The Goal of the Analysis Process

The business activity that revolved around books during the early modern period is a widely researched topic. Despite earlier qualitative research in this area, quantitative attempts to analyze printing in the early modern period have been few and limited in scope due to the large amount of tedious manual work that is required for such analysis. Recent developments in improvement of historical bibliographical metadata quality achieved by the Helsinki Computational History Group (COMHIS) have made highly automated approaches possible. Also, successful Named Entity Recognition (NER) of individuals appearing in the metadata has enriched it with a vantage point from which automated estimation of book trade can start. The aim of the analysis pipeline presented here is to significantly decrease the amount of work needed when doing such estimates with bibliographical metadata.

The Evaluation Method

In the iterative development to make the process work, we tested its output for London printers from 1614 to 1618 against the largely manually derived estimates of Gants finally reaching the level presented in this poster. The dataset we used was the English Short Title Catalogue, but the process can be applied to any collection, and it’s not ad hoc adjusted to the period of the replication, although the comparison point was vital for the development of the process. The process does some of the harder parts like merging of duplicate records in the metadata automatically, but the steps of merging significantly differing variants of printer names (detected automatically) and the merging of individuals into Printing Houses still needs to be done manually.

Visualisation of the Process

Subset of the ESTC

Manual cleaning of NER-based table of printers

Greedy parsing of the imprint

Matching the cleaned table with the imprint

Missing value estimation

Duplicate removal

Merging individuals to Printing Houses

Aggregation of variables by Printing Houses

Edition Sheet Distribution

Comparison

Errors for the Physical Variables

Normalised Errors of Edition Sheet Estimates

Comparison of Total Numbers

Co-printing Network

Methods and Validation

As the graphs suggest, the replication of Gants estimations was largely successful for the physical variables. As illustrated at the figure 2, the average errors were at the range of 8 to 23 percent for these, well at the domain of being reliable enough for statistical analysis of the general structures of the trade, although caution is needed when making assessment about individual printing houses. Only a couple of works were attributed to people entirely outside of printing in London during this period. The manual parts, cleaning of the NER-based individuals and merging them to printing houses, required to make the final replication required approximately days worth of work.

Conclusions

We conclude that the method is already useful for obtaining accurate enough results that it can be applied. However, further development is needed in the NLP methods used to make progress in automated topic or self-publishing classification of editions.

The process is not tied to printers, and it can be used to gain similar insight into publishers and booksellers. Also, the possibility of using the process for the research of interrelationships within and between different parts of the book printing industry, as illustrated in the figure Co-printing Network, needs further examination.