Letters, Ideas and Information Technology: Using digital corpora of letters to disclose the circulation of knowledge in the 17th century

Roorda, Dirk
dirk.roorda@dans.knaw.nl
Data Archiving and Networked Services, Netherlands

Bos, Erik-Jan
erik-jan.bos@phil.uu.nl
Department of Philosophy, Utrecht University, Netherlands

van den Heuvel, Charles
charles.vandenheuvel@vks.knaw.nl
Virtual Knowledge Studio, Netherlands

1. Circulation of Knowledge and Letters

The scientific revolution of the 17th century was driven by countless discoveries in Europe and overseas in the observatory, in the library, in the workshop and in society at large. There was a dramatic increase in the amount of information, giving rise to new knowledge, theories and world images. But how were new elements of knowledge picked up, processed, disseminated and – ultimately – accepted in broad circles of the educated community? A consortium of universities, research institutes and cultural heritage institutions has started a project called CKCC[1] to meet this research question, building a multidisciplinary collaboratory to analyze a machine-readable and growing corpus of letters of scholars who lived in the 17th-century Dutch Republic. Until the publication of the first scientific journals in the 1660s, letters were by far the most direct and important means of communication between intellectuals. Therefore the 17th-century Republic of Letters offers an ideal case for exploring the answers to this question.

Researchers want to uncover patterns in letters that are indicative for the circulation of knowledge, patterns that reveal the emergence of complex, collective phenomena in modern science. However, they face some fundamental problems with finding such patterns in letters. One cannot know in advance the nature of these patterns, and only few categorical hypotheses can be tested by simply data mining the letters. Purported patterns cannot be tested against the letters, because the heterogeneous information on which these patterns are based cannot be gleaned from the texts, but need considerable interpretation and contextualization.

Here is a short list of the problems: (i) the letters are not uniformly available; (ii) the 17th century language varieties are not standardised and pose a challenge for language technology; (iii) much interpretation is needed to resolve references to people, places, dates, ideas and instruments; interpretations are complicated by the heterogeneity of annotations; (iv) it is not clear how to set up visualisations of patterns that are really informative to the historian of science. These four types of problems will be used to report on the methodology of the project and on its results so far.

2. Information technology as a humanities' observatory

2.1. Availability of the Letters

CKCC limits itself to the ca. 20,000 letters written by scholars that were active in the Netherlands: René Descartes, Hugo Grotius, Constantijn Huygens, Christiaan Huygens, Caspar Barlaeus, Jan Swammerdam and Anthony van Leeuwenhoek. Modern editions of these correspondences—already published or in an advanced state of production by members of CKCC—form the basis of the digitised texts. The letters, once converted to a minimal TEI format, will then be made available through e-Laborate,[2] a web-based philological annotation tool that will be transformed into a collaboratory for the history of science and the humanities in general. It serves three purposes: (a) providing scholarly access to the letters; (b) allowing researchers to enrich existing datasets and annotate the letters; (c) using the letters and the input of

researchers to visualise patterns meaningful for the circulation of knowledge.

2.2. Use of other datasets

We will incorporate a particular database of (meta)data, the *Catalogus Epistularum Neerlandaricum* (CEN), or the Catalogue of letters in Dutch repositories. It is a relatively old database, already available via Telnet in the early 1990s, before the world wide web came into being.

CEN is an exhaustive database of letters in the collections of five Dutch university libraries, the Royal Library, and four other important libraries. It contains more than 265,000 descriptions of approximately 1,000,000 letters, dating from 1600 until the present day (of which ca. 100,000 from the 17th century). It supplies the following metadata: sender, recipient, place of sending, year, language, repository and shelf mark.

The format in which this database will be made available to the project is to be negotiated with the owner, OCLC.[3]

Usage of this database will enable us to make assertions about the fraction of the selected letters with respect to the total body of letters. Moreover, it allows us to increase the density of the networks we are interested in, leading to unprecedented research opportunities.

2.3. Language technology

In order to find meaningful patterns in social networks of scholars and in circulation of knowledge language technology is needed. For this, CKCC is cooperating with CLARIN.[4] The mission of CLARIN is to make language technology interoperable and to make linguistic resources accessible on a European infrastructure, so that all the arts and humanities can make use of it. The Netherlands pillar of CLARIN, CLARIN-NL,[5] has already obtained funding for constructing such infrastructure, and has issued a call for proposals for adding existing resources to this infrastructure and writing demonstrator services. Aided by expertise provided by CLARIN members, in particular by the University of Lancaster,[6] CKCC is developing such a demonstrator. A proposal to this end has been accepted by CLARIN-NL. The demonstrator, comprising the correspondences of Grotius, Const. Huygens and Descartes (ca. 15,000 letters in all), is planned to be completed by October 2010. It will perform a time-sensitive keyword extraction, which can be visualised by means of a dynamic word cloud. As the source languages are 17th century Dutch, French and Latin, one needs at least spelling normalisation and harmonisation of keywords across languages.

2.4. Interpretation and Enrichment

References to people, places and times are often implicit and can only be retrieved by studying contextual material or by using secondary sources. Named Entity Recognisers are helpful, but it is not possible to rely on technology alone. In order to get an accurate picture in sufficient resolution, interplay between manual work and automatic tools is needed. The collaboratory based on e-Laborate gives researchers the opportunity to collect their interpretations of the texts, compare them to others and to annotate them with their insights. Over time, the results of this hand/mind work might be automatically gathered and incorporated in enriched transcriptions of the texts.

2.5. Visualisation

By offering meaningful visualizations of the data, the CKCC will enable humanities researchers in a wider context to use the tools and the results yielded. Not only the relationships between corresponding authors will be made visible in time and space, but CKCC also aims at visualizing the dynamics of knowledge production by focusing on the emergence of themes in scholarly debates and social networks of 17th century natural philosophers.

The dynamic word clouds based on keyword extraction is just a first step. CKCC will subsequently explore several approaches of gathering and visualising meaningful patterns, which are deliberately different in nature. The first approach (a) is a sophistication of keyword extraction, and the second one (b) is based on associations in text. Both methods can be used to evaluate the results of each other.

(a) Concept analysis. This requires considerable more analysis than keyword extraction. For example: the many surface expressions of a concept must be linked into one entity, preferably part of an ontology or thesaurus. Existing subject indices and reference corpora will be used. Visualising the behaviour of concepts over time will yield a good approximation of knowledge circulation.

(b) Associative neural network technology. This is an application of a recent effort, ANNH,[7] to apply the idea of neural networks to the humanities. This approach enables the automatic comparison of texts, based on the degree of associative similarity. Concepts or themes occurring in the letters could thus be made visible, either by focussing on single terms (e.g. 'observations') or word pairs (e.g. 'soul' and 'matter'). Moreover, it is possible to query for letters associated to a given one, and rank the results by the degree of association. The facility to track the circulation of knowledge will then be within reach.

3. An infrastructure for the digital humanities

Infrastructure is of particular interest in view of current developments in Europe as testified by the ESFRI roadmap.[8] The roadmap funds the preparation for several research infrastructures in the humanities, among which CLARIN, as demonstrated above, is most relevant for the CKCC project. CKCC will take care that the materials residing in the collaboratory can be exported in such a way that it is available on the CLARIN infrastructure, thus contributing to the much grander vision to have all scholarly letters of the 17th century uniformly available for research, including the results of related work.[9]

In due course, CKCC will not only contribute to the understanding of the circulation of knowledge in the 17th century, but also generate useful technologies for cross-disciplinary collaborations involving data-sharing and data-enrichment in the Humanities. As such, this web-based humanities collaboratory on correspondences is a valuable prototype for possible future research collaborations focusing on large, heterogeneous datasets in the Humanities.

References

Holthausen, K and Ziche, P. (2007). 'Neuronale Netze für die Geisteswissenschaften'. *Akademie Aktuell*. **20**: 32-35. http://www.badw.de/aktuell/akademie_aktuell/2007/heft1/10_Holthausen.pdf .

Notes

1. Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic. A Web-based Humanities' Collaboratory on Correspondences. Participants: Descartes Centre for the History and Philosophy of the Sciences and the Humanities, Utrecht University; National Library of the Netherlands; Huygens Institute, http://www.huygensinstituut.knaw.nl/__eng/ (accessed 2010-03-19); DANS; Virtual Knowledge Studio. Start date: November 2008. Duration: 4 years. Budget: 1 M€.
2. e-Laborate, http://www.e-laborate.nl/en/ (accessed 2010-03-19), Huygens Instituut.
3. Online Computer Library Center, http://www.oclc.org (accessed 2010-03-19).
4. Common Language Resources and Technology Infrastructure, http://www.clarin.eu/ (accessed 2010-03-19). CKCC is on the list of supported projects by which CLARIN is reaching out to the humanities, http://www.clarin.eu/wp3/wp3/wp3-documents/call-for-full-proposals-for-collaboration-with-humanities-and-social-sciences (accessed 2010-03-19).
5. CLARIN-NL (Netherlands), http://www.clarin.nl/node/2 (accessed 2010-03-19).
6. University Centre for Computer Corpus Research on Language, http://ucrel.lancs.ac.uk/ (accessed 2010-03-19), in particular Paul Rayson, http://www.comp.lancs.ac.uk/~paul/ (accessed 2010-03-19).
7. Associative Neural Networks for the Humanities. An application developed by the Department of Philosophy (P. Ziche, E.-O. Onnasch, E.-J. Bos), Utrecht University, and Dr. Holthausen GmbH, Bocholt. See (Holthausen et al., 2007).
8. European Roadmap for Research Infrastructures, http://cordis.europa.eu/esfri/roadmap.htm (accessed 2010-03-19).
9. Mapping the Republic of letters, http://shc.stanford.edu/collaborations/supported-projects/mapping-republic-letters (accessed 2010-03-19). Cultures of Knowledge, http://www.history.ox.ac.uk/cofk/ (accessed 2010-03-19).