October, when around 100 former workers and supporters gathered for drinks and speeches. A particular pleasure was that Prof. Samuels was able to attend and give a short talk on his reasons for starting the project. Those who had known him as a lecturer earlier in their careers were seen to sit up straight and start taking notes. In fact, the creation of the *HTOED* was precisely bookended by two of Prof. Samuels' speeches: the project formally began on 15 January 1965 at an address to the Philological Society in London, where he announced that the work would be undertaken by himself and his colleagues at Glasgow, and it ended at the launch party on 22 October 2009. The project thus consumed 44 years, 9 months and 1 week exactly (or 16,351 days). It cost £1.1million in grants (when adjusted for inflation approximately £2.2m/$3.4m in 2010), plus a good deal of uncosted academic time; a bargain at a little over 1p per word and around 340 words a week.

Even at the launch, there were hints that *HTOED* was selling much better than expected, and in November we learned that the first printing had sold out. A second printing was rushed through in time for Christmas, and there were two more in 2010. It was gratifying that the book seemed to appeal to word-lovers generally as well as to its primary target of academic users. Another unexpected market was the surprising number of historical novelists who appear to see in *HTOED* a means of providing a ring of authenticity to their dialogue.

As well as offering parties and press interviews, the period around publication provided an opportunity to speak at conferences on the full work. After many years of papers based on available excerpts, the first talk to use the full *HTOED* database in its complete form was given in March 2009. Between then and summer 2010, over thirty papers were presented in the UK, Europe, and North America by team members alone. More importantly, we have begun to see papers appearing in print and at conferences by scholars who were not involved in the creation of the work itself, but have gratifyingly grasped even at this early stage the opportunities it offers them to advance their own research.

Last Words

One of the worst aspects of finishing a major project is the requirement to clear up after it. By the end of the project, we had accumulated well over a million paper slips, and a set of OED volumes recording all the meanings which had been included. These had been treasured over the years, especially after the aforementioned fire, when they were saved only because they were stored in metal drawers in metal cabinets. On the one hand, since all the data were now held on computer, it seemed pointless to keep a paper version. On the other hand, some of the slips contained discussions about where they should be classified, recording many changes of mind, or comments on the OED data, or other interesting marginalia in the manner of medieval manuscripts. It was not impossible that future scholars might wish to track our mental processes through these stages, or debate whether the more sinister-looking stains might be blood or merely coffee. In the end a compromise was reached, and the slips were sent to the Glasgow University archive, with a note to review the situation in ten years' time.

Preparing the slips for departure and moving out of the workroom the project had occupied for many years were sad as well as strenuous occasions. Even sadder was saying goodbye to most of the people who had worked with such dedication and good humour during the final phase of the project. Luckily, this sadness was overcome by excitement at seeing the whole work complete and published at last. Plans were set in motion by OUP to incorporate *HTOED* into the relaunched *OED3* website, so that the two could be searched in tandem. Clicking on an *OED* meaning accesses the list of *HTOED* synonyms, while clicking on a word in an *HTOED* list connects the user to the full range of information in the *OED*.

Finally, a few months after publication, we were able to announce the *HTOED* Scholarships at Glasgow. Funded entirely from the print edition's royalties, these scholarships for new postgraduates researching any part of the English language were agreed long before publication to be the most appropriate reinvestment of any income earned by the Thesaurus. The first four *HTOED* scholars began their studies at Glasgow in September 2010.

Future Plans

One of the most frustrating things about working on a major project is that it leaves the editors little time for exploiting its research potential. Over the years we responded to requests for data from over 40 colleagues around the world, and in due course these became articles or dissertations. But at home, apart from the theses produced by our own students, research output was restricted by lack of time, and papers were sometimes written with at least half an eye to potential sources of funding. Now we have been liberated by publication, and have time to contemplate the full richness of *HTOED*, and to plan how to exploit the features which give researchers a unique new perspective on the history of the English language. We will maintain a revised version of the project on the University of Glasgow website. Like the print *HTOED*, but unlike the one on the *OED* website, which follows the *OED* policy of not including words which died out before 1150, this version will contain all the vocabulary from *A Thesaurus of Old English* (Roberts and Kay 2000).

Furthermore, as the *HTOED* project was one of the first to use computers in humanities research[2], it has progressed side-by-side with the evolution of what is now known as the digital humanities. There are many potential applications of its data in the computational sphere alongside the use of the originally-envisaged print volumes. As the project was coming to an end, the UK Joint Information Systems Committee funded the Enroller project at Glasgow, designed to bring together various datasets in the study of language and literature into a single online platform. *HTOED* will be in Enroller alongside various corpora and dictionaries, allowing researchers not only to easily browse the data, but also to carry out meaning-based searches using two or more of the resources together (such as "find all the words in the *SCOTS* corpus to do with drunkenness" or "find all the contemporary words for excellent in this eighteenth century text").

Beyond this, there are scholars planning to use *HTOED* in many different and previously unthought-of manners. Projects are being set up to investigate the history of metaphor in English by computationally finding the source and target domains of transferred or figurative uses throughout the entire *HTOED* database; to tag political texts both modern and historical with the precise semantic domain of each word form used; to examine spelling variation by semantically disambiguating word forms and then analysing their orthographic reflexes; to analyse modern historical fiction through charting its lexical authenticity (or lack thereof); or even to demonstrate visually historical trauma in semantic fields throughout the past 800 years of English. Taking into account con-

[2] See Kay and Chase (1987) and Wotherspoon (1992).

ference papers, publications and symposia both completed and planned, the volume of research activities arising from *HTOED* barely half a year from its launch day is encouraging.

These possibilities, more than anything else, represent the culmination of what over 230 people have worked on for the past 44 years and more. In these implementations and investigations into the data, we see scholars who have spent almost their whole career creating *HTOED* working alongside younger academics who can see their own careers consisting of using it. We feel there is no greater or more fitting end to the long history of this project for the *HTOED* to be used, to be exploited, and to be improved — both in the ways which were originally intended and in others as yet unimagined.

References

Collier, Leslie and Christian Kay. 1980. The Historical Thesaurus of English. *Dictionaries* 2/3: 88–112.

Kay, Christian, Jane Roberts, Michael Samuels and Irené Wotherspoon. 2009. *Historical Thesaurus of the Oxford English Dictionary.* Oxford: Oxford University Press.

Kay, Christian. Forthcoming. Classification: Principles and Practice. In *Cunning Passages, Contrived Corridors: Unexpected Essays in the History of Lexicography*, edited by Michael Adams. Monza: Polimetrica.

Kay, Christian and Thomas J. P. Chase. 1987. Constructing a Thesaurus Database. *Literary and Linguistic Computing* 2: 161–163.

Roberts, Jane and Christian Kay with Lynne Grundy. 1995. *A Thesaurus of Old English.* (= *King's College London Medieval Studies XI.*) Second edition, 2000. Amsterdam: Rodopi.

Wotherspoon, Irené. 1992. Historical Thesaurus Database Using Ingres. *Literary and Linguistic Computing* 7: 218–225.

Wotherspoon, Irené. Forthcoming. The Making of the Historical Thesaurus of the Oxford English Dictionary. In *Cunning Passages, Contrived Corridors: Unexpected Essays in the History of Lexicography*, edited by Michael Adams. Monza: Polimetrica.