CHAPTER 26

The ePistolarium: Origins and Techniques

Walter Ravenek, Charles van den Heuvel and Guido Gerritsen
Huygens ING, Postbus 10855, 1001 EW Amsterdam, The Netherlands

ABSTRACT

The *Circulation of Knowledge: A Web-based Humanities' Collaboratory on Correspondences and Learned Practices in the 17th-century Dutch Republic* (CKCC) project was an NWO project aimed at developing an infrastructure for researchers. Its main goal was to gain insight into the Dutch share of the circulation of knowledge in the 17th-century 'Republic of letters' by means of analysis and visualisation tools. A database of 20,000 letters in TEI-format offered the possibility to falsify hypotheses that were often based on extrapolations from limited numbers of letters. The complexity of this collection of data – caused by the presence of multilingual letters, often with several languages within a letter, extensive spelling variation, and early modern language variants – was a challenge for the researchers and IT specialists. With the support of CLARIN-NL and the EU, however, we were able to overcome these linguistic problems.

26.1 Introduction

This chapter deals with various aspects of the *ePistolarium*, a virtual research environment for browsing and analysing a corpus of letters written by and sent to 17th-century scholars who lived in the Dutch Republic. Firstly, we describe the project from which the ePistolarium arose, named *Circulation of Knowledge: A Web-based Humanities' Collaboratory on Correspondences and Learned Practices in the 17th-century Dutch Republic* (CKCC; ePistolarium, 2013). Secondly, we give an overview of the analysis methods that are available to the users of the ePistolarium, emphasising the role of Natural Language Processing techniques.

How to cite this book chapter:
Ravenek, W, van den Heuvel, C and Gerritsen, G. 2017. The ePistolarium: Origins and Techniques. In: Odijk, J and van Hessen, A. (eds.) *CLARIN in the Low Countries*, Pp. 317–323. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi.26. License: CC-BY 4.0

26.2 Project

Whereas internationally renowned projects such as *Cultures of Knowledge* (CofK; 2016) and *Mapping the Republic of Letters* (RofL; 2013) which both more or less started at the same time as the CKCC project focused on metadata to get insight into the circulation of knowledge, the CKCC project focused on analysing the content of the letters themselves. The project aimed at providing researchers with tools for answering questions related to the dissemination and appropriation of knowledge:

1. How did knowledge circulate in the 17th-century Dutch Republic? How were elements of knowledge – generated in workshops, at sea, in the colonies overseas, on the battlefield and in libraries – picked up and used by the learned community? How was this new knowledge processed, disseminated, theorised and ultimately accepted, or, for that matter, rejected?
2. How can we combine and structure various sets of letters of 17th-century scholars and their correspondents in such a way that we can analyse the circulation and appropriation of knowledge production in a wider international context and recognise the development of themes of interest and scholarly debates in space and time?
3. How can we search and contextualise this information on knowledge production and its appropriation to make it accessible to interdisciplinary research in the Humanities?

As will be discussed in Section 26.5, researchers also use the ePistolarium to answer their own research questions, which are more specific than the general questions formulated at the start of the project.

Wijnand Mijnhardt (Descartes Centre University of Utrecht) was the principal investigator of the CKCC project, while Huygens ING was responsible for the technical development and implementation of its infrastructure, its analytical and visualisation tools and its user interface. CKCC was funded from November 2008 to February 2013 by the Netherlands Organisation for Scientific Research (NWO) as part of the Investment Grant NWO Medium programme to provide a large group of researchers with suitable tools to analyse and visualise the circulation of knowledge in the Dutch Republic. CLARIN-EU selected the CKCC project as a 'flagship' demonstrator to show the potential of the CLARIN infrastructure, not only for linguistics but for other humanities disciplines as well. CLARIN-NL provided extra funding for the adaptation of language and knowledge representation technologies for keyword extraction and concept extraction. Finally, after its completion in 2013, CKCC was one of the projects selected by CLARIN-NL for the development of an educational module. In June 2013, the ePistolarium was officially launched and made public. Its web location is ckcc.huygens.knaw.nl/epistolarium.

26.3 Corpus

The CKCC corpus currently contains 20,020 letters, ignoring duplicates. It consists of the correspondence of the 17th-century scholars Caspar Barlaeus, Isaac Beeckman, René Descartes, Hugo Grotius, Christiaan Huygens, Constantijn Huygens, Antoni van Leeuwenhoek, Dirck Rembrantsz van Nierop, and Jan Swammerdam. Most of this correspondence was already digitised at the start of the project, but the formats of the letters differed widely, requiring us to convert the letter texts to a standardised format, for which we chose TEI. In addition, the metadata of the letters needed to be standardised, and concordances of person and place lists to an aggregated dataset needed to be prepared. Even though we use a limited number of metadata tags (date, correspondents and sender/recipient locations) considerable effort was needed to standardise and enrich the metadata.

From a language perspective our corpus has a number of characteristics that are important when it comes to processing the letters:

- The corpus contains letters in various languages, the most important ones being Dutch, French and Latin. As can be seen from Table 26.1, these three languages account for almost 95% of the text.
- Many letters are multilingual. In order to apply language resources and technology we have to segment the letters to at least the paragraph level.
- The letters often contain elaborate opening and closing phrases that contribute little to the subject matter of the letters. Currently we have some 10,000 opening and 17,500 closing sections marked up. It is worthwhile to exclude such sections from content extraction.
- Finally, 17th-century writing exhibits a large degree of spelling variation, which has a negative effect on the performance of analysis techniques.

26.4 Analysis Techniques

The ePistolarium offers its users a range of analysis techniques, the most advanced ones being topic modelling and cocitation analysis, both of which depend on natural language processing. In this section we give a concise overview of the techniques employed in the ePistolarium.

26.4.1 Language Identification

For language identification – applied at the paragraph level – we used the N-gram based cumulative frequency addition algorithm (Ahmed, Cha and Tappert, 2004). The text is preprocessed by removing punctuation, quotes, mathematical symbols, digits and Roman numerals. The algorithm requires language profiles which are constructed using a selected set of monolingual letters from the corpus.

26.4.2 Spelling Normalisation

Of the three major languages in the corpus, Dutch exhibits most spelling variation; it is also the language that differs most from its modern counterpart. In addition it is the language of the lesser educated correspondents, e.g. Antoni van Leeuwenhoek.

Language	Paragraphs	Tokens	Rel. size
Dutch	37,570	2,491,730	30.33%
English	865	85,290	1.04%
French	26,747	2,810,484	34.21%
German	2,728	106,180	1.29%
Greek	46	709	0.01%
Italian	2,011	63,203	0.77%
Latin	38,331	2,458,403	29.92%
Portuguese	2	614	0.01%
Spanish	26	2,110	0.03%
Not Assigned	27,202	196,672	2.39%
CKCC corpus	135,528	8,215,395	

Table 26.1: Corpus size by language.

Correspondence	Recognised	Identified
Barlaeus	3,073	1,460
Beeckman	139	120
Descartes	4,096	3,917
Grotius	76,790	57,286
Chr. Huygens	21,647	17,411
Const. Huygens	17,354	12,632
Van Leeuwenhoek	899	869
Van Nierop	394	326
Swammerdam	567	532
CKCC corpus	124,959	94,553

Table 26.2: Results of NER for person references. *Recognised* names are annotated in the letter texts; *identified* names are used in cocitation analysis.

We decided to use the sophisticated spelling normalisation application VARD 2 (Baron and Rayson, 2008) for Dutch, and handled spelling variation in French and Latin with a rule-based approach. The basic philosophy of VARD 2 is to normalise text to modern spelling, allowing existing linguistic tools to be used unmodified. It was developed and trained to deal with spelling variation in Early Modern English, but can also be trained to deal with spelling variation in other languages. We used a version of VARD 2 that was adapted by its author Alistair Baron to allow integration in our text-processing pipeline.

26.4.3 Named Entity Recognition

We used Named Entity Recognition (NER) to label person names in the letter texts – the availability of identified names is a prerequisite for cocitation analysis.

We used an iterative, rule-based approach to build gazetteers (lookup lists of names), which were extended with hand-annotated names and names from indexes of book editions. More names were generated by applying rules to Latinised names (for instance, if 'Grotius' and 'Grotio' occur the names 'Grotium' and 'Grotii' are also generated). For the actual matching the well-known Aho-Corasick (1975) algorithm is used on a normalised representation of the gazetteers and the letter texts. The normalisation involves removing diacritical marks and applying the character mappings j → i, y → i, v → u, and w → u; this normalisation is language-independent and works well for 17th-century texts. (See Table 26.2 for NER statistics.)

26.4.4 Keyword Analysis

We performed a keyword analysis for the three main languages in the corpus, following an approach similar to the one implemented by Rayson in the Wmatrix corpus analysis tool (Rayson, 2008). The analysis is based on frequency profiling of the individual letters and comparing the obtained profiles with the corresponding profile of the full letter collection as a reference corpus. Keywords are determined with a log-likelihood estimator, using a threshold of 99% confidence of significance. We thus obtained keywords for 82% of the letters; these results are displayed in the ePistolarium.

26.4.5 Topic Modelling

Topic modelling constitutes a statistical approach to content extraction. The major approaches to topic modelling are able to identify hidden variables that can be interpreted as 'topics'. We tested three topic modelling methods: Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Random Indexing (RI). We found that RI performed best in the task of reproducing topic labels assigned by human experts for a randomly selected subset of letters (Wittek and Ravenek 2011). From a computational point of view RI has the benefit that it does not rely on computationally intensive matrix operations as, for example, LSA does. Instead, RI builds an incremental word space model that scales very well with increasing corpus size. For these reasons we employed RI as the topic modelling method in the ePistolarium.

Preprocessing for the actual calculation of the topic model involves some general processing (e.g., elimination of opening and closing phrases, of formulas, and of words with a length smaller than three characters) and some languagespecific processing (e.g., removal of stop words, spelling normalisation). We calculated a single topic model for all languages combined.

In the ePistolarium the topic model is used for calculating similarities between letters and between words. To illustrate the latter we describe the query term suggestion feature (see Figure 26.1): the ePistolarium offers a fulltext search, implemented with the *Lucene* search library. The user enters search terms and can request query terms that have the largest cosine similarity with the terms entered. These new, suggested terms can be transferred to and used in the regular fulltext search. Two examples of such query term suggestions are:

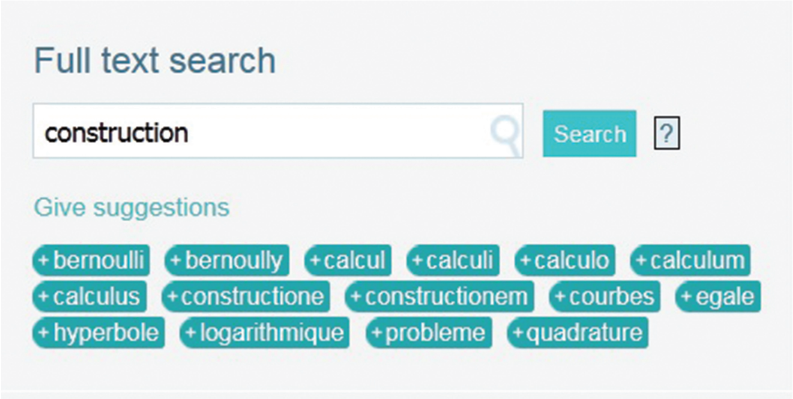
- construction → bernoulli, bernoully, calcul, . . ., courbes, egale, hyperbole, logarithmique, probleme, quadrature; and
- roi de france → affaires, angleterre, espagne, espagnols, gens, guerre, paix, reine

Based on the qualitative judgement of users (see Section 26.5) we find that the best results are obtained for terms that pertain to a topic that has a specific terminology.

26.4.6 Cocitation Analysis

One of the visualisations offered by the ePistolarium is a cocitation network graph. It shows individuals that play a role in an intellectual debate and their connections. The graph is constructed using references to individuals in the same *paragraph* of a letter: the more often individuals are mentioned together, the stronger their connection.

In our analysis we excluded the opening and closing phrases of the letters. Obviously, only identified person names can be taken into account. The analysis is performed by combining data for the selection of letters made by the user with the faceted search and/or the fulltext search (see Figure 26.2 for an example of a cocitation graph).



Full text search

construction

Figure 26.1: Example of query terms suggested with topic modelling.

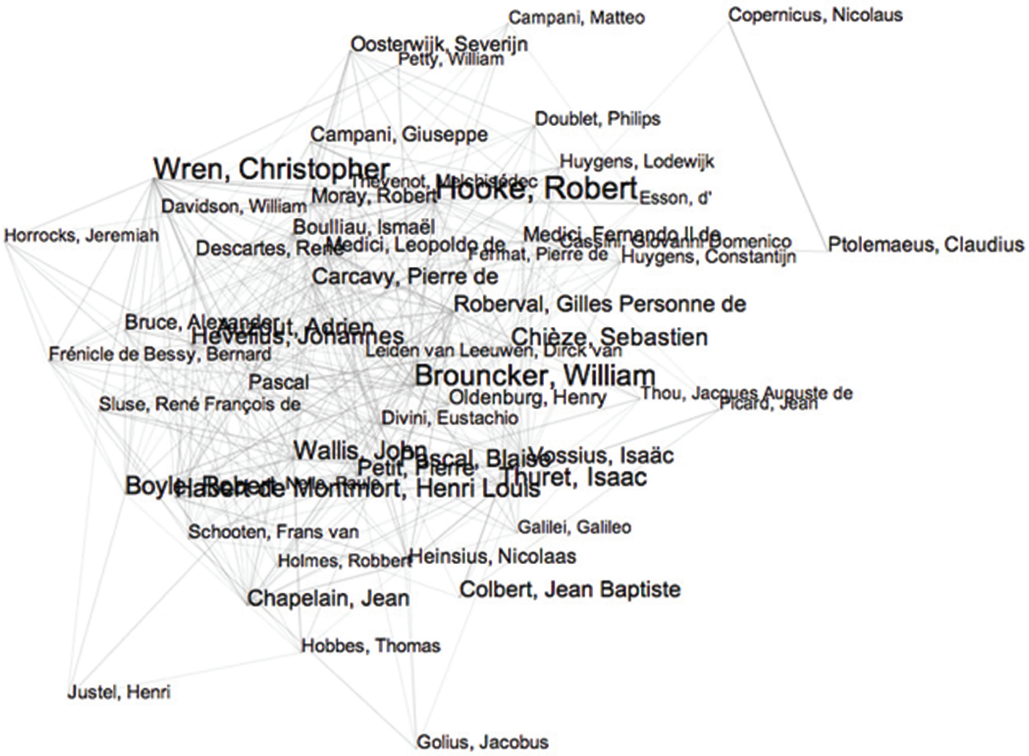


Figure 26.2: Example of a co-citation graph.

26.5 Usability Tests

During the development of the ePistolarium we organised various usability tests by historians of science, both with researchers involved in the project and with participants of the NIAS-Lorentz workshop *Mathematical Life in the Dutch Republic* (Leiden, 2010) The feedback led to improved versions of the faceted search and to a shift of focus in the application of topic modelling. The interpretation of the 'topics' proved hard and ambiguous; therefore we decided to use topic modelling to calculate similarities between (paragraphs of) letters and for search term suggestions.

When the ePistolarium was launched in June 2013, new experiments were set up that are all fully documented (ePistolarium, 2013) The use of co-citation analysis proved quite succesful. The use of topic modelling requires knowledge of its limitations: firstly, it seems that our corpus is still too limited in size, with a relatively small overlap in subject matters. Secondly, it seems that the unit of modelling (paragraphs) is not fine-grained enough, as paragraphs in the letters tend to be long and to cover various different subjects. Thirdly, search term suggestion yields the mostusable results for subjects with a specific terminology (Heuvel et al., 2016).

Several research projects using the ePistolarium were set up outside the CKCC consortium. For instance, Wouter Klein and Toine Pieters (2016) used the ePistolarium tool to reveal the hidden history of an exotic therapeutic drug in the correspondence of Constantijn and Christiaan Huygens.

26.6 Outlook

To remedy the problem of the small size and imbalance in the composition of the corpus we intend to include more data in the CKCC corpus with the last three volumes of the Van Leeuwenhoek correspondence and the correspondence of Pierre Bayle and Carolus Clusius to be added; other correspondence will follow. In order to obtain more data the CKCC project plays an active role in the European project *Reassembling the Republic of Letters* (RROL; 2015).

One of the major disadvantages in our current approach to topic modelling is that the texts in the various languages each 'live' in their own subspace of the overall word space. Using a translation to a common language, preferably English, would allow us to build a unified topic model covering the bulk of the text in the corpus.

Although the analysis methods in the ePistolarium are dynamic in the sense that they are applied to selections made by the user, the texts themselves and the annotations made on them (e.g. person identifications) are static. We intend to extend the ePistolarium by allowing user annotations to be made. This will be accommodated by using the storage mechanism provided by Alexandria, an annotation environment that is currently being developed at the Huygens ING. Alexandria will be compliant with the CLARIAH infrastructure, the largest digital infrastructure for the humanities in the Netherlands.

References

Ahmed, B, Cha, S, and Tappert, C 2004 *Language identification from text using n-gram based cumulative frequency addition*. In: *Proceedings of Student/Faculty Research Day*, CSIS, Pace University.

Aho, A V and Corasick, M 1975 Efficient string matching: An aid to bibliographic search. *Communications of the ACM* 18, 333–340.

Baron, A and Rayson, P 2008 *VARD 2: A tool for dealing with spelling variation in historical corpora*. In: Proceedings of the Postgraduate Conference in Corpus Linguistics, Birmingham, UK.

CofK 2016. Available at http://www.culturesofknowledge.org.

ePistolarium 2013. Available at http://ckcc.huygens.knaw.nl.

Heuvel, C van den, Weingart, S, Spelt, N and Nellen, H 2016 Circles of Confidence in Correspondences Confidentiality in seventeenth-century knowledge exchange in networks of letters and drawings. *Nuncius* 31, 78–106.

Klein, W and Pieters, T 2016, The Hidden History of a Famous Drug: Tracing the Medical and Public Acculturation of Peruvian Bark in Early Modern Western Europe (c. 1650–1720). *Journal of the History of Medicine and Allied Sciences*, DOI: http://dx.doi.org/10.1093/jhmas/jrw004.

Rayson, P 2008 From key words to key semantic domains. *International Journal of Corpus Linguistics* 13, 519–549.

RofL 2013. Available at http://republicofletters.stanford.edu.

RROL 2015. RROL- ISCH - COST-action IS1310 *Reassembling the Republic of Letters*. Available at http://www.republicofletters.net.

Wittek, P and Ravenek, W 2011 *Supporting the Exploration of a Corpus of 17th-Century Scholarly Correspondences by Topic Modeling*. In: B. Maegaard (Ed.), *Supporting Digital Humanities 2011: Answering the unaskable*. Copenhagen, Denmark.