Mapping the Bentham Corpus

Estelle Tieberghien, Frédérique Mélanie-Becquet, Pablo Ruiz Fabo, Thierry Poibeau, Melissa Terras, Tim Causer

Mapping the Bentham Corpus

Estelle Tieberghien, Frédérique Mélanie-Bécquet, Pablo Ruiz and Thierry Poibeau
Laboratoire LATTICE, Paris

Melissa Terras and Tim Causer
University College London

1 Introduction

University College London (UCL) owns a large corpus of the philosopher and social reformer Jeremy Bentham (1748-1832). Until recently, these papers were for the most part untranscribed, so that very few people had access to the corpus to evaluate its content and its value. The corpus is now being digitized and transcribed thanks to a large number of volunteers recruited through a crowd-sourcing initiative called Transcribe Bentham (Causer and Terras, 2014a, 2014b).

The problem researchers are facing with such a corpus is clear: how to access the content, how to structure these 30,000 files, and how to get relevant access to this mass of data? Our goal has thus been to produce an automatic analysis procedure aiming at providing a general characterization of the content of the corpus. We are more specifically interested in identifying the main topics and their structure so as to provide meaningful static and dynamic representations of their evolution over time.

2 Comparison with other works

The exploration of large corpora in the Humanities is a known problem for today's scholars. For example, the recent PoliInformatics challenge addressed the issue by promoting a framework to develop new and original research in text-rich domains (the project focused on political science but can be extended to any sub-field within the Humanities).

Specific experiments have recently been done in the field of philosophy, but they mainly concern the analysis of metadata, like indexes or references (Lamarra and Tardella, 2014; Sula and Dean, 2014). Different experiments have nevertheless involved an exploration of large amounts of textual data (see e.g. Diesner and Carley, 2005 on the Enron corpus) with relevant visualization interfaces (Yanhua et al., 2009).

In this paper, we propose to explore more advanced natural language processing techniques to extract keywords and filter them according to an external ontology, so as to obtain a more relevant indexing of the documents before visualization. We also explore dynamic representations, which were not addressed in the above-mentioned studies.

3 Corpus exploration strategy

3.1 The Text analysis module

Different scripts have been developed to filter the corpus[1]. Then documents are assigned a date whenever possible: Since the corpus mostly contains notes and letters, the first date mentioned in the document often refers to the date of the document's composition (even if this assumption is of course not always true). A large number of documents cannot be assigned a date and are thus not used for the dynamic analysis of the corpus.

To index the corpus and identify meaningful concepts, we first tried to directly extract relevant keywords from the texts. However, traditional techniques like the use of tf-idf (Salton et al., 1983) and c-value (Frantzi et al., 2000) do not seem very efficient in our case. This is not too surprising: it is well known that texts are too ambiguous to provide a sound basis for a direct semantic extraction. Surface variations, the use of synonyms and hyponyms, linguistic ambiguity and other factors constitute strong obstacles for the task. We thus decided to use natural language processing techniques that provide relevant tools to overcome some of these limitations. The tools we employed are either web-based or possible to execute on a personal computer with average specs.

We tried to refine concept extraction by confronting the text with an external, structured database. We used DBpedia (Auer et al., 2007) as a source of structured knowledge (DBpedia is a database made of information extracted from Wikipedia). DBpedia is not a specialized source of information but this guarantees that the approach is not domain or author specific and could be easily used for other corpora. We then used the DBpedia Spotlight Web Service (Mendes et al., 2012; Daiber et al., 2013) to make the connexion between the corpus and DBpedia concepts. This leads to a much more fine grained and relevant analysis than possible with an entirely data-driven keyword extraction.

Based on the outputs of Spotlight, only concepts that occurred at least 100 times, and with a confidence value of at least 0.1 were kept. Spotlight outputs a confidence value between 0 and 1 for each annotation; a 0.1 threshold removes clearly unreliable annotations while maintaining good coverage. Tagging the full corpus with Spotlight (ca. 30,000 documents) took over 24 hours. We called the Spotlight service one document at a time; parallelizing the process can decrease processing time.

3.2 The visualization module

Once relevant concepts are identified, one wants to produce relevant text representations so as to provide a usable interface to end users. We present here three different kinds of interfaces that show the possible exploitation of the analysis described above.

---

[1] For example, Bentham sometimes used French in his correspondence and these texts are eliminated via automatic language detection, since we focus on English in this experiment.

The corpus is first indexed in a Solr search index[2] and accessible through a graphical end-user interface. It is possible to query the corpus by date, using Solr's faceted search functions[3] (see figure 1).

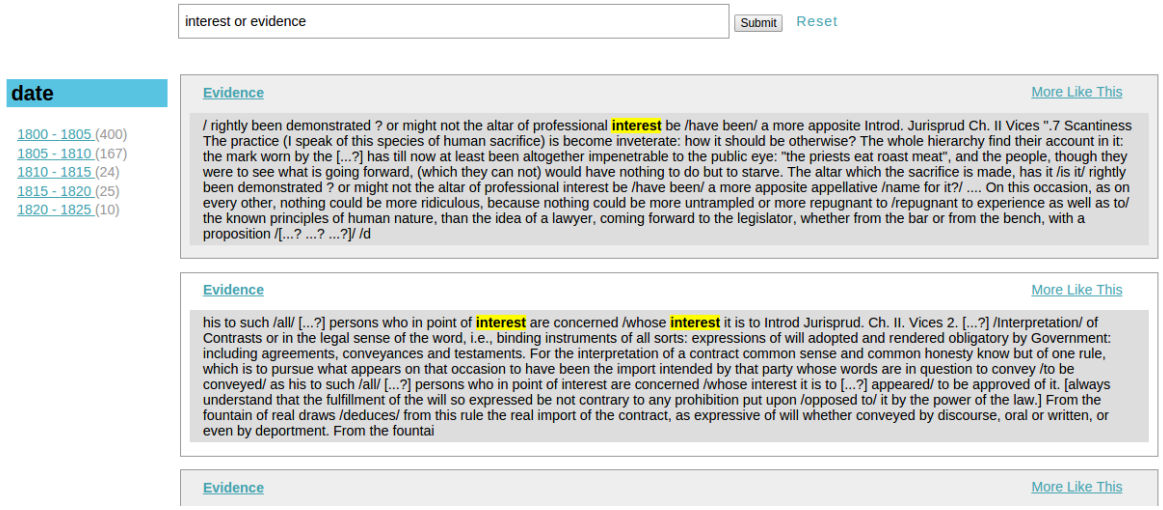

Figure 1: Search interface: users can search via year extracted from the text, which in most cases is the year of writing, allowing users to see texts (especially correspondence) in chronological order.

It is also possible to cluster together related keywords, so as to get access to homogeneous sub-parts of the corpora representing specific subfields of Bentham's activity (see figure 2).

[2] https://lucene.apache.org/solr/
[3] https://wiki.apache.org/solr/SolrFacetingOverview

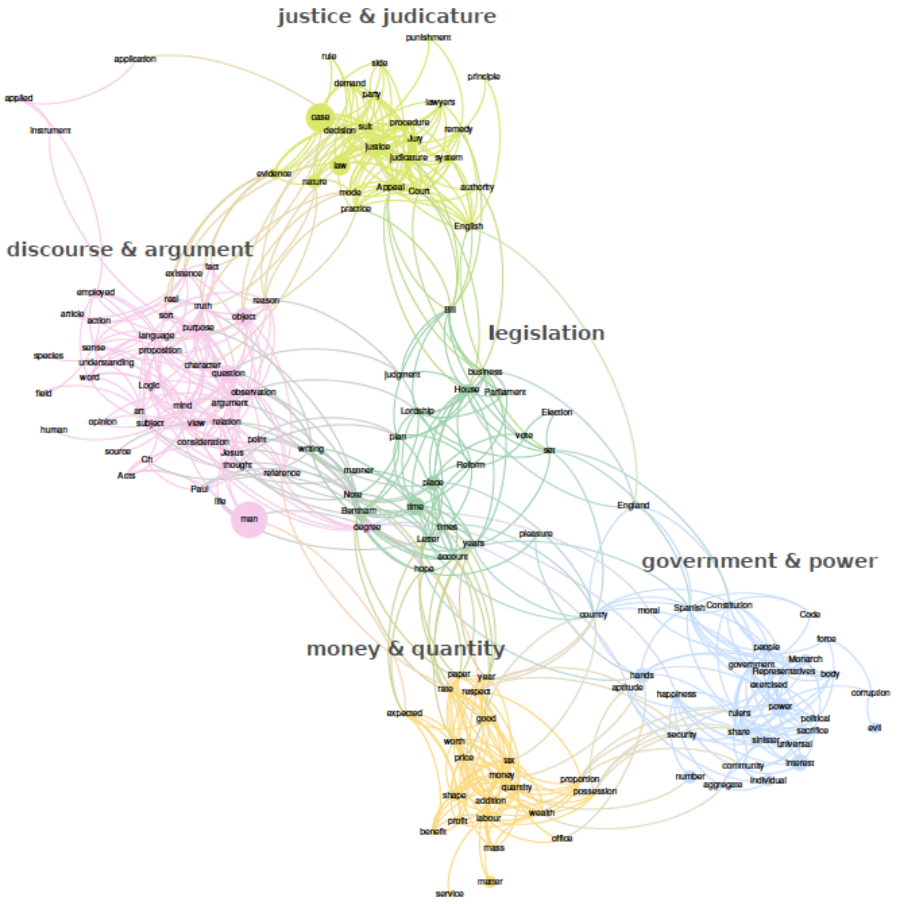

Figure 2: the main topics addressed in the corpus, based on clusters of concepts, showing the main concerns of Bentham's writings, which map closely onto established research areas in Bentham studies. The network was produced by Cortext; colours and fonts were reformatted in Gephi based on Cortext's gexf-format export[4]

Dynamic maps are also possible, to see for example the evolution of the different topics addressed in the corpus over time (see figure 3).

[4] https://gephi.github.io/

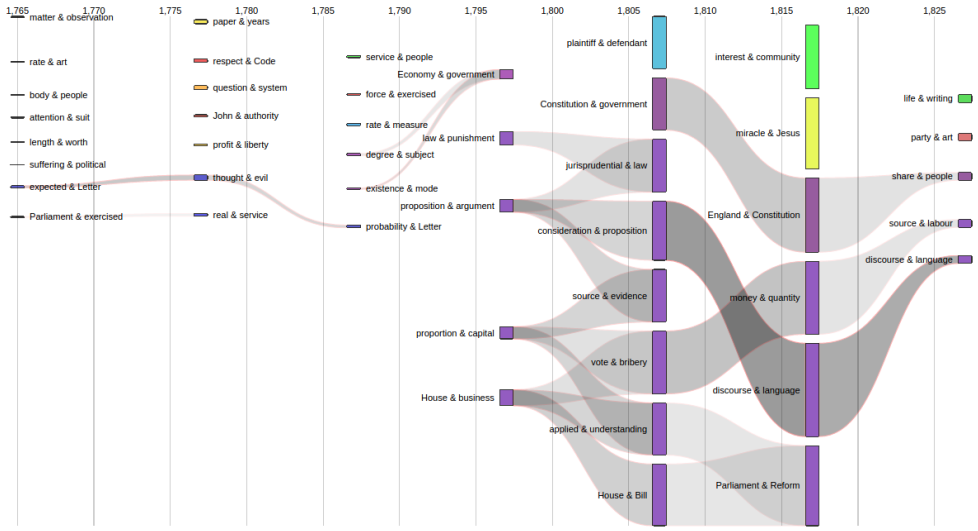

Figure 3: A dynamic view of the corpus, computed with the Cortext plarform *(tubes layout)*, with the evolution of the main topics addressed over time

4 Scholarly benefits of these tools for the Transcribe Bentham project

Since 1958, UCL's Bentham Project has been producing the new, critical edition of the "Collected Works of Jeremy Bentham". The edition is expected to run to some eighty volumes, the thirty-third of which has recently been sent to the press. The "Collected Works" is based upon texts, which Bentham published during his lifetime, and unpublished texts, which exist in manuscript. It is a major task: UCL's Bentham Papers runs to some 75,000 manuscript pages, while the British Library's has a further 25,000 or so pages. About 40,000 pages have been transcribed to date and, while UCL's award-winning 'Transcribe Bentham' initiative has helped to significantly increase the pace of transcription, and great deal more work needs to be done.

The first task in producing a volume of the "Collected Works" based upon texts in manuscript is to identify all the relevant pages. Bentham Project editorial staff use the Bentham Papers Database Catalogue, which indexes the manuscript collection by sixteen headings, including date, main heading, subject heading(s), author(s), and so forth. It is, however, entirely possible to miss relevant manuscripts using this method. The subject maps produced for this research promise to complement traditional Bentham Project methods; for instance, Bentham's work on political economy encompasses topics as varied as income tax to colonisation, and the subject maps will make it more straightforward to investigate the nexus between these, and other, subjects.

The dynamic corpus view, showing the evolution of topics addressed over time, could also prove useful in editorial work as can be shown in two examples. First, an editor at the Bentham Project is currently working on Bentham's writings on convict transportation, though there is some confusion over when exactly Bentham first broached the topic. The dynamic corpus view could help to clear up whether it was only around 1802 when Bentham wrote about transportation, or if he had investigated the subject in any great detail during the 1790s. Second, Bentham became more radical as he aged, and several Bentham scholars have sought to identify the point at which

Bentham abandoned his earlier conservatism and 'converted' to political radicalism, and representative democracy; an analysis of Bentham's language at

5 Conclusion

In this paper, we have presented a first attempt to give a relevant access to a large interdisciplinary corpus in the domain of philosophy, law and history. We have shown that using tools in concept clustering and visualization can provide an alternative way to navigate large-scale corpora, and confirm and visualise scholarly approaches to large scale textual corpora. Exploring how these tools can be effectively used with a corpus such as Bentham's indicates these methods are applicable to other sources as well.

In the near future, we are planning to refine the linguistic analysis in order to give better representations of the textual content of the corpus. We are also planning experiments with end-user to evaluate in more details the solution and the visualisation techniques used so far in this project.

References

Auer, Sören, et al. (2007). DBpedia: A nucleus for a web of open data. *The Semantic Web*. Springer Berlin Heidelberg.

Causer, Tim, and Terras, Melissa (2014a). Many hands make light work. Many hands together make merry work: *Transcribe Bentham* and crowdsourcing manuscript collections, in *Crowdsourcing Our Cultural Heritage*, ed. M. Ridge, Ashgate, 2014.

Causer, Tim, and Terras, Melissa (2014b). Crowdsourcing Bentham: Beyond the Traditional Boundaries of Academic History, *International Journal of Humanities and Arts Computing*, vol. 8 (1), pp. 46-64, 2014.

Daiber, Joachim ; Jakob, Max ; Hokamp, Chris and Mendes, Pablo N. (2013). Improving Efficiency and Accuracy in Multilingual Entity Extraction. Proceedings of the 9th International Conference on Semantic Systems (I-Semantics).

Diesner, Jana and Carley, Kathleen M. (2005). Exploration of Communication Networks from the Enron Email Corpus. Proceedings of SIAM International Conference on Data Mining, Workshop on Link Analysis, Counterterrorism and Security, pg. 3- 14, Newport Beach, CA, April 2005.

Frantzi, Katerina ; Ananiadou, Sophia and Mima Hideki (2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, Volume 3, Number 2, Page 115.

Lamarra, Antonio and Tardella, Michela (2014). Theophilo. A prootype for a thesaurus of philosophy. Digital Humanities 2014, Lausanne, Switzerland.

Mendes, Pablo N. ; Daiber, Joachim ; Rajapakse, Rohana, Sasaki, Felix and Bizer, Christian (2012). Evaluating the Impact of Phrase Recognition on Concept Tagging. Proceedings of the International Conference on Language Resources and Evaluation, LREC 2012, 21-27 May 2012, Istanbul, Turkey.

Salton, Gerard, Edward A. Fox, and Harry Wu. (1983). Extended Boolean information retrieval. *Communications of the ACM 26.11*: 1022-1036.

Sula, Chris Alen and Dean, Will (2014). Visualization of Historical Knowledge structures: an Analysis of the Bibliography of Philosophy. Digital Humanities 2014, Lausanne, Switzerland.

Yanhua, Chen; Lijun, Wang; Ming, Dong and Jing, Hua (2009). "Exemplar-based Visualization of Large Document Corpus". IEEE Transactions on Visualization and Computer Graphics (InfoVis2009), vol.15, no.6, pp.1161-1168, Nov.-Dec. 2009.