Getting More Out of the *Oxford English Dictionary* (by Putting More In)

[preprint]

David-Antoine Williams

St Jerome's University in the University of Waterloo

david.williams@uwaterloo.ca

Note: this is the author's archived preprint copy of an article appearing in *Dictionaries: The Journal of the Dictionary Society of North America* 38.2 (2017) pp. 106–113. The published version may be accessed at https://muse.jhu.edu/article/679603. All quotations should cite the copy of record.

ABSTRACT

This report describes a project to annotate the background coded text of the *Oxford English Dictionary* (*OED*) with metadata, as well as various current and projected outcomes. The focus now is on the 2.436 million quotations in the 1989 Second Edition (*OED2*), which are being marked for author gender, textual genre, publication type, and edition of inclusion. A future phase will address the additional 1.150 million quotations added in the current version of *OED Online* (*OED3*), as well as the non-quotation text of all three editions.

Keywords: *Oxford English Dictionary*, *OED* quotation evidence, genre, Digital Humanities, metadata, corpora.

INTRODUCTION

Readers of this journal will need no introduction to the *OED*, nor will they require detailed reminder of the various search functions facilitated by *OED* software (on CD-ROM or online) or

the kinds of research it has made possible, much of it discussed in these pages. However, in the face of the bounty of information delivered by *OED* software concerning the history of the English language, of English-language literature and culture, and of English lexicographical theory and practice, it is worth remembering a few questions that the software was not able to address in any systematic way. In her study of loanwords, Sarah Ogilvie (2013) had to rely on the (not always reliable) counts of main entries and "alien words" given in the introductions to *OED1* volumes (1894–1928) (Ogilvie, 80–81) because "alien" status has never been a search parameter in any version of *OED* software, even though the information exists in the underlying formatting code. Software is not always to blame, however: because author gender has never been recorded in the underlying data, an earlier study of female authors in the *OED* (Baigent et al. 2005) had to limit its analysis to twelve highly quoted women identified from a manual inspection of several samples.[1] Similarly, attempts to come to grips with the gnarly question of

[1] The authors consequently fail to record the contributions of two female authors who should have made their top-twelve list: "Ouida" (pseudonym of Maria Louise Ramé, aka Marie Louise de la Ramée), whose quotations number 795; and, unaccountably, Ann Radcliffe, who has 1,119. They also badly underestimate Charlotte Brontë's quotations, presumably because they did not include the 314 *OED2* citations of "C. Bronte" with their 698 references to "C. Brontë." In a proper reckoning, Radcliffe and Mary Russell Mitford would equal each other as the seventh-most quoted female authors, Jane Austen would go from eight to ninth, Brontë would be tenth (displacing Mary Whortley Montagu, Harriet Beecher Stowe, and Elizabeth Gaskell), and Ouida would be fourteenth (though still she would have tallied ahead of Baigent et al.'s lowest-ranking

textual genre in *OED* quotation evidence—most of them centered on literary genres, often lumped together—have tended to focus on highly quoted authors (e.g., Brewer 2010, Considine 2009, Willinsky 1994), giving only a partial picture of generic distribution in the dictionary. Most cripplingly for comparative dictionary study, researchers have never been able to differentiate and compare the various editions, supplements, and additions systematically, a problem Charlotte Brewer has discussed at length (Brewer 2013).

When J. C. Gray began in the late 1980s to experiment with the newly digitized *OED* to see what it could reveal about the works of John Milton, he imagined that the new resource would soon provide answers to questions "limited only by a scholar's imagination and ingenuity" (Gray 1989, 73). But this is not, in fact, what happened. Instead, in anticipating what kinds of searches scholars (and others) would most want to perform, *OED* editors had to limit "all possible searches" to "many of the most probable searches." Accordingly, most advanced studies of *OED* data have had to infer their conclusions from what evidence could be gathered from these interfaces, rather than basing them on what ideally should be gathered.

The project I describe here has roots in what I thought was a simple research question, which occurred to me about ten years ago, as a postdoctoral researcher in Oxford. It was, "what role does poetry play in the *OED*'s quotation evidence?" I soon came to realize that the question was not at all simple. For one, I could not get past a very impressionistic and anecdotal account,

---

author). There are another dozen or so women authors among the 1,000 most quoted who receive no mention at all, and more than 5,000 in the corpus as a whole.

largely based on highly cited authors and the well-known opinions of editors. I turned repeatedly to *OED Online* for some objective, quantitative basis on which to build my subjective analyses, but no proxy I could think of seemed quite valid. The labels *poet.* and *poetic* were of no use, since poetry isn't always "poetic," and non-poetry sometimes is, and anyway they were employed inconsistently and infrequently (in just 1,565 definitions in *OED2* [0.17%], illustrated by only around 8,525 quotations [0.34%]).[2] An analysis of the most-quoted authors was so partial and skewed that it was difficult to avoid question-begging and baking-in. I ruled out a sampling methodology because I wanted to make a finely grained accounting, broken down by year, or at least by century, and I suspected that, for some periods at least, what I needed to measure would be close to or smaller than the margin of error. I also wanted to know how the Second Supplement treated poetry differently than *OED1*. And finally, for all of these things, I realized that anything I wished to learn about poetry would have to be contextualized with the same information about fiction, expository prose, scientific and technical writing, and many more different kinds of text. I put the idea aside.

Shortly after that, I arrived at St Jerome's University in the University of Waterloo, where (I remembered) colleagues at the Cheriton School of Computer Science had developed the markup language originally used to encode the digital *OED* into a searchable form. The original 1989 file of *OED2* still existed and was available to University of Waterloo researchers, a legacy of the original agreement with Oxford University Press. Frank Tompa, who in the 1980s led the project

[2] As a comparison, 0.93% of definitions, containing 1.9% of quotations, are labelled *slang*.

to create this marked-up *OED*, showed me the thing: 546MB of plain text in one continuous string of 573 million characters, beginning "<E><HG><HL><LF>A</LF>." I made a copy to my mini thumb drive, and on my way back to my office at St Jerome's, I thought again about poetry in the *OED*. Now the solution really was simple, if somewhat daunting: in order to perform a valid quantitative analysis, I would have to mark the textual genre of all 2.436 million quotations in *OED2*.

A number of small grants from St Jerome's University, the University of Waterloo, and the Social Science and Humanities Research Council of Canada allowed me to do some initial development work with the *OED2* file. Then in 2015 I received an Early Researcher Award (ERA) from the Ontario Ministry of Research and Innovation—a large, five-year grant—which is funding a small team of undergraduate research assistants, normally one full-time and one part-time each term, whom I have trained to research and categorize *OED*'s text references. In the sections that follow I describe our work, our current status, and what lies ahead.

PROGRAM OF WORK

The 2.436 million quotations in *OED2* are represented by about 370,300 unique author/title combinations. At its most basic, our metadata enhancement program involves assigning to each reference one of nineteen genre categories, and noting if the work is female-authored (or co-authored), and/or if it is a serial publication (others are assumed to be books), and/or a translation. Practically speaking this involves extracting references, sample quotations, and other information from the file, creating spreadsheets in Microsoft Excel to be filled in by RAs, and then updating the file (rather, rewriting it from scratch) once the sheets have been completed and verified. The *OED* edition in which a quotation first appears is assigned automatically, based on comparisons to the 1987 Tri Star CD-ROM version of the 1928 *OED1* (quotations added in the

First Supplement will be marked out at a later date). All processing of the file itself is done with custom programs written by me in Python, a popular "high-level" programming language (similar to Java) especially suited to text applications.

The work of my RAs immerses them in the long history of English textual production. It comes with various challenges, not least the volume and variety of references to research. Readers will appreciate that, although the works of Shakespeare are cited 327 different ways in *OED2*, it is a brief task to categorize their 33,281 citations as either poetry or verse drama.[3] Given a list, an undergraduate could do this in a few minutes, and with minimal research (she might look up "Pass. Pil." and one or two other references). By contrast, marking up a like number of quotations by the least-quoted authors in *OED* would require roughly 33,000 look-ups (since they are cited only once) in one or more digital libraries: Google Books, Internet Archive, HathiTrust Digital Library, or Early English Books Online. Each look-up could take from a few seconds to a few minutes, and many would come up short. Like many projects of this kind, the "long tail" is highly inefficient.

For this reason, we employ a number of techniques to increase the efficiency of our tagging. One is simply to re-organize the list of references so as to group similar texts together (e.g., texts

[3] We do not currently differentiate among multiple genres contained in the same text: all quotations of *Macbeth*, for example, are marked as verse drama. If *OED* quotations from a text contain significant amounts of more than one genre, it is assigned one of several "mixed" categories, which may be further differentiated in a future phase.

with titles ending in a year tend to be annual reports, proceedings, and reference works; those ending in "Times" or "Herald" tend to be newspapers, and so on). Another is to include automatically generated "genre scores" to suggest a category based on characteristics of the reference or quotation text. We also try to match references to texts in other corpora, in order to glean whatever metadata might be present there. In late 2017 an Advanced Collaborative Support Award from the HathiTrust Research Center allowed us to match *OED2* and *OED3* references to the 15 million volumes in the HathiTrust Digital Library. Eventually this will allow us to regroup our list by, for example, Library of Congress Call Number (LCCN), bringing together works on similar subjects, or sorting them by textual metrics such as words per line of text, or ratios of word types (e.g. part of speech). Every method has its advantages and limitations. Call numbers, for instance, are not uniformly good predictors of genre: "PR" ("English Literature") collects texts of all literary genres, as well as the scholarship that pertains to them. In combination, however, these various methods will allow us to complete all of *OED2* by the time the ERA grant terminates in 2020.

The team's thinking about textual genre has developed considerably over the course of the project. We have always taken an inductive approach, working with the peculiarities of the *OED*'s quotation corpus to delineate areas of concentration, then thrashing out "hard cases" among team members to sharpen up the edges. Yet genre is by nature a fuzzy concept, and even people with identical training can reasonably disagree (indeed, it is not unheard of when revisiting a text to disagree with one's former self). There is also the problem of chronological range. For example, because of their prominence in the *OED* corpus, especially among nineteenth-century texts, we have separate categories for scientific and theological expository writing, but at a remove of only a few centuries these can easily converge (and meld with other

genres: take, e.g., early modern alchemical verse). The project started with five genre categories, which soon became seven. A later review increased this to nineteen, grouped within five broad registers. Even so, in the virtual Great Library that is the *OED* bibliography, there will always be the borderline, the indeterminable, and the *sui generis*.

CURRENT PROGRESS AND WHAT'S AHEAD

As of the end of 2017, we have annotated 2.137 of 2.436 million *OED* quotations (88%) under one or another of our genre categorization schemas. All quotations have been marked as belonging either to the 1928 *OED1* or the Supplements (thus we can say that the 88% coverage overall represents 92% of *OED1* quotations, and 73% of quotations added after 1928), and virtually all quotations by female authors have been identified and annotated. The program is expected to reach completion before the ERA grant terminates in 2020.

Even in its current state, the enhanced *OED2* is giving over new insights. Some of these I presented at the 2017 meeting of the DSNA in Barbados. A book chapter offering a quantitative assessment of poetry in the *OED* (Williams 2018) is my fullest answer yet to the question I asked myself ten years ago in Oxford. An earlier piece (Williams 2016a) looked closely at T. S. Eliot's presence in the Second Supplement. Several other articles and a monograph are in planning or in draft. From time to time I post preliminary analysis and discussion on my research blog (http://thelifeofwords.uwaterloo.ca/). Recent posts have discussed "alien or not fully naturalized" words (Williams 2017a), Shakespeare's first citations (Williams 2017b), author gender (Williams 2016b, 2016d), and the Second Supplement's uses of pre-1928 sources (Williams 2016c).

I mentioned that in collaboration with the HathiTrust Research Center, I have been matching *OED* references to volumes in the HathiTrust Digital Library. In addition to organizing and

deduplicating some *OED2* references, this will give us instant access to additional metadata, such as place(s) of publication, LCCN, and other bibliographical information. This is only the beginning, however. Because HathiTrust has equivalent metadata on virtually every book in any big library, when the project is complete we will be able to compare information about *OED* sources to information about all the books published after 1800 that would have been available to *OED*'s contributors and compilers. Future phases will conduct similar analyses with Early English Books Online, Internet Archive books, and the English Short Title Catalogue. Then we can begin to contend with John Considine's persuasive stance that "representative sampling" is not the proper concern of the dictionary maker (Considine 2009, 632), not least because we will know just how representative or not the *OED* is, in various ways.

In 2017 Oxford University Press provided, under licence, the XML files that form the background code of *OED3*. They have been immediately useful, both in tidying up some of *OED2*'s bibliographical and labeling inconsistencies, and in opening up *OED3* to advanced custom queries. These can extract any kind of data from and about *OED3*—information that can be used on its own (as a model for the history of the English language, for instance), or to compare new and/or revised *OED3* entries to *OED2*. Thus aspects of lexicographical practice over the long history of the dictionary project can be illuminated by quantitative analyses. Antedating is one such aspect, which can be addressed with current resources. The distribution and characteristics of *OED* sources is another, but this will have to wait until the next phase of the project.

Indeed, the plan for future work entails a full metadata enhancement of *OED3*, along with a revision of deprecated genre categories in the enhanced *OED2* (i.e., from the old five- and seven-genre schemas to the current nineteen-genre one), and refinement of the edition tag to

differentiate among Supplements and Additions and to include non-quotation text (e.g., definitions). This all will require significant new funding, which is actively being sought.

Readers who wish to explore opportunities for collaboration, either on the development of the enhanced *OED*, or on its exploitation for research, are invited to write to me by email. I am also happy to respond to specific queries about *OED* that cannot be addressed using the online interface.

REFERENCES

Baigent, Elizabeth, Charlotte Brewer, and Vivienne Larminie. 2005. Gender in the archive: Women in the *Dictionary of National Biography* and the *Oxford English Dictionary*. *Archives* (Journal of the British Records Association) 30: 13–35.

Brewer, Charlotte. 2010. The use of literary quotations in the *Oxford English Dictionary*. *Review of English Studies* 61.248: 93–125.

Brewer, Charlotte. 2013. *OED Online* re-launched: Distinguishing old scholarship from new. *Dictionaries: The Journal of the Dictionary Society of North America* 34: 101–26.

Considine, John. 2009. Literary classics in *OED* quotation evidence. *Review of English Studies* 60.246: 620–38.

Gray, J. C. 1989. John Milton and the *OED* as electronic database. *Milton Quarterly* 23.2: 66–73.

Ogilvie, Sarah. 2013. *Words of the World: A Global History of the* Oxford English Dictionary. Cambridge: Cambridge University Press.

Williams, David-Antoine. 2016a. T. S. Eliot in the *Oxford English Dictionary*. *Notes & Queries* 63.2: 296–301.

Williams, David-Antoine. 2016b. Sex in the OED. *The Life of Words* (blog). https://thelifeofwords.uwaterloo.ca/sex-in-the-oed/ (2 December).

Williams, David-Antoine. 2016c. Burchfield's reach-backs. *The Life of Words* (blog). https://thelifeofwords.uwaterloo.ca/burchfields-reach-backs/ (6 December).

Williams, David-Antoine. 2016d. OED gender genre. *The Life of Words* (blog). https://thelifeofwords.uwaterloo.ca/oed-gender-genre/ (8 December).

Williams, David-Antoine. 2017a. ||-Tripping over tramlines-||. *The Life of Words* (blog). https://thelifeofwords.uwaterloo.ca/tripping-over-tramlines/ (3 March).

Williams, David-Antoine. 2017b. Shakespeare's earliest citations in the OED. *The Life of Words* (blog). https://thelifeofwords.uwaterloo.ca/shakespeare-earliest-oed/ (6 March).

Williams, David-Antoine. 2018. Poetry in the *Oxford English Dictionary*: A quantitative profile. In *Poetry and the Dictionary*, edited by Andrew Blades and Piers Pennington. Liverpool: Liverpool University Press.