AOR Data Release 2.0: Status of the transcriptions

13 July 2016

Compared to the previous AOR Data Release 1.0 (November 2015), the only major difference is the inclusion of another book, Thomas Tusser's *Fiue hundred pointes of good husbandrie* (for more information on this book, see the blog by Kristoff Smeyers and the one written by Jaap Geraerts). All the reader's interventions in this book have been transcribed and translated into English (in the case of non-English marginal notes) adding 94 marginal notes, 13,160 underscored words, 1,495 marks and 59 symbols to the AOR dataset. In total, the AOR dataset consists of 5,827 marginal annotations; 241,111 underscored words; 31,402 marks; 2,857 symbols; 223 numerals and 6 drawings.

Another significant difference is that since the last data release we have continued checking and cleaning up our data. *All transcriptions have been cross-checked*, but we have not completely finished the work on the AOR dataset. Currently, we are still in the process of dealing with around 75 remaining issues, ranging from the translation of enigmatic marginal notes to trying to figure out yet undecipherable marks and unfamiliar symbols. A tiny minority of Harvey's annotations thus are shrouded in uncertainty, but all of his annotations are transcribed and, more importantly, the uncertainty the transcribers might have about some of his annotations is highlighted in the transcriptions themselves. We aim to solve all these issues before the end of the first phase of AOR, and any successful resolutions of these issues will be included in AOR Data Release 3.0, which is scheduled for late August. Any unresolved issues will be carefully documented and put on the AOR website. Please also note that a set of applied statistical analyses will be conducted using the dataset represented in this Data Release 2.0, and will be further represented by a report and useful visualizations shortly thereafter.

For more information on the dataset, please consult the documentation accompanying the first data release or, for even more detailed information, the Transcriber's Manual. Both documents can be found on the downloads page of the AOR website.

NB: Some of the .csv files which are part of our data analysis are encoded in UTF-8 and the normal Windows encoding will result in gibberish when opening them. In order to avoid this, open Microsoft Excel, go to the tab 'Data', select 'From text', select the .csv file you want to open, press 'Import' and then select for file origin 'Unicode (UTF-8). This applies to all the .csv files of which the name starts with 'vocab_'.

Suggested citation:

Archaeology of Reading Data Release, version 2 (July 2016), based on data from www.bookwheel.org