Digital Methods for Corpus Expansion in Early Modern Philosophy Research

Andrea Sangiacomo, Raluca Tanasescu, Silvia Donker, Hugo Hogenbirk
University of Groningen, Faculty of Philosophy

This poster presents the initial corpus expansion stage and its attendant digitally-inflected methodologies of the ERC-funded research project "The Normalization of Natural Philosophy: How Teaching Practices Shaped the Evolution of Early Modern Science." The main contention of our project is that these teaching practices were socially embedded and had a decisive 'normalising' impact on the progressive dissemination, adaptation, and selection of rival conceptions of natural philosophy (Sangiacomo 2018). Two major axes of research are followed to this end. The first one involves social network analysis, which reconstructs the networks of authors and sources that were involved in the debate on early modern natural philosophy and their evolution over time. The second one uses semantic network analysis, which rebuilds the networks of concepts of natural philosophy, their linguistic context, and vocabulary and explains why certain concepts and approaches become accepted as standards, and which elements determined this output.

The poster outlines the methodology used for the initial building of the corpus on which the two kinds of network analyses will be carried out. The team have departed from the existing dictionaries of Dutch (van Bunge et al. 2003), British (Yolton et al. 1999 and Pyle 2000), and French (Foisneau 2008) early modern philosophers (1600-1800) and created a dataset of associated relevant works divided into three broad categories meant to differentiate between their contribution to the dissemination and normalization of teaching practices: 'primary'—clearly systematic in nature, most comprehensive, most likely to be used as teaching materials and offer less doubts about the fact that they concern natural philosophy as a whole; 'secondary'—similar to primary works except that they are not necessarily systematic (i.e. student disputations), often offering a glance at more specific core issues that are debated in the discipline across time and space; and 'tertiary'—not necessarily connected with natural philosophy but show a relevant use of key notions, debates or trends in natural philosophy for related topics and discussions. In order to expand these sets of canonical writers and works and to explore the 'unread' debate on early modern natural philosophy, we have derived lists of frequent words and frequent collocations from the titles as follows: for each language (Latin, French, and English), for each corpus (Dutch, British, and French), and, within each corpus, for various time clusters (periods during which works of natural philosophy have been published regularly, without significant gaps between two consecutive titles). Since we have not performed "whole-text" searches (Tangherlini and Leonard 2013) and we have rather faced small-sized departing corpora of titles in the existing bio-bibliographical dictionaries (several hundred words each), simple word frequency counts were considered as sufficient (Graham et al. 2012). We retained both keywords that were semantically

related to ‘physics’ or ‘natural philosophy’ or two-word collocations consisting of a more general term, such as the Latin philosophia, and a more specific one (e.g. (e.g. naturalis). We set the bar to three occurrences for single keywords and to two for collocations and lowered the threshold to two occurrences in the case of small corpora, such as the first French time cluster of only seven titles.

The lists of keywords and collocations obtained via the open-source concordancer AntConc were then used for focused web-crawling and scraping the WorldCat catalog by means of a Python script, which collected a total of over 74,000 titles published in Latin, English, and French between 1600 and 1800. As expected, the corpus contained a very large number of duplicates, as well as numerous instances of titles not pertaining to the field of philosophy due to the semantic loading of several keywords. The large amount of scraped information was first cleaned, simplified and tokenized via the pandas, numpy, os, string, and NLTK Python libraries. Subsequently, the .csv files created for each keyword and collocation were further analyzed and processed to remove all duplicates, as well as to remove the data already present in the three existing dictionaries. The data deduplication was done by substring similarity matching via the FuzzyWuzzy library, which assigned similarity scores to authors and titles and removed those with values generally over 60%. The remaining data were then analyzed using close reading in order to eliminate any works not pertaining or not related to the field of philosophy, to break down the final results into the three working categories and, finally, to divide the resulting lists of authors by nationality and retain those of interest.

The proposed poster will present the proof of concept for corpus expansion and will demo the Python scripts used throughout this initial stage of the project, which we consider relevant for any kind of data-intensive bibliographic research.

Works cited:

Foisneau, Luc. 2008. *Dictionary of Seventeenth-Century French Philosophers*. London; Oxford; New York; New Delhi; Sydney: Bloomsbury.

Graham, Shawn; Weingart, Scott; Milligan, Ian. 2012. “Getting Started with Topic Modeling and MALLET.” *The Programming Historian* 1, Web: http://bit.ly/2OomNba.

Pyle, Andrew. 2000. *Dictionary of Seventeenth-Century British Philosophers*. London; Oxford; New York; New Delhi; Sydney: Bloomsbury.

Sangiacomo, Andrea. 2018. “Modelling the history of early modern natural philosophy: the fate of the art-nature distinction in the Dutch universities,” *British Journal for the History of Philosophy*, DOI: 10.1080/09608788.2018.1506313.

Tangherlini, Timothy R.; Leonard, Peter, 2013. “Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research.” *Poetics*, 41(6), pp.725-749.

van Bunge, Wiep; Krop, Henri; Leeuwenburgh, Bart; Schuurman, Paul; van Ruler, Han; Wielema.

Michiel. 2003. *Dictionary of Seventeenth and Eighteenth-Century Dutch Philosophers*. London; Oxford; New York; New Delhi; Sydney: Bloomsbury.

Yolton, John; Price, Valdimir; Stephens, John. 1999. *Dictionary of Eighteen-Century British Philosophers*. London; Oxford; New York; New Delhi; Sydney: Bloomsbury.