Simon Hengchen (STSM)

The Hartlib Papers and the COST action “Reassembling the Republic of Letters” The Hartlib Papers is the collection of correspondence, personal papers, and documents of interest to the ‘intelligencer’ Samuel Hartlib (c.1600-1662). The epistolary corpus consists of 4,833 letters written by some 325 correspondents, spread mostly over Western and Central Europe. The largest part of Hartlib’s papers is kept in the University of Sheffield Library, which has published online scans and transcription of their own collection as well as of those in other libraries and archives. Devoted to an ideal of a reformed society, Hartlib had his correspondents send him information on developments in philosophy, science, technology, medicine, theology, agriculture and education. His letters are known to be an extremely rich source for the study of intellectual life in seventeenth-century Europe. In addition to covering a wide spectrum of topics, the letters span over a relatively long period of 42 years, from 1620 to 1662. The letters are also particularly informative of the network through which this information found its way to Hartlib. Given the number of people mentioned in the letters, the reach of his network must have been more than double the number of his correspondents. It consisted of people with diverse socio-economic backgrounds, such as merchants, clergymen, academics, courtiers, artisans, etc., some of whom held important political, ecclesiastical, academic, diplomatic or military positions. The geographical spread of the network and the large number of non- academics among the correspondents also account for the multilinguality of the corpus, consisting of 3,015 English letters, 527 German letters, 859 Latin letters, 84 French letters, 3 Dutch letters, and 345 letters mixing multiple languages.

The size of the collection, its multilinguality, and the great variety of subjects discussed in the Hartlib Papers make the collection difficult to access for the individual researcher, and offer methodological challenges which make them a very useful testbed for application to other epistolary corpora. Methods from Natural Language Processing and computational linguistics to analyse the letters’ topic distribution and semantics have not yet been applied to the Hartlib Papers and would enable large-scale content analyses of this rich corpus for the first time. Exploring this large quantity of epistolary data by using topic modelling and semi-automatic semantic change detection techniques will help researchers to mine the content of the Hartlib Papers and to better understand the interests shared by Hartlib’s correspondents. It will also make this epistolary corpus more accessible and help to map the network behind it. More broadly, the development of such automatic methods for topic extraction and semantic change detection, while being tested in the challenging context of the Hartlib Papers, will offer us the opportunity to develop a general methodology that can be applied to other epistolary corpora. This research will thus contribute to the objectives of the COST action both by developing innovative scholarly methods and by allowing impactful research questions to be answered.

The project within which the mission will take place has two main scientific aims. The first aim is to extract topics contained in the letters, using a combination of Natural Language Processing tools for automatic topic extraction and Dr Buning’s expert input in the validation and refinement phases. Particular attention will be given to the letters in English and in Latin, as those are the languages more represented in the letters. This will allow us to answer important questions regarding what subjects were discussed most in the letters, if the subjects changed over time and which subjects were discussed in which parts of the network. It will also offer a general methodology to be applied to other similar collections. The second aim of the project is to use these topics as a basis to study semantic and conceptual shifts in the English and Latin letters of the Hartlib Papers collection, and this will be the focus of the scientific mission, as detailed below. This will have two main benefits for the research community. First, it will allow us to answer historical research questions that will shed new light into conceptual change over the time span covered by the Hartlib Papers. Semantic change detection in the Hartlib Papers would show to what extent the content of the terms that were central to their ideas and ideals, and which they discussed extensively, such as ‘pansophy’, ‘college’, ‘desiderata’, changed over the years. Second, this research will create the algorithmic basis to allow researchers to explore the letters from a semantic point of view. The computational detection of semantic change will assign meaning to the words in the letters, thus making it possible to add a semantic search layer to the letters at a later stage. This contribution will also have the potential to be extended to other collections, and therefore special attention will be devoted to the generalizable methodological outcomes of this process. Scope of the mission The project started in late 2016 and has already produced a set of automatic processes for cleaning and pre-processing the Hartlib Papers. Moreover, topic extraction methods for the English letters have already been tested and refined, and will form the methodological basis for the next analyses. Before the start of the mission, we plan to conduct the pre-processing and topic extraction phase for the Latin letters. This way, we will be able to focus our efforts during the mission on the semantic change modelling for the English letters, as well as the general methodological contributions.

Following the mission, we will concentrate our attention on applying the semantic change detection methods to the Latin letters and writing up the results of the research for publication in a chapter for the COST volume in the section “IV.7 Text-mining the Republic of Letters”. This plan has already been discussed with the lead author of the section, Charles van den Heuvel, who fully supports this project. A more extensive publication will follow and will be submitted to the journal Digital Scholarship in the Humanities published by Oxford University Press. Whilst the first phase of the project has been and will be conducted remotely, it is the study of semantic change that will require an active, focussed collaboration, and is therefore particularly suited for the STSM. Indeed, as a careful study of the state of the art undertaken during my doctoral dissertation (handed in on August 29, 2017) demonstrates, the discovery of semantic change is a complicated task that draws from various domains: history, (computational) linguistics, but also computer science, data science and, in some sense, artificial intelligence.

With that in mind, it is important to stress that undertaking such an interdisciplinary work requires an interdisciplinary team. This short-term scientific mission would allow this interdisciplinary team to work optimally towards the completion of the project. The discovery of semantic change in dirty, OCRed, historical texts in the Dutch language is the focus of my PhD research and my expertise, which I will continue to develop in the next years in the context of my post-doctoral position on semantic and conceptual change analysis at the University of Helsinki. At the same time, the multilingual nature of the Hartlib Papers corpus and the computational challenges involved in processing the texts in Early Modern English and neo-Latin require additional skills. Dr Buning is an expert of neo- Latin and has extensive knowledge of the Hartlib Papers, which were the focus of his post-doctoral research in Oxford (Buning 2014; 2017; forthcoming). Dr McGillivray is a computational linguist with a strong background in mathematics, and is an expert in the field of Latin computational linguistics (McGillivray 2013). Her research at the Alan Turing Institute focuses on computational models for semantic change detection in historical texts. Additionally, the Alan Turing Institute, as a leading hub for high-profile data science research, is an optimal institution to employ state-of-the-art computational techniques to answer important research questions in the Humanities. This is further strengthened by the recent creation of a Data Science and Digital Humanities Interest Group, led by Dr McGillivray, which provides an ideal context for this kind of research, particularly given its methodological focus (cf. also Jenset & McGillivray, 2017). A close collaboration during the mission, in an environment optimised for such research, is critical for the completion of the project.

This mission’s contribution to the project and to the COST action “Reassembling the Republic of Letters” will deliver the following results: Develop code for computational models for semantic change detection from the English and Latin letters; Automatically create lists of words that shifted from one topic to another and from one meaning to another in the corpus; Compile a set of methodological recommendations for carrying out this research in the context of epistolary corpora As detailed above, the research conducted as part of the mission will form the basis for the publication of the chapter in the volume which constitutes one of the main outputs of the COST action. This research will also give rise of an article publication, which will further extend the chapter published in the COST action's volume. The paper will make use of learned lessons and best practices from previous work using topic modelling to tackle historical text, especially in those languages and periods (Wittek & Ravenek, 2011; Roorda et al., 2010), thus furthering the state of the art both in history and in historical computational linguistics and digital humanities.

References:

- Buning, Robin (2014). 'A Circle and Its Centres: A Glimpse into the Structure of the Hartlib Circle'. Presented at 'The Practise of Scholarly Communication'.
- Buning, Robin (2017). 'Collecting Biographies of the Members of Samuel Hartlib's Circle: A Prosopographical Approach to Networking the Republic of Letters'. Presented at 'Reception, Reputation and Circulation in the Early Modern World, 1500-1800'.
- Buning, Robin (forthcoming). 'Scholarly Communication between England, the Netherlands and Central Europe: Samuel Hartlib's Dutch Agent Caspar Streso (1603–1664)', in Correspondence Networks between Central and Western Europe, 1550–1700 (forthcoming).
- Greengrass, Mark, Robertson, Alexander M., Schinke, Robyn, & Peter Willett (1996). 'Processing morphological variants in searches of Latin text'. Information Research 2:1.
- Jenset, Gard & Barbara McGillivray (2017). Foundations of quantitative historical linguistics. A corpus framework. Oxford: Oxford University Press.
- McGillivray, Barbara (2013). Methods in Latin Computational Linguistics. Leiden: Brill.
- Rogers, Heather J. & Willett, Peter (1991). 'Searching For Historical Word Forms In Text Databases Using Spelling-Correction Methods: Reverse Error And Phonetic Coding Methods', Journal of Documentation 47:4 (1991), 333-353
- Roorda, D., Bos, E. J., & van den Heuvel, C. (2010). Letters, ideas and information technology: Using digital corpora of letters to disclose the circulation of knowledge in the 17th century. Proceedings of Digital Humanities, London, UK.
- Schinke, Robyn, Greengrass, Mark, Robertson, Alexander M. & Willett, Peter (1997). Retrieval Of Morphological Variants In Searches Of Latin Text Databases, Computers and the Humanities 31:5, 409–432.
- Wittek, P., & Ravenek, W. (2011). Supporting the Exploration of a Corpus of 17th-Century Scholarly Correspondences by Topic Modeling. Presented at the SDH 2011 Supporting Digital Humanities: Answering the unaskable.