ACS Award Report – *The Life of Word*s

January 2018

Researcher: David-Antoine Williams

ACS Support: Eleanor Dickson, Boris Capitanu

Research Context

The Life of Words (LOW) is a research project in its third year (currently funded to 2020), led by me at St Jerome's University in the University of Waterloo (Canada). The aim of the project is to enhance the *Oxford English Dictionary* (OED) with metadata concerning its corpus of 3.5 million quotations.

The OED is widely considered to be the English language dictionary of record. As an historical dictionary it documents the meanings and usages of English words from the beginning of the language to the present day. The heart of the dictionary is its 3.5 million textual quotations, roughly 2 million of which were gathered, analyzed, and published between 1880 and 1928. The bulk of the remainder were added during two revisions: OED2 (1989), and OED3 (2000-ongoing). Although these quotations make up a formidable corpus of textual, linguistic, and lexicographical data, little metadata was ever included alongside it, meaning that systematic comparisons of, e.g., the gender of quoted authors, or the textual genre of quoted texts, has never been possible.

LOW is in the course of adding back metadata to the OED, principally by manual inspection of its 580,000+ textual references, and the quotations drawn from them. Having been compiled over 140 years, these are highly fragmented and heterogenous. Thus while to date we have labelled ~90% of OED2 quotations, this represents only ~60% of references: the "long tail" is laborious.

Additionally, some potentially valuable types of bibliographical metadata are too research intensive or granular (e.g. "subject area", "place(s) of publication") to be assigned from scratch on such a vast scale.

Aims and objectives

Using the heterogenous and fragmented bibliographical data in OED2 and OED3, this ACS project aimed to match OED references to HathiTrust volume IDs, in order then to draw down associated metadata. This metadata will serve four basic research objectives of the project:

1. Optimization of manual tagging regime.
   Using HT metadata, matched references can be grouped together to facilitate the assignation of LOW metadata. E.g., HT LC classifications can suggest or verify LOW genre values, and HT author full names can suggest or verify LOW author gender.

2. Supplementation of manually assigned metadata.
   Matched references can be further enhanced by additional HT metadata.

3. Automatic analysis of volumes using Extracted Features

Underwood (2009) has demonstrated the utility of the HTRC Extracted Features dataset in automatically determining textual genre. Similar methods could be applied to OED-matched volumes to assign, suggest, or verify LOW genres. Additionally, the LOW dataset, which includes texts from a longer timespan, categorized into more genres, could be used as a training set alongside the HTRC Extracted Features dataset in future automatic classification experiments.

4. Wide-scale comparison of corpora
   "Representativeness" has been a major research question for OED scholars. Metadata from matched OED references can be compared on the same basis to the entire HT corpus, or subsets thereof, to provide a truer sense of how the OED citation corpus reflects the available potential documentation.

Methods

OED quotation data typically includes some combination of author name, work title, publication date, and quotation text, each of which may appear in a number of formats (e.g., author name might be a full name, or initials and a last name, or just a last name, or even just an abbreviation). Some texts, such as periodicals, have no author information. Complicating matters further, the OED's idea of a "text" varies, and often has no direct analogue in the HT Digital Library or any other textual corpus. Sometimes a text is a poem, sometimes a Collected Poems; sometimes it's an authored article in a periodical, other times the periodical *tout court*.

We implemented a three-step matching process, primarily to achieve computational practicability, and secondarily to reduce false positives while minimizing false negatives as far as possible. As false negatives cannot be recuperated, and false positives may in theory be eliminated on further review, on balance we erred on the side of capturing more false positives.

For each step I provided a datafile of OED references to search against the HT Digital Library, generated from my OED2 and OED3 reference data. Searches were performed separately on HT "books" and "non-books", depending on whether an author name was present in the OED reference. Only works published after 1749 were searched.

Step 1 matched OED Author Last Name (if present), Sequential (but not necessarily consecutive) OED Title Words, and Publication Date. Positive matches from this search were considered high confidence unless the number of matched volumes exceeded an arbitrary threshold, on the basis that only common or low-information data would tend to overmatch, given these three criteria.

Step 2 was performed on OED references left unmatched in Step 1. This matched OED Author Last Name (if present) and Sequential OED Short Title Words. Publication Date Range (+/- 5 years) was used to limit results only if an Author Last Name was not present (i.e. for serials). Step 2 matches were considered medium-confidence unless the number of matched volumes exceeded a second arbitrary threshold, in which case they were considered low-confidence.

Step 3 was performed using all matched references from Steps 1 and 2 that exceeded the low-confidence threshold for that step. Step three searched for a number of word-level trigrams from the OED quotation text within the full text of the low-confidence matched volumes.

Results

Of the ~450k OED2 and OED3 references searched, this process yielded 25% high-confidence matches, 4% medium-confidence matches, 16% low-confidence matches, and 55% null matches. Book references fared better than nonbook references, with 29% high confidence-matches and 49% null matches, vs. 12% high-confidence matches and 80% null matches.

Null matches tended to be for works published after 1930. On manual inspection of a small sample of these, the majority were True Negatives, in the sense that no volumes were found to exist in the HT Digital Library. The majority of these, however, were found in other repositories, usually Google Books.

With additional post-hoc reconciliation of OED references, using these matches I have been able to provide a HT volume ID for 68% of OED2 references 1750-1990, and 72% of OED2 references 1800-1928. These references are responsible for 82% and 86% of the quotations from their time periods, respectively.

Outcomes

Using the HT metadata associated with matched volumes, I have been able to reorganize and reprioritize the project's OED reference inspection and tagging program (objective 1). Student RAs have recently begun using the new system and are evaluating its contribution to their work.

I have also been able to use associated HT metadata to identify ~5k fugitive female-author quotations in OED2 (just 0.2% of all quotations, but +6% of previously identified female-author quotations) (objective 1).

Future work

There is a substantial amount of additional manual verification work to do on the matched volume dataset before it can be used for reliable analysis (objectives 2, 3, 4). This will continue apace, with the goal of having a highly reliable match profile for OED1 works published 1800-1928 by the end of the current project phase (2020), as well as at least a sketch of profiles for 1928-1990 works in OED2 and post-1990 works in OED3.

In order to capture references absent from HT at the time of the search, but present, e.g., in Google Books, it would be highly desirable to repeat the process with a smaller set of outstanding OED references after some future expansion of the HT library.