Mapping the Bentham Corpus: Concept-based Navigation

Pablo Ruiz, Thierry Poibeau

Journal of Data Mining and Digital Humanities

Mapping the Bentham Corpus: Concept-based Navigation

Pablo Ruiz Fabo[1] **and Thierry Poibeau**[2]

[1]Laboratoire LiLPa, Université de Strasbourg
[2]Laboratoire Lattice, CNRS, École normale supérieure / PSL, Université Sorbonne nouvelle

Corresponding author: Pablo Ruiz Fabo, Thierry Poibeau
ruizfabo@unistra.fr, thierry.poibeau@ens.fr

Abstract

British philosopher and reformer Jeremy Bentham (1748–1832) left over 60,000 folios of unpublished manuscripts. The Bentham Project, at University College London, is creating a TEI version of the manuscripts, via crowdsourced transcription verified by experts. We present here an interface to navigate these largely unedited manuscripts, and the language technologies the corpus was enriched with to facilitate navigation, i.e Entity Linking against the DBpedia knowledge base and keyphrase extraction. The challenges of tagging a historical domain-specific corpus with a contemporary knowledge base are discussed. The concepts extracted were used to create interactive co-occurrence networks, that serve as a map for the corpus and help navigate it, along with a search index. These corpus representations were integrated in a user interface. The interface was evaluated by domain experts with satisfactory results, e.g. they found the distributional semantics methods exploited here applicable in order to assist in retrieving related passages for scholarly editing of the corpus.

Keywords

Jeremy Bentham; manuscripts; corpus navigation; entity linking; keyphrase extraction

I INTRODUCTION

With the development of digital technologies, communication networks and large storage capacities, a large effort has been done to digitize all kinds of content, especially cultural heritage ones. This is true for public libraries, private companies (especially Google with the Google Books initiative) but also for more modest institutions archiving or interested in specific collections.

The Bentham project, launched in 1959 at University College London (UCL), aims at producing a new edition of the work and correspondence of Jeremy Bentham, especially the previously unpublished texts archived at UCL and the British Library. Thirty-three volumes of the new Collected Works have been published so far, and 50 years have been necessary to scholars to transcribe 20,000 folios. This edition is thus an important but slow process due to the amount of work to be carried out by a rather small team.

However, instead of waiting for another 50 years or more to see the end of the project, people involved in it observed that the transcription process could be sped up a lot thanks to modern technologies. This gave birth to a new project called Transcribe Bentham, that started in 2010: the main goal of Transcribe Bentham is to use a crowdsourcing platform to help with the task. A very precise and controlled workflow has been defined: Anybody can transcribe a manuscript, but the result is checked by a Bentham scholar, corrected and added to the list of transcribed

documents and rejected otherwise (which means that, if the proposed transcription is not of a good enough quality and would require too much work to be corrected, the folio is just kept in the list of manuscripts to be transcribed until a new, satisfactory, transcription is produced). Note that nothing is fully automatic in this process, and thus all the manuscripts are carefully checked by an expert before being marked as transcribed. The idea is to speed up the transcription process without decreasing the quality of the output. The work of the Bentham project team members has just changed: instead of transcribing all the material themselves, they now check the candidate transcriptions produced by others, besides still transcribing when necessary.

Since 2010, 20,000 folios have been transcribed, which means as many folios have been transcribed in 8 years than in the previous 50 years. Around 650 people have transcribed something but only 30 transcribers are "super transcribers", doing most of the work. The project fortunately succeeded to get some press coverage, which helped attract people, make them visit the website and participate in the project. As always with this kind of projects, one needs to gain a lot of interest (the website has received nearly 100,000 visits since its beginning) to be able to recruit only a handful of very active participants. These are highly motivated people, generally producing high quality work since more than 94% of the transcribed texts have been added to the database after being checked and corrected (which means less than 6% of the transcribed texts are rejected, mainly because they have been only partially transcribed). The transcription effort is however far from complete and several years will still be necessary to complete the whole task, hopefully within two decades from now, if the work continues at the current speed.

The corpus is nonetheless still largely unpublished and, even if available, hard to access given the number of files and the amount of text to read. Then practical questions arise, such as how to access this corpus and more generally large collections of texts from a specific author or a set of different authors. General purpose search engines may be useful, but they are not sufficient as there is also a need to address specific questions, like the following:

- What topics did Bentham address during his career as a philosopher?
- How have these topics evolved over time? On what topic was Bentham working at any specific period of time?
- What are the connections between the different topics?

Answering these questions means implementing relevant text analytics tools, in order to be able to extract relevant pieces of information, cluster and structure this information into meaningful representations and provide conceptual maps of the philosopher's ideas over time. We also need to evaluate these representations and check with domain experts that they really make sense and can be useful for experts as well as for a more general audience.

Performing this analysis means having recourse to content analysis tools. These tools have progressed a lot in the last two decades, thanks to new techniques based on machine learning and the large amount of data available for training. The consequence is that natural language processing (NLP) techniques are now mature enough to analyze large textual corpora. Our bet is that these tools, although they were initially not intended to process eighteenth-century texts, are now robust enough to process non-standard texts, including philosophical ones, even if such material is expected to pose difficulties for the technology. This is the sense of the set of preliminary experiments we have done over the Transcribe Bentham manuscript collection. The challenge is thus to evaluate the adequacy of existing NLP techniques, develop specialized techniques whenever possible (especially for lexical and terminological extraction, which is

central in our approach) and apply them to the Bentham corpus. A thorough evaluation of the result with specialists of Bentham will then ensure the quality of the output, its interest and the possibility to generalize the approach to the rest of Bentham's works, and to other works from other philosophers in the future.

The rest of the paper is structured as follows. Section II describes the corpus, our sample, and the preprocessing carried out before the corpus can be analyzed with text mining tools. Section III covers prior work on the corpus. Section IV describes the technologies we applied in order to create concept networks. Two concept annotation technologies were exploited (4.1): Entity linking to DBpedia and keyphrase extraction. We discuss the difficulties in applying a contemporary knowledge-base like DBpedia to historical text. The navigable corpus maps created on the basis of the lexical extraction are also discussed (4.3). Section V presents the user interface that integrates the corpus maps with a search index. Finally, section VI consists in a qualitative evaluation of the interface with domain experts.

II THE CORPUS: BENTHAM'S MANUSCRIPTS

2.1 Digitization and Transcription by UCL's Bentham Project

Jeremy Bentham (17481832) was a British philosopher and reformer, known as the founder of utilitarianism, which proposes that the ethical measure of an action corresponds to the extent to which it promotes the greatest happiness of the greatest number. He developed a theory of punishment in agreement with this principle, stressing deterrence and rehabilitation. He was also a proponent of female suffrage and a theorist of representative democracy [Causer and Terras, 2014a]. He wrote on a vast range of subjects, from political economy to religion and sexual morality. Expressing some of Bentham's ideas would have been punishable in his days, and such content remained unpublished during his lifetime. However, Bentham produced over 60,000 folios of manuscripts, thanks to which we are aware of his views on topics like the above. The Bentham Project,[1] at University College London (UCL), is creating a new edition of Bentham's Collected Works [Bentham, 1968 – ongoing], taking into account input from these manuscripts. Bentham Project scholars started transcribing the material and catalogued the corpus, adding metadata like dates, document types and others. Since 2010, the manuscripts are being digitized at UCL and are being transcribed by volunteers thanks to the Transcribe Bentham crowdsourcing initiative [Causer and Terras, 2014b, i. a.], also coordinated by UCL. As a collaboration between the LATTICE Lab and UCL, we had the chance to run text mining tools on a subset of those transcripts, which our user interface allows navigating.

2.2 Corpus Structure

Here we describe the structure of the complete corpus, from which we selected a sample for our analyses. The manuscripts are organized in boxes, containing several folios each. Folios are divided into one or more pages. Each page is encoded as a TEI-compliant XML file.[2] As of January 2017, when the results of our experiments were written up, 17,513 folios had been transcribed (45.17% of the total amount of then digitized material).[3] Each folio is identified by a combination of a box-ID and a folio-ID, and based on those IDs, unique identifiers for each page can be created. The document types are heterogeneous. In an effort of several years, Bentham

[1] https://www.ucl.ac.uk/bentham-project

[2] For TEI, see http://www.tei-c.org/index.xml

[3] The Transcription Desk shows current progress: http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham

Project scholars went over each folio to determine document types and to add to the corpus other metadata (details below). The most frequent document categories are the following:[4]

- **Text Sheets**: Draft material for works in progress. Their interest lies in the fact that Bentham usually destroyed drafts after publication, so that extant text sheets contain unpublished works, or material excluded in the published version of a work.
- **Marginal summary sheets**: They summarize the content of text sheets and are useful to restore their order when unclear.
- **Fair copies**: Final version of a work that would be handed to a publisher, after a cycle of corrections.
- **Collectanea**: Material copied by Bentham's amanuenses from newspapers or other sources, so that he could cite it.
- **Correspondence**: Either received by Bentham or drafts of letters he sent.

A distribution of documents across these main categories in the subset of the corpus selected for our analyses is in Figure 2. Besides the document type, Bentham Project scholars recorded several metadata for each folio,[5] when related information was available on the manuscripts, such as: Date or estimated date of composition, headings and subheadings (set apart from the body of the page), titles (in the body of the page), watermarks, penner (Bentham or one of his assistants), and, for the correspondence, sender and addressee.

2.3 TEI Encoding

The TEI markup encodes information about the writing process, like additions and deletions (crossed-out material). Document structure elements like headings, breaks and marginal notes are also encoded. Other markup identifies stretches of foreign language, and uncertain or illegible text. Features like superscribed or underlined text are also annotated. See [Causer et al., 2012, 123ff.] and the transcription guidelines for a detailed description.[6] We did not exploit this information in our analyses, but it would be useful to do so, for purposes like restricting corpus searches to deleted passages only. Figure 3 shows an example of a manuscript and the information annotated in its TEI transcription, once rendered as HTML.

Most corpus documents are in English, but some are in French, or contain long Latin passages. Language metadata was not annotated at folio or page level. The annotation scheme uses a foreign tag to identify foreign passages, but without specifying the language.[7]

2.4 Corpus sample in this study

2.4.1 Document Selection

As a collaboration between the LATTICE Lab and UCL's Centre for Digital Humanities, we had access to a large subset (29,928 XML files) of the then transcribed material, to perform automatic text analyses on it. Each XML file corresponds to a transcribed page. For our analyses, we did not use all the files made available to us, but about 55% of them, for reasons detailed in following.

[4] See http://www.benthampapers.ucl.ac.uk/help.aspx?subject=category for more details.

[5] See http://www.benthampapers.ucl.ac.uk/search.aspx?formtype=advanced

[6] http://transcribe-bentham.ucl.ac.uk/td/Help:Transcription_Guidelines#Core_Guidelines

[7] http://transcribe-bentham.ucl.ac.uk/td/Help:Transcription_Guidelines#Supplementary_Guidelines

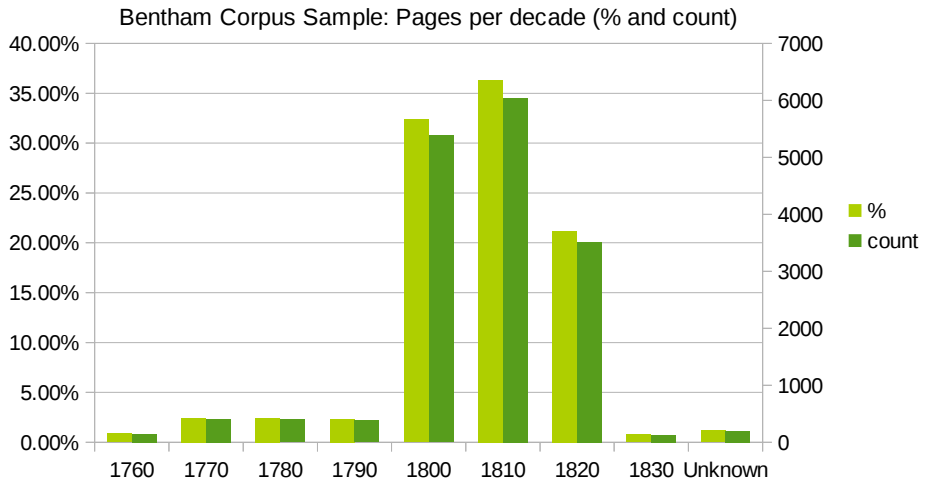


Figure 1: Percentage and count of pages per decade in our sample of 16,618 pages from the Transcribe Bentham corpus. Decades correspond to dates assigned to the documents by the Bentham Project.

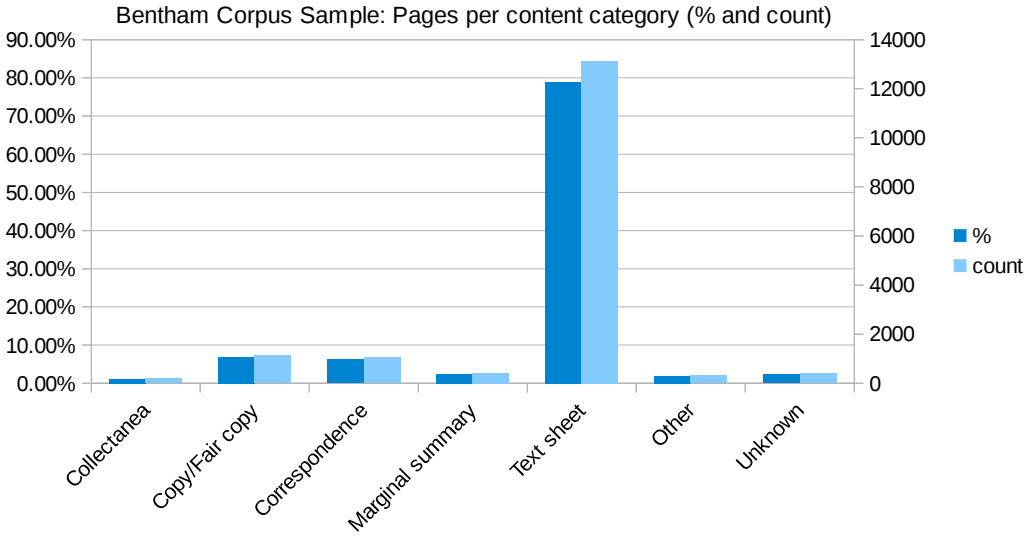


Figure 2: Distribution of pages (count and percentage) in our 16,618 page sample across the main content categories, besides infrequent categories grouped as *Other*.

At first, we did not have access to the corpus metadata, which include document dates.[8] Since we wanted to analyze the temporal evolution of the corpus, we needed to assign a date (a year) to each file. We used a simple heuristic to assign years: If the first sequence of four digits in the file was between Bentham's year of birth and death, this was considered as the text's year of production. The years obtained with this simple rule correlate very strongly with the actual years identified by Bentham Project scholars, which were made available to us more recently (Pearson's $r = 0.976$). However, the heuristic was not applicable to ca. 44% of the XML files we received, as they contained no sequence of four digits.

To identify non-English files, we ran a language identification tool (Lingua-Identify, a Perl module).[9] This classifies text against pre-trained models for many languages, using features

[8]The data originally available in the context of the collaboration between the LATTICE Lab and UCL included the TEI documents only, not the master metadata table. The metadata was made available towards the end of the work described here.

[9]Using the default options, as documented in `http://search.cpan.org/~ambs/Lingua-Identify-0.56/lib/Lingua/Identify.pm#langof`

JB/027/047/001

Click Here To Edit

C
Preface

Though I have to speak of Rebellion and Hate-Libels,
I hope I shall be forgiven, though I ~~have not~~ should
be found not to have declaimed against a mob-monarchy; nor called
Kings Tyrants because they might could be so. One
thing only I will avow maintain without reserve that where the many
govern it can only might only be for the (sake of) the more;
and that where [it is] but one [that] governs the case
is still the same. the power is in the many, the benefit belongs only to the more I
fear not much the being disavowed
by anyone in the circle of sovereigns ~~by whom Europe~~
~~is now~~ who now [wield] fill any of the thrones the monarchies of Europe.
There is one at least whom I am sure of: she her who has spoken
out and said Path[?]. 2. Instruct. art. 500. " ... This perhaps will not be much to the
"taste of those flatterers who are incessantly whispering
"into the ears of sovereigns ~~that [. . .] [. . .]~~ "the nations
"are ~~our~~ your property and created ~~only~~ for ~~those [. . .]~~ >your use. For
"our parts we make it a ~~point~~ matterof duty to remember
and glory to avow, that [it [is] we who
are made for [our people people's] Russia's sake, and not they
. . . Prussia for ours'.

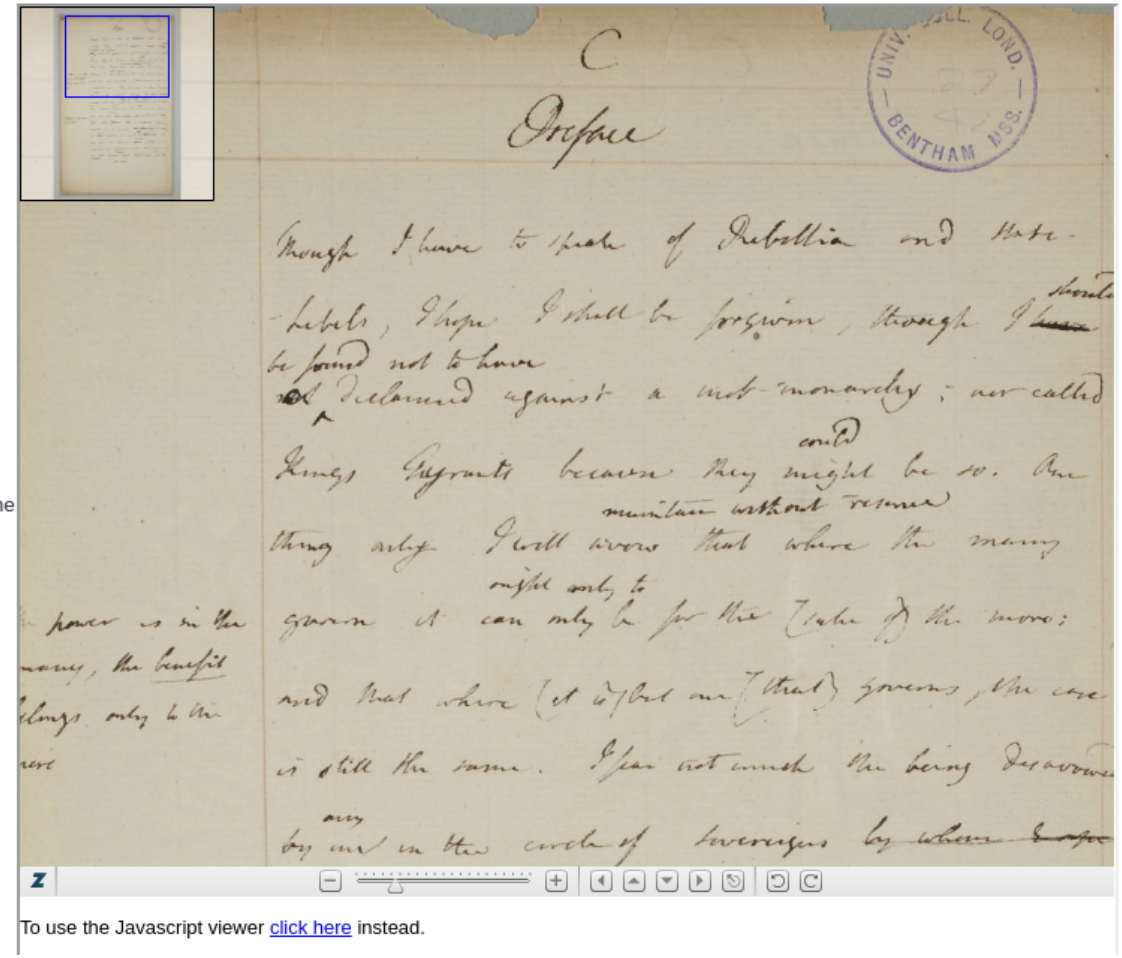


Figure 3: UCL Transcribe Bentham interface. A manuscript digitized by UCL is on the right, which shows some of the characteristics encoded in TEI by volunteer transcribers, like added and deleted text, and marginal notes. The left pane shows an HTML rendering of the TEI, reflecting those annotations: Added superscribed text, crossed-out deleted text, and boxes for marginal notes. The screenshot was taken from http://transcribe-bentham.ucl.ac.uk/td/JB/027/047/001

like words, prefixes and suffixes common in each language, and character n-grams. Approx. 400 files were identified as not in English.

After eliminating the files which our dating heuristic could not find a year for, as well as non-English files, the sample size kept for our analyses was 16,618 pages (i.e. 55.53% of the documents originally sent to us). A side effect of our document-dating heuristic is that our sample mostly contains documents from after 1800, when Bentham started regularly dating the manuscripts.[10] In consequence, our content evolution analyses yield clearer results from 1800 onwards (see Figure 6). The distribution of pages per decade in our sample, using dates provided by the Bentham Project (not those output by our dating heuristic) is in Fig. 1. The sample's distribution of pages across main document types is in Fig. 2.

2.4.2 Corpus Preprocessing

Regarding text preprocessing performed before feeding text to the NLP tools, recall that the transcripts are encoded in TEI, providing information like added or deleted text, marginal notes etc. The NLP tools we applied take unannotated text as input, rather than TEI-encoded text. The TEI-information encoded was not essential for our text mining work, even if exploiting this information could be useful, e.g. for a special treatment of deleted items, added items, or of the small proportion of material not authored by Bentham, like correspondence he received, or material that he collected from other sources. A TEI document's two outermost elements are a `teiHeader` tag, for metadata, and a text tag, for the document content. Our preprocessing ignored the `teiHeader` tag, as it only encodes information relevant internally at Transcribe Bentham–the metadata mentioned above like dates or document types are available in a database and had not been rendered in our source TEI. To obtain unannotated text, text inside a deletion tag was removed. The textual content of all other tags in the document body was kept for text mining. As such, the text used for our analyses corresponds to the content of each page, barring deletions, and including all additions. This is a safe choice, in the sense that all non-deleted text is kept, but it is not a detailed way to represent the corpus. In future work, preprocessing could be improved. For instance, deleted text could be retained but indexed in a separate field, to be able to perform searches (or co-occurrence analyses) involving deleted text.

III PRIOR WORK ON THE CORPUS

Before giving some examples of scholarly work based on the body of transcripts produced by Transcribe Bentham, a first thing to note is that years of effort have been invested in creating the corpus by Transcribe Bentham, its "crowdsourced" volunteers, and the Bentham Project itself, whose scholars catalogued the corpus and were transcribing the manuscripts prior to the crowdsourcing initiative. Several works analyze this corpus creation process [Causer et al. 2012, Causer and Terras 2014a,b]. Issues discussed include the transcription platform and the TEI encoding choices and crowdsourcing. As regards crowdsourcing, topics addressed are methodology, productivity evaluation and a discussion of larger implications of the initiative, like public participation in cultural heritage, and engaging a non-specialized audience into creating valuable resources for scholarly work. The potential of integrating handwritten text recognition technology in the transcription process is also analyzed [Causer and Terras 2014b; Toselli and Vidal 2015].

Besides these accounts of how the corpus was created, including cataloguing (metadata), digitization (photographs of the manuscripts), and transcription, there are two platforms that offer

[10] See `http://www.benthampapers.ucl.ac.uk/help.aspx?subject=estimated_date`

some corpus navigation functions for those outputs. The Bentham Papers Database allows searching in metadata fields, returning matching records, and providing a link to the record's transcript when available.[11] UCL Libraries' Digital Collections created a platform where, besides metadata, the full text of available transcripts can be searched. It returns links to the document's image and its TEI transcript if available.[12]

Regarding Bentham studies work based on the transcriptions themselves, a major output the transcripts are contributing to is the Bentham Project's edition of Bentham's Collected Works [Bentham, 1968 – ongoing]. These are being produced by a team led by Prof. Schofield. Input from the volunteer-transcribed manuscripts is being considered as a source for editorial comment within the new Collected Works, and material from the manuscripts is being published as part of them. Causer and Terras [2014a] mention several examples of previously unknown content in fundamental areas of Bentham's work, like legal reform and political economy, including new information about his strong opposition to convict transportation. Also, on his support of a fair treatment of animals. This new material is relevant for debates in Bentham scholarship, such as the origin of his notion of *sinister interest* (i.e. when private interest, rather than common good, is leading rulers' actions) and the timeline of his conversion to political radicalism.[13] The Bentham Project and Transcribe Bentham are authors of a large number of significant outputs, the work just mentioned refers to the most important contributions to Bentham scholarship based on the transcriptions of the manuscripts. A systematic account of those projects' achievements is maintained at their respective websites.[14]

IV OUR APPROACH: CORPUS CARTOGRAPHY

The analyses just mentioned involve a detailed reading of the transcripts in order to find new evidence that can contribute to Bentham studies. By contrast, the technologies we have applied seek to provide new evidence from connecting aggregated data. More precisely, from an overview of the corpus in the shape of a network or map, that can potentially provide new insight. We are not aware of previous automatic text analyses of the Bentham corpus, which increases the interest of the experience reported here.

To create corpus maps, Natural Language Processing and graph visualization tools are applied, performing three steps: First, an extraction of expressions to model the corpus with. Second, a clustering of those lexical sequences based on words shared across their contexts of occurrence, as an indication of semantic relatedness between the expressions. Finally, since clustering computes semantic distances between terms, the corpus can be visualized as a network of related expressions, thanks to spatialization algorithms that take those distances into account. The network thus created serves as a map of the corpus. Each of these steps is discussed below.

4.1 Lexical Extraction

For lexical extraction, we used two technologies, Entity Linking and Keyphrase Extraction, with a view to comparing the results of each. In following, our use of both technologies is discussed: tools, settings, and the process employed to select a set of expressions to analyze the corpus, based on the output of each technology.

[11] http://www.benthampapers.ucl.ac.uk/

[12] https://www.ucl.ac.uk/library/digital-collections/collections/bentham

[13] *Radicalism* in this context involved positions like defending universal suffrage and a representative parliament, as Causer and Terras [2014a] point out.

[14] Bentham Project: https://www.ucl.ac.uk/bentham-project/; Transcribe Bentham: http://blogs.ucl.ac.uk/transcribe-bentham/

4.1.1 Entity Linking

Entity Linking (EL) looks in a corpus for mentions to terms from a knowledge base (KB), i.e. a repository like Wikipedia or its semantic web version, DBpedia. The mention is then annotated with the relevant KB term. This is used to relate passages referring to the same KB term to each other, abstracting away from variability in the ways of referring to that term in the corpus. For instance, textual mentions *amount* and *quantity* will be mapped to the *Quantity* concept in DBpedia.[15]

In our Entity Linking workflow, we are targeting knowledge-base terms that correspond to conceptual mentions as well as terms expressed by named entities. Conceptual mentions are usually noun phrases. Named entities are lexical sequences corresponding to a set of predefined types (like people, places and organizations) and are often proper nouns.[16]

For Entity Linking, we used the DBpedia Spotlight tool [Mendes et al. 2011, Daiber et al. 2013]. This tool employs DBpedia Auer et al. [2007] as its knowledge base. DBpedia's content is extracted from Wikipedia. A question to ask is whether Wikipedia, as a general-domain encyclopedia created in the 21st century, is a relevant source of knowledge to analyze specialized texts from the 18th and 19th centuries. Using DBpedia as the KB gave good results in some cases, but posed some problems too. The results and a way to work around these difficulties are discussed below; the limitations of using DBpedia as the KB was a reason to use Keyphrase Extraction as a second source for identifying important expressions in the corpus.

Spotlight's algorithm proceeds in essence as follows. It first identifies concept-mentions, i.e. corpus sequences potentially referring to DBpedia terms, as well as the set of DBpedia candidate terms for each sequence. This "mention-spotting" relies on a pre-defined dictionary which maps expressions to DBpedia pages, based on page titles, Wikipedia link anchor texts, etc. Then, it compares the context of an expression in the corpus with the context vectors for each candidate term. A context vector is the concatenation of all paragraphs mentioning the term in Wikipedia. The similarity between the context of an expression in the corpus and each DBpedia candidate term's context vector is computed, whereby context tokens are weighted according to their discriminative power to tease candidates apart. The term whose similarity with the mention's context is largest is selected, if the score is above a configurable threshold. The similarity score (among other factors) is used to output a confidence score for the annotation, which gives an indication of the extent to which it is likely correct.

We called Spotlight's web service, using default settings, to analyze our corpus sample, after the preprocessing steps described above. From the results returned, only annotations whose confidence was above 0.1 were kept. Besides, we only kept an annotation if at least one of its textual mentions (i.e. the span of text the annotation covers) occurs at least 100 times in the corpus. Mentions occurring less than 100 times were also removed from each annotation's mention-set. The appropriateness of these thresholds was determined empirically.

These thresholds yielded a list of 285 terms. Each term could have one or more textual variants. For instance, the term *Judiciary* had been assigned by Spotlight to mentions *judicatory*, *judicial*, and *judicature*, but the term *Doctrine* had been used to tag occurrences of one textual variant only (*Doctrine*).

[15] Information about a DBpedia concept can be accessed by prepending `http://dbpedia.org/page/` to the concept label, e.g. `http://dbpedia.org/page/Quantity` for concept *Quantity*

[16] When Entity Linking targets conceptual information, some scholars speak of Wikification instead of EL. In this paper, both terms are used interchangeably.

The first step after obtaining this initial list was manually verifying the terms, both the textual mentions and the DBpedia terms they had been annotated with by Entity Linking. This revealed several errors. Some errors were anachronisms, when a mention had been annotated with a DBpedia concept for senses which started existing after Bentham's life, such as the mention *quantum*, annotated as the physics concept *Quantum*, or the mention *application*, which in about 25% of its ca. 1,000 occurrences was annotated as *Application_software*. Anachronisms are easy to spot and remove from the term list before creating corpus maps. Some other errors are harder to find, since determining the correctness of the annotation requires looking at corpus examples. For instance, the mention *execution* is used in the corpus in a sense of *application of a judicial decision*. However, it had been annotated by Entity Linking as DBpedia term *Capital_Punishment*. If we accept this automatic tagging, we would be misrepresenting the corpus content, as all the contexts where the word *execution* appears would be considered as contexts where the death penalty is discussed, and co-occurrents of this word would be considered as terms mentioned in discussions around the death penalty. This would be false.

To avoid such errors, instead of labeling nodes in the corpus map with the DBpedia concept for the set of textual variants whose occurrences are aggregated in the node, the nodes were labeled with the most frequent variant in the set. When a textual variant had been disambiguated as more than one DBpedia concept, the variant set for both concepts was generally the same. In case of a discrepancy, the set containing the most frequent variant was kept.

The implication of the labeling procedure we chose is the following: An Entity Linking tool was used, with the original intention to use DBpedia concepts to model the corpus. However, in view of incorrect disambiguations, yielding anachronisms or other errors, the mention-spotting step was the main source of information to annotate the corpus, rather than the full results of linking to DBpedia. The concept structure of DBpedia was still used, in the sense that mentions that had been disambiguated as the same concept were kept as variants of each other (choosing the most frequent variant in the set as the label to represent them all). But this does not exploit the disambiguation fully, in the sense that, by not using the DBpedia label, the annotation does not claim that the related DBpedia concept is mentioned in the corpus.

Besides the label modification, Spotlight's original results were manually filtered to remove weaker results. Recall from above that annotations whose confidence was below 0.1, and variants whose corpus frequency was below 100 had been removed automatically. A list of 285 ⟨*variant set, label*⟩ pairs was thus obtained. Among those, some items express a general meaning that is unlikely connected with core notions in the corpus, e.g. variants or labels like *time* or *place*. For this reason, about 25 pairs were filtered out, yielding a final list of 258 pairs, which were then used to create concept networks, and corpus maps based on them. The final list is available on our user interface.[17]

4.1.2 Keyphrase Extraction

Keyphrase extraction [Turney 2000, Kim et al. 2010] identifies sequences of words representing the most important concepts in a text. The technology has been used for purposes like bibliographic indexing, or improving retrieval in search engines via keyphrase-indexing. In Digital Humanities applications, it is sometimes used to give an overview of a corpus (e.g. Moretti et al. 2015, Rayson 2008) which is the use intended here.

[17] http://apps.lattice.cnrs.fr/bentham/lexical-extraction.html

Keyphrase Extraction was performed with Yatea [Aubin and Hamon, 2006], a rule-based keyphrase extractor.[18] It takes as its input part-of-speech tagged text in Treetagger output format.[19] Part-of-speech tagging (PoS-tagging) was done with Treetagger [Schmid, 1994]. Based on the PoS tags, Yatea first chunks text in order to identify noun phrases, according to configurable PoS patterns. The tool then filters the resulting noun phrases, in order to eliminate candidates, which, although matching one of the expected patterns, contain uninformative sequences. For instance, terms containing the preposition + noun sequence *of course* would be filtered out. We configured the tool to output both phrases with several words and single-word phrases.

Keyphrases with at least 10 occurrences in the corpus were initially kept, giving a list of ca. 2550 terms. This list was filtered further with regular expressions to eliminate ill-formed terms. An example of such terms are terms containing punctuation, given tokenization errors coming from irregular corpus formatting. Also, uninformative terms not previously filtered, like phrases containing the demonstrative *such* or the determiner *certain*. After applying regular expressions, the list was finally filtered manually to eliminate remaining irrelevant terms. This yielded a final list of approx. 1,950 terms. From these, the most frequent 250 terms were used to create corpus maps. The list of terms is shown on our user interface.[20]

The minimum frequency selected for keyphrases (10) is smaller than the one for Entity Linking mentions (set at 100). Keyphrases are generally multi-token and will consequently reach a smaller frequency than single-word items. A frequency of 10 for multi-token expressions was considered sufficiently representative for the corpus. As regards single-word keyphrases, the minimum frequency allowed for them was also 10, but their relevance was verified manually.

Since we were planning on a manual verification of keyphrases and the number of keyphrases to verify was sufficiently small, the choice of the keyphrase extraction tool was not crucial. We chose Yatea since we had worked successfully with it in earlier projects, for French and English texts [Mélanie et al. 2015, Ruiz Fabo et al. 2016]. A newer tool that may require less manual cleanup of results is Keyphrase Digger (KD), by Moretti et al. [2015].[21]

4.2 Lexical Clustering

The lists of terms described above, including their variants in the case of terms derived from Entity Linking, were clustered with the CorText Manager platform.[22] CorText Manager is a browser-based tool, able to perform all three steps in corpus cartography: lexical extraction, clustering and visualization. The tool can also be used to create the network only, as we did. In this case, lexical extraction and visualization are performed with other tools: CorText Manager accepts standard import formats for term lists (e.g. CSV), and it exports the network visualization as a GEXF file,[23] which can then be visualized with network analysis tools like Gephi.[24]

[18] http://search.cpan.org/~thhamon/Lingua-YaTeA/lib/Lingua/YaTeA.pm

[19] See http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger

[20] http://apps.lattice.cnrs.fr/bentham/lexical-extraction-keyphrases.html

[21] https://dh.fbk.eu/technologies/kd. This opinion is based on an informal inspection of the tool's results on the Bentham corpus. The tool provides a weight for each term extracted, indicative of the term's importance in its document and in the corpus. Like in Yatea, the ratio of single-word to multitoken terms extracted is configurable. KD has been tested on the SemEval 2010 keyphrase extraction dataset [Kim et al., 2010], ranking 2th to 4th depending on the evaluation mode.

[22] https://docs.cortext.net/

[23] GEXF stands for Graph Exchange XML Format. This format was created by the Gephi project (https://gephi.org/gexf/format/)

[24] Gephi is a social network analysis tool, available at https://gephi.org/. The tool reads and exports networks in several popular formats, and has functions for editing the networks.

Prior to importing term lists, the corpus needs to be indexed in the platform, so that the terms can be searched in the corpus, and their context vectors computed and compared for clustering. The corpus is importable in several standard formats, we chose a CSV format with fields for each documents title, text, date of composition and decade.

Clustering starts with selecting the **number of nodes** to create the networks with. We chose to create networks of approx. 150 and 250 nodes. A maximum of 250 nodes was chosen since we thought that the network would be easily readable at this (limited) level of detail. The 150-node network was created to see the differences in informativeness between it and the 250-node network. Since the network creation algorithm eliminates some of the weakly connected nodes (see the discussion of network filtering below), the actual number of nodes in the networks was 141 and 233 for the EL-based ones, and 133 and 240 for the keyphrase-based ones.

When assessing the number of nodes to include in the networks, we also created a network with 1,000 keyphrases.[25] The corpus overview in this larger network is comparable to the smaller networks we created, since similar topics are covered, but in greater detail. In this sense, we consider that the smaller networks are an appropriate representation of the corpus, in that they do not seem to leave out essential parts of its content, and are more easily navigable than larger networks. Domain-expert feedback (below) did not suggest that the networks with approx. 250 or 150 nodes lacked coverage of corpus areas either.

Terms are clustered based on their distributional similarity, i.e. based on the overlap between tokens in the terms' contexts in the corpus. The context length is configurable: a range of sentences can be chosen, or the whole document. In our settings, the context is five sentences around the term. The score for similarity between two terms relies on pointwise mutual information, using a measure defined in [Rule et al., 2015, Supporting Information, p. 1]. This measure is inspired by Weeds and Weir [2005], and, for reasons discussed there [p. 443ff.], it is asymmetric.[26] This asymmetric measure results in a directed network, where edge weights correspond to the similarity score between the terms linked by an edge.

The **network is filtered** during its creation, to obtain relevant clusters, by removing unmeaningful edges that may obscure more important connections. The filtering steps are configurable from CorText Manager's UI. The first filtering step we applied consists in a similarity threshold; links whose weight is below it are deleted. The threshold can be fixed by the user, or an optimal threshold can be computed automatically, with the goal of obtaining a connected network, with no single disconnected nodes, and where a disconnected component contains maximally three nodes [Rule et al., 2015, Supporting Information, p. 2]. We chose for optimal thresholds to be computed automatically. The optimal threshold for the networks whose nodes were based on Entity Linking was 0.41. For the networks based on keyphrase extraction, the thresholds were 0.33 for the 231-node network and 0.28 for the 240-node network. A further filtering consists in restricting edges to those connecting each node to their top-N neighbours, as ranked by the similarity measure mentioned in the paragraphs above. The number of top neighbours was set to 10 for all networks.

[25] https://documents.cortext.net/lib/mapexplorer/explorerjs.html?file=https://assets.cortext.net/docs/8ce9f27f43b4d0952fca99e1f1eb73dc

[26] The notion of similarity is based on lexical substitutability, and one example illustrating asymmetry is that *dog* can generally be replaced by *animal* in a sentence, but not vice-versa. Other examples involve the different senses of homonyms.

Communities are computed on the network, i.e. groups of highly interconnected nodes. The algorithm is Louvain [Blondel et al., 2008]. In visualization (below), nodes are coloured according to their community. The **communities are labeled** using the names of their two most central nodes. The highest centrality is defined here as receiving the most inlinks from other nodes in the community. We speak of *in*-links, since, as mentioned above, the networks are directed. The labeling algorithm intends to select labels that capture the main themes represented by lexical items in the cluster. An example showing communities and a legend with their labels is in Figure 5.

4.3 Network Visualization

The CorText Manager platform uses a force-directed layout to spatialize the networks. This type of layout simulates a physical system where repulsive forces push the nodes apart (like charged particles), whereas attractive forces exerted by the edges pull the nodes together (like a spring), until the forces are stabilized [Rule et al., 2015]. In CorText, a notion of gravity pulling nodes towards the center of the graph also applies. Since the network edges encode semantic similarity, nodes closer in the network are thematically related, sharing common contexts. Nodes receiving links from nodes in two different clusters share contexts with nodes from both of those clusters, and represent concepts related to the themes of both clusters.

The spatialized network is encoded as a GEXF file.[23] Besides the positions of nodes and edges, other network attributes encoded in the GEXF, that are exploited for visualization, are the following:

- **Node Weight**: this is represented by node size in the networks. This weight is based on the sum of the node's co-occurrents, using CorText's default setting.[27]
- **Community**: this is rendered as different colours in the visualization.

Other measures to characterize nodes' importance in the network are also encoded, among others a node's degree, in-degree and out-degree, i.e. its total number of connections, incoming connections and outgoing connections. However, we did not exploit these measures in the visualizations.

Networks and other visualizations that allow examining the temporal evolution of the corpus can also be created with the CorText platform. Whereas the platform has more complex functions to analyze the evolution of lexical cluster structure (see [Rule et al., 2015]),[28] the simplest way to get information about how the corpus content changes across time is by using what is known as the *Heatmap* function. This will highlight which areas of the network are salient in each of a series of pre-defined corpus periods. We had divided the corpus into decades, adding a decade field to the documents before indexing them. To establish salient areas per period, the tool gives a choice of statistics to find nodes whose occurrences are overrepresented in a period, i.e. having a statistically unlikely high frequency. We chose the tool's default option (the χ^2 statistic) to compute overrepresentation. Heatmaps are discussed further below.

Besides providing a GEXF file, the platform also renders the network as a PDF file. These outputs can be displayed on the platform, but we preferred to create a UI to give access to the Bentham corpus, complementing these networks with other navigation tools like a search index. This UI is discussed in following.

[27] https://docs.cortext.net/analysis-mapping-heterogeneous-networks/mapping-node-selection/

[28] Or the documentation at https://docs.cortext.net/analysis-mapping-heterogeneous-networks/mapping-dynamical-analysis-options/

V USER INTERFACE: CORPUS NAVIGATION VIA CONCEPT NETWORKS

The user interface[29] (UI) gives access to our sample of the Transcribe Bentham corpus via a full-text search index and thanks to a navigable rendering of the concept networks described above. The search index can be used to access contexts for the network nodes. Besides, a type of map called *heatmap* depicts the temporal evolution of the corpus content. The goal of the UI is to provide an overview of the corpus. As a first requirement, the networks should reflect a domain expert's knowledge of the corpus. Optimally, the networks and the connections between concepts in them might suggest new research ideas to a scholar (see the UI evaluation section for discussion).

Our UI complements prior platforms to navigate the corpus: The Bentham Papers Database[11] and UCL Library's platform.[12] The Bentham Papers Database offers a detailed metadata-based search. Both UCL Libraries' tool and our UI search for query terms in the complete text of transcripts. Whereas their application returns the image and transcribed text for manuscripts matching a query, our UI returns the text of each matching manuscript, with the query terms highlighted, and with date facets. The other way in which our UI complements the prior ones is by allowing us to navigate the corpus using concept networks; this possibility was not available in the tools just cited.

5.1 User Interface Structure

The default view of the UI is the search index, displayed on Figure 4. The `Search` menu points to the search interface. The `Corpus Maps` dropdown gives access to the navigable concept networks and heatmaps. The `Lexical Extraction` menu provides information to the user on how the term-lists to model the corpus with were created. Information for users on the types of maps created and how to use them can be reached at the `Introduction` page under the `Corpus Maps` menu. The following paragraphs describe the search interface and each type of corpus map.

5.2 Search Interface

The search backend is Solr (Lucene-based),[30] which is a widely used and easily configurable search server, with HTTP requests for indexing and retrieval. As Lucene, it features field-specific queries, e.g. searching in titles or document body only. Logical operators, proximity search and fuzzy matching are also possible.[31] For relevance scoring, the main factors used by Solr are tf-idf weighting with raw term-frequency counts,[32] the number of query terms found in a record, and the length of the matching field (matches in a short field are scored higher than in a larger one).[33] The tool returns a set of documents matching the query, ranked by relevance, with the query matches highlighted. It also performs faceting on the results returned, i.e. aggregation over a field of each document in the result-set. We faceted results over dates (year of manuscript composition); results can be filtered by 5-year periods in our UI.

[29] http://apps.lattice.cnrs.fr/bentham

[30] https://lucene.apache.org/solr/

[31] https://lucene.apache.org/core/4_0_0/queryparser/org/apache/lucene/queryparser/classic/package-summary.html

[32] The tf-idf (*term frequency – inverse document frequency*) for a term t in a document d is defined as $tf \cdot idf$, where $tf = frequency\ of\ t\ in\ d$ and $idf = 1 + log\frac{|documents\ in\ corpus|}{|documents\ containing\ the\ term|+1}$

[33] For a short description of relevance scoring in Solr/Lucene, see https://wiki.apache.org/solr/SolrRelevancyFAQ#How_are_documents_scored. For a more principled explanation, see https://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html

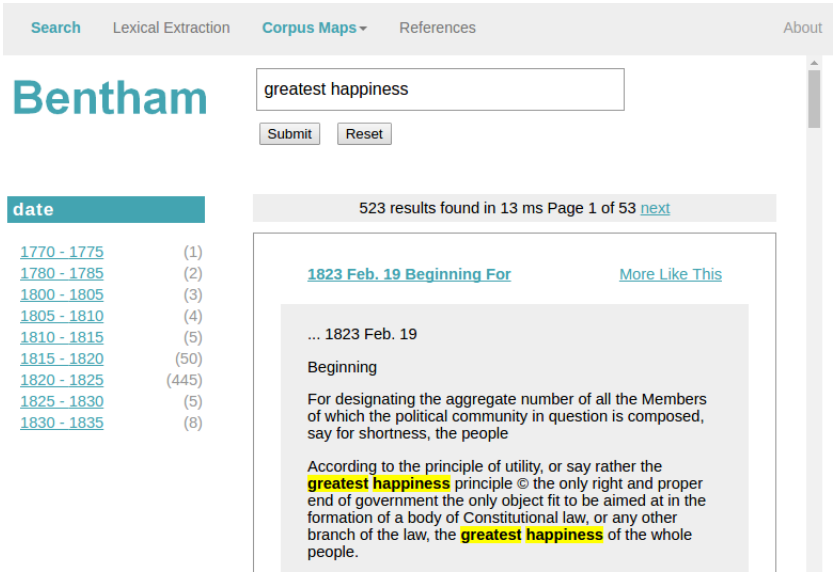


Figure 4: Structure of our User Interface to navigate transcripts of Bentham's manuscripts using concept networks and full text search. The screenshot displays the search index, showing results for query *greatest happiness*. Date facets are available on the left, to filter results per 5-year period (the number in parenthesis indicates records returned per period). Concept maps are accessible from the `Corpus Maps` menu.

5.3 Navigable Corpus Maps

Concept networks were created with the CorText platform, based on concept mentions obtained via Entity Linking, and based on keyphrase extraction. As will be described in the paragraphs below, the networks were then exported in GEXF format, and were rendered navigable with two libraries, which offer different navigation functions each. All these networks are accessible on the UI from the `Corpus Maps` menu, under the `Static` option.

Based on the concept-mention list on the one hand, and on the keyphrase list on the other, we had obtained networks of approx. 150 and 250 nodes. Two navigable versions of each network were created, using two tools. The first tool is the Gephi Sigma JS exporter plugin[34] and the second one is the TinawebJS project explorer.[35] Both tools rely on the Sigma JS graph drawing library.[36]

The navigable networks obtained with both tools allow us to search for a node in the network. Upon clicking on a node, both tools display a list of its neighbours on an interactive panel: By clicking on a node in the neighbour list, we can locate it in the network.

Besides the navigable maps just presented, **heatmaps** were created, that show salient areas in the corpus map per decade, using the method to calculate lexical saliency described on p. 13. In the UI screen captures in Figure 6, the areas shaded in red show how in the 1810s the manuscripts focused on human reasoning and religion, i.e. the communities labeled as

[34] https://github.com/oxfordinternetinstitute/gephi-plugins/tree/sigmaexporter-plugin/modules/sigmaExporter. The tool was created at the Oxford Internet Institute.

[35] https://github.com/moma/ProjectExplorer. The tool was created at two CNRS labs, the ISC-PIF and the CAMS.

[36] http://sigmajs.org/

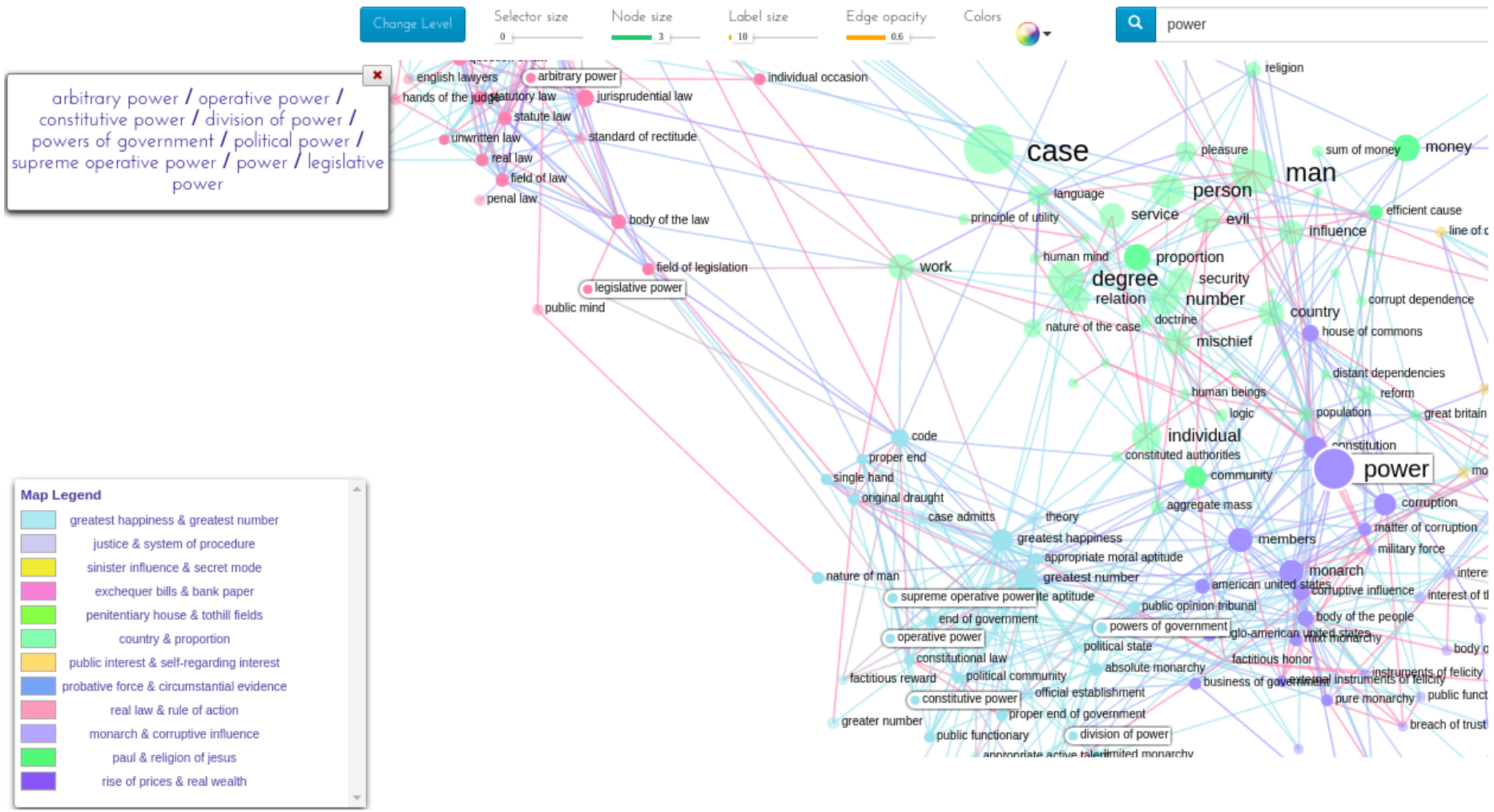


Figure 5: TinawebJS rendering of the 250 keyphrase network. Nodes matching *power* have been searched for, and the tool highlights them in the network. Terms belonging to different communities are highlighted, providing an overview of how power is discussed in the corpus. The communities are identified by colour. A legend by colour shows the two most representative nodes of each community, indicating each community's main themes. Node highlighting shows that the notion of power is treated in relation with the monarch (Community *monarch & corruptive influence*, antepenultimate in the legend). Also, in the context of what Bentham deemed the ruler's proper goal, i.e. maximizing the greatest happiness of the greatest number (turquoise community, first in the legend). Besides, power in a legislative context is also discussed (pale pink community containing nodes *arbitrary power* and *legislative power*). The TinawebJS library was created by the Complex Sciences Institute and the CAMS lab in Paris.

discourse & proposition and *God & Jesus*, whereas in the 1820s, the manuscripts turn their attention to the *Constitution & government* cluster.

In the case of heatmaps, on the UI we rendered them as image files based on the PDF files generated by CorText and navigability is restricted to choosing the decade. It would be useful to improve this in the future.

VI USER INTERFACE EVALUATION WITH EXPERTS

This section describes user validation work around the Bentham corpus interface. The evaluation task and its expected outcomes are described, and the results are discussed.

6.1 Introduction and basic evaluation data

Feedback was gathered from one Bentham scholar and one Digital Humanities (DH) expert. This can be seen as a basic preliminary validation with only two users. The evaluation sessions took around one hour each, and were carried out at University College London (UCL) in December 2016. Basic data about the users who contributed their feedback follows:

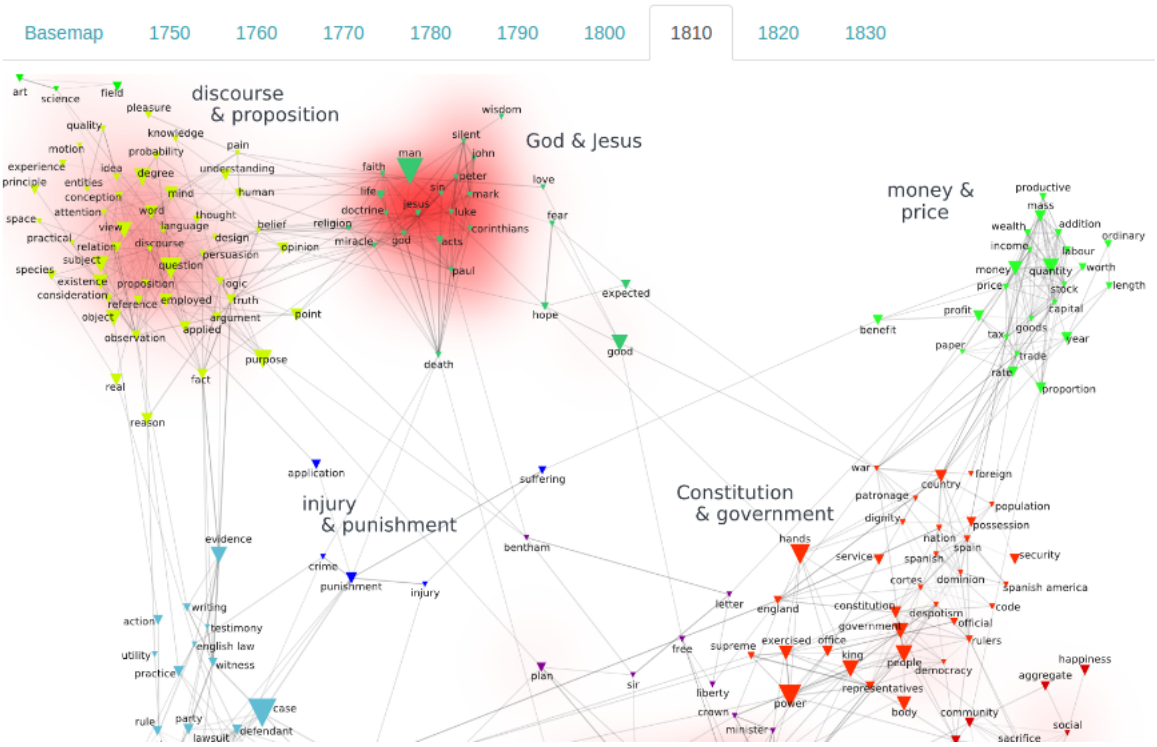
- **Domain expert**: Dr. Tim Causer, a historian and Bentham scholar working at UCL's Bentham Project and Transcribe Bentham, with deep knowledge of the corpus and its crowdsourcing transcription initiative, having published research on both, and who is working on Bentham's editions. Formerly he was the coordinator of Transcribe Bentham.
- **DH Researcher**: Professor Melissa Terras, who was at the time Director of University College London's Centre for Digital Humanities and Professor of Digital Humanities at UCL's Department of Information Studies. She has also published research within the Transcribe Bentham project.

6.2 Expected outcomes

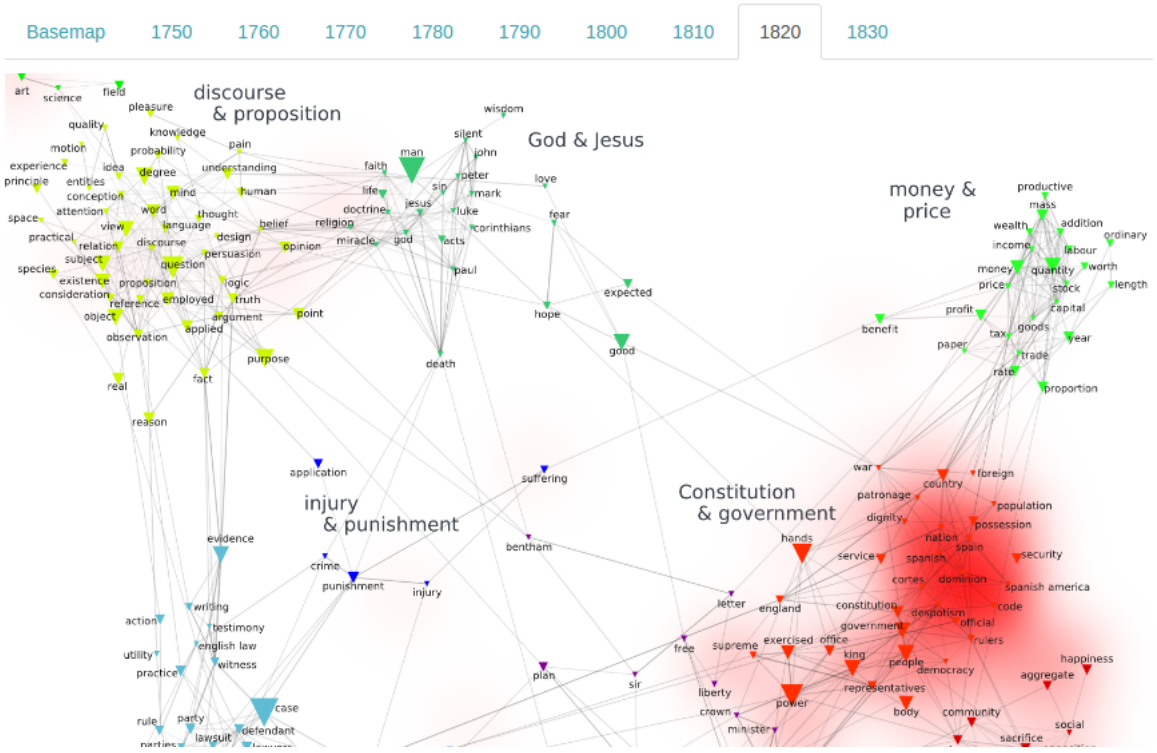
As the evaluation took place with only one domain-expert and one DH researcher, it was expected to provide preliminary feedback about the validity of the interface. No formal hypotheses were defined. The task, particularly the session with the Bentham domain-expert, was expected to provide information about the following issues:

- **Plausibility of the representations**: Are artifacts observed that would compromise the usefulness of the concept networks?
- **Usefulness of types of corpus terms extracted**: Two types of corpus terms had been extracted. First, mentions to DBpedia concepts, via Entity Linking/Wikification. Second, keyphrases salient in the corpus, regardless whether they are covered by DBpedia or not. We expected that DBpedia concept mentions would be perceived by the domain-expert as clearer for a non-expert public. Conversely, we expected the terms obtained via keyphrase extraction to be more informative for the domain-expert than the DBpedia mentions. The reason for these expectations was that the keyphrases, unlike the wikification mentions, are often technical terms referring to precise notions in Bentham's work.
- **Potential for new insight**: Whether using the networks may provide new ideas for research, e.g. about less commonly studied aspects of the corpus suggested by connections in the network.

The session with the DH researcher was mainly intended to provide general feedback about potential usefulness of the interface, ways to improve it, and the relevance of the approach chosen.



(a) 1810s



(b) 1820s

Figure 6: Heatmaps per decade, based on concept mentions obtained with Entity Linking: Red shading becomes darker as saliency of the corpus area in the decade increases. **Top:** In the **1810s**, two areas the corpus focuses on are the *discourse & proposition* and *God & Jesus* clusters. **Bottom:** In the **1820s**, the manuscripts strongly focus on notions around the *Constitution & government* cluster. A Bentham expert confirmed that the heatmaps correspond to the temporal evolution of Bentham's writings (p. 20): In the 1820s, Bentham wrote or commented constitutional code for several countries. See `http://apps.lattice.cnrs.fr/bentham/heatmaps-more.html` for high resolution maps.

6.3 Evaluation task

The feedback sessions involved the following steps. First, the methods for obtaining the visualizations were explained to the users, i.e. details about lexical extraction, term clustering, cluster labeling, visualization layout and heatmaps. Users were also given some examples how to use the networks to look for information. Then, the experts used the networks to look for information. The task was audio-recorded and a transcript summarizing the content of the audio was produced by the experimenter (the first author here).

The explanations given to the experts, before they used the networks to look for information, were the following.

- **Terms that connect two clusters**: This means that their contexts of occurrence overlap with the contexts of certain nodes in both of those clusters. The idea is to see if these connections are informative for a scholar. E.g. in the network with 150 terms obtained via wikification,[37] *degree* and *aptitude* connect the discourse-related purple cluster and the government-related green cluster (Figure 7). Is this relevant for a scholar?[38]
- **Verifying the corpus contexts where terms co-occur**: This can be useful for connections that seem interesting (or even suspicious). The corpus context can be verified with the `Search` menu.[39] E.g. in the 250 node wikification-based map,[40] *vote* and *bribery* are connected; we can verify the contexts connecting both terms with the search index as described (Figure 8).
- **Using the maps' search functions**: The *Navigable* maps can be searched. E.g. searching for *power* we see[41] that there's a term for *power* in a cluster related to the government, and another term *powers*, related to legislation. The *powers* node may then refer to *separation of powers* (Figure 9).

After these explanations, the experts were asked to look for information in the maps, or comment on how they would use the maps (if they would). The following maps were shown:

- Maps based on **Entity Linking**: both the 150-node and the 250-node versions
- Maps based on **Keyphrase extraction**: both the 150-node and the 250-node versions
- **Heatmaps**: For time reasons, only the 250-node ones, based on EL, were shown. (Note that the remaining heatmaps do not provide information conflicting with the heatmaps chosen).

The experts were asked explicitly about their perception of the differences between each version of the maps. The steps above were followed more closely with the Bentham expert, but more loosely with the DH researcher, whose feedback was less-corpus specific than the Bentham scholar's.

[37] http://apps.lattice.cnrs.fr/bentham/bentham-js.html

[38] An example where the connection seemed irrelevant to the experimenter was chosen, thinking that this may help not bias the expert towards thinking that these connections *will* be relevant. See p. 23 for discussions of possible biases in experts' feedback in a visualization evaluation task.

[39] http://apps.lattice.cnrs.fr/bentham

[40] http://apps.lattice.cnrs.fr/bentham/bentham-js-more.html

[41] The example comes from the 250-term wikification-based navigable map, http://apps.lattice.cnrs.fr/bentham/bentham-js-more.html

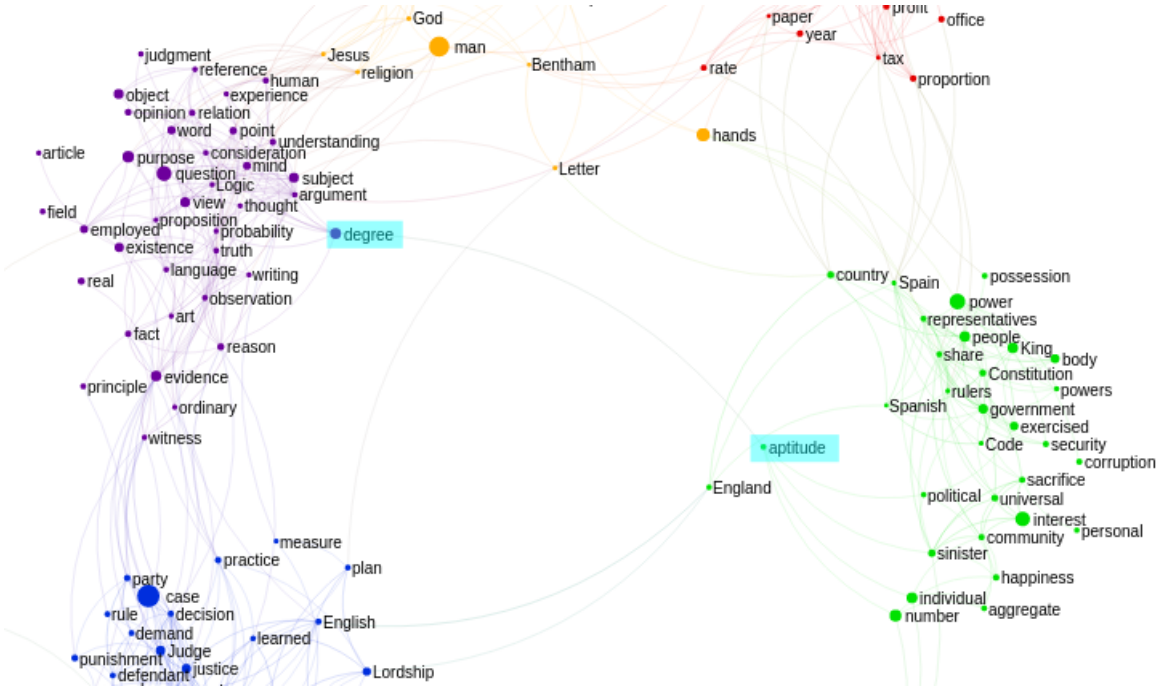


Figure 7: Nodes *degree* and *aptitude* connect two clusters in the 150-concept-mention map (Gephi Sigma JS export)

6.4 Results, discussion, and possible UI improvements

6.4.1 General comments by users

The DH researcher suggested to give more details to the users about the methods to create the corpus maps. This recommendation was followed by adding such information on th UI.[42]

Both the Bentham scholar and the DH researcher pointed out that being able to access the corpus contexts containing a network-node would be valuable. The current workaround is to search the node(s) in the search index,[43] but improving this would be useful future work.

6.4.2 Plausibility of the representations

The domain-expert expressed that the maps agree with his knowledge of the corpus, as suggested by some of his comments, documented below. Regarding the heatmaps per decade, he found that the corpus areas shown as salient in each decade correspond to Bentham's interest in that decade.

6.4.3 Applicability perceived by domain-expert

The domain expert found the corpus maps useful for the following application: When editing, they're interested in finding passages where Bentham discusses a given concept, even if he does not use the same words in each passage. For instance, around 1800, Bentham introduced the concept of *sinister interest*, which is at play when those in power act in their own interest, rather than for the benefit of society. However, Bentham may have referred to this concept earlier on, with phrases like *vested interest* or *sinister motivation*. The expert perceives that these networks help find terms that co-occur with *sinister interest* (or simply *interest*), and that, in

[42] E.g. in http://apps.lattice.cnrs.fr/bentham/maps-intro.html or http://apps.lattice.cnrs.fr/bentham/lexical-extraction.html

[43] http://apps.lattice.cnrs.fr/bentham/index.html

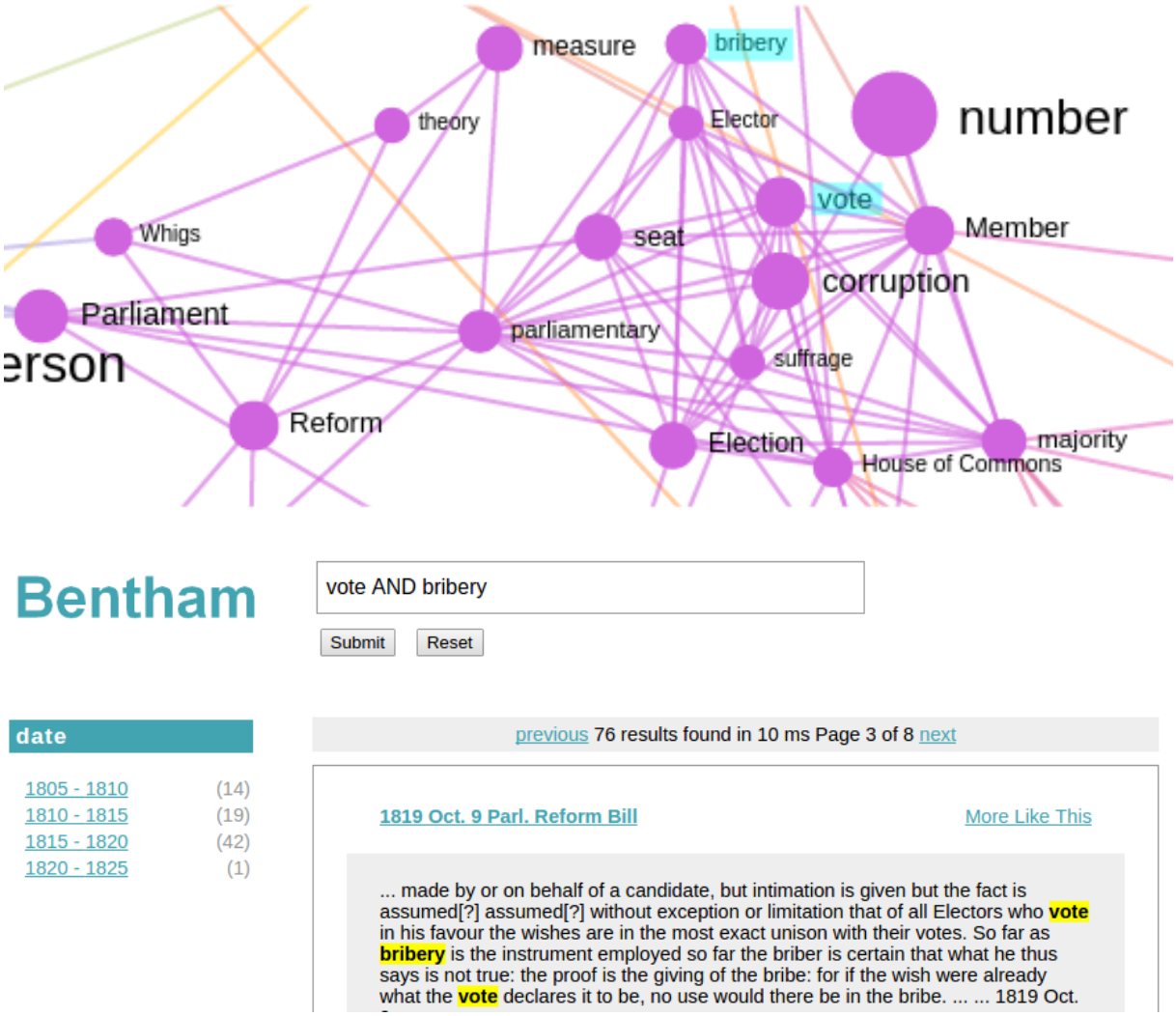


Figure 8: The concept-network (top) shows *vote* and *bribery* connected. The search index (bottom) allows us to find those connection's contexts

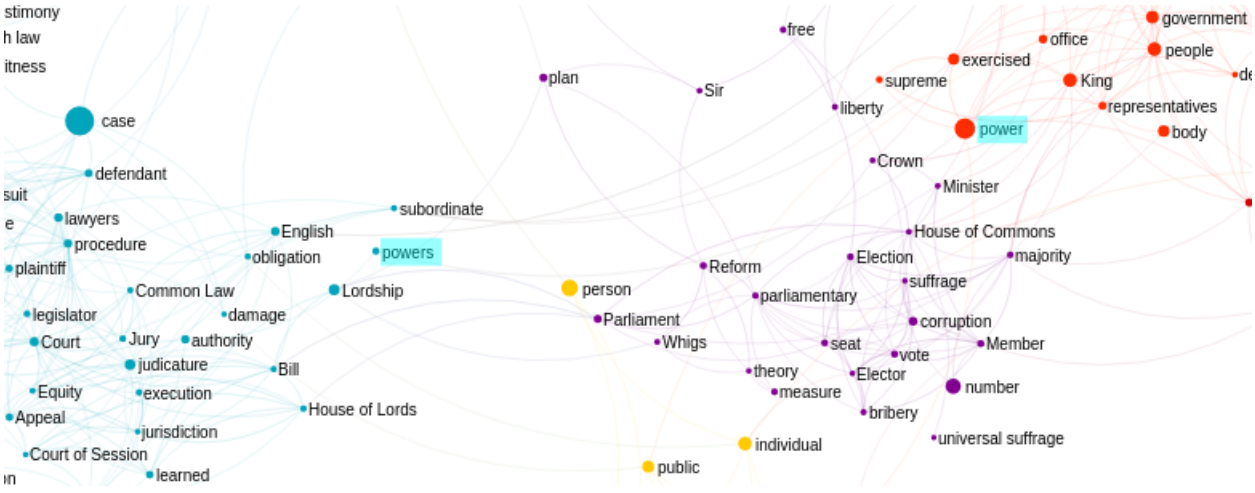


Figure 9: Nodes matching query *power* in the 250 concept-mention map (Gephi Sigma JS export)

turn, searching for these terms in the corpus may bring up contexts where the notion of sinister interest is discussed, albeit with different words. In fact, looking for *interest* in the navigable maps returned some terms that the expert found useful in the way just described (Figure 10), e.g. *private interest* and *self-regarding interest* (near-synonyms of *sinister interest*), or *general interest* and *interest of the people* (near-antonyms to that term).

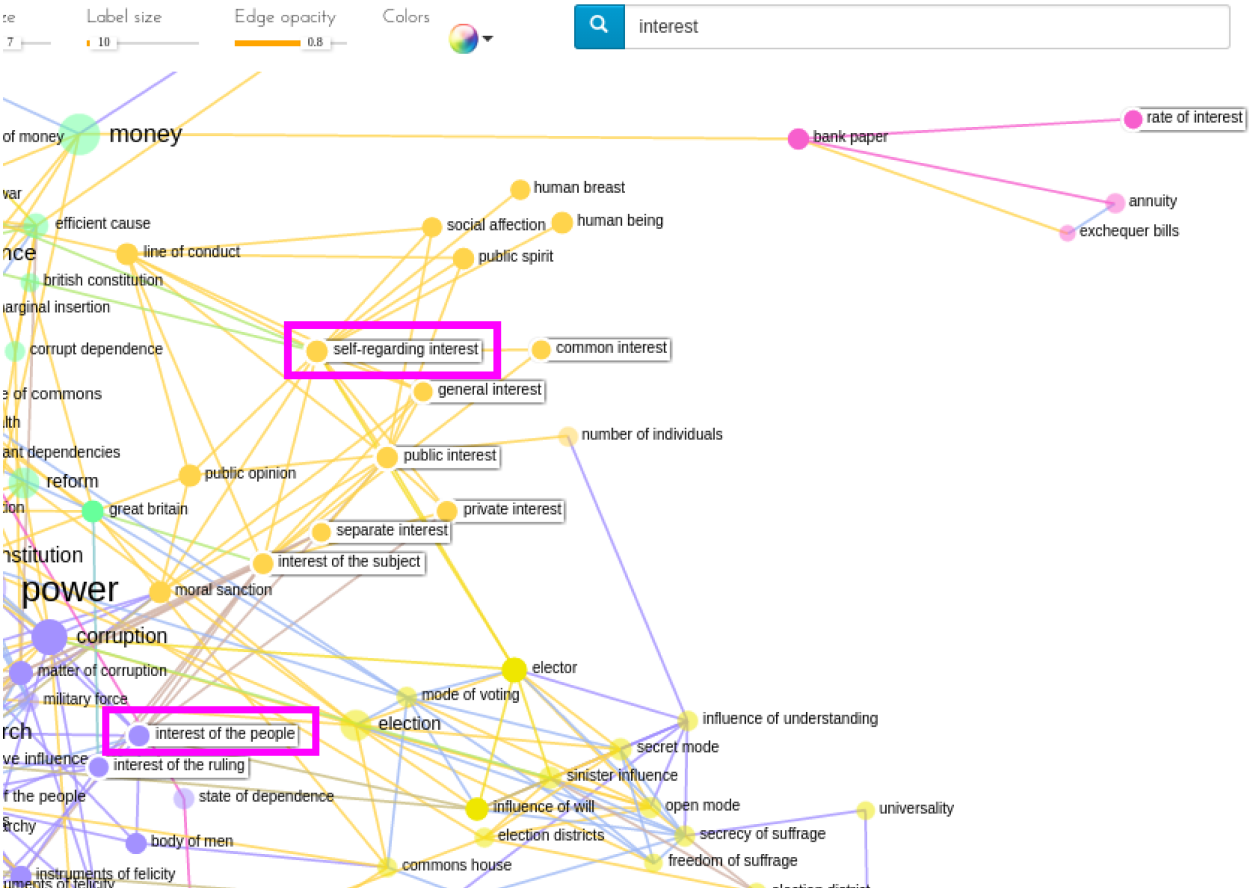


Figure 10: Results for query *interest* in the 250-keyphrase network (Tina export). Given a core corpus term like *sinister interest*, the domain-expert identified in these results near-synonyms (e.g. *self-regarding interest*) and near-antonyms (e.g. *interest of the people*) for that concept (pink squares were added on the image to highlight these two examples). This suggests the usefulness of the maps to find alternative formulations for a concept, and to examine terms in the context of those formulations

These user comments suggest a potential for gain of insight in the corpus representations created, thanks to their clustering of semantically similar terms together, based on the distribution of words in those terms' contexts. Note that, in order to find corpus contexts semantically related to a given term, other means would also be helpful, complementing the approach presented here: Tools from the "textometry" corpus-analysis school would help, like TXM [Heiden 2010, Heiden et al. 2010] or Le Trameur [Fleury and Zimina, 2014], which compute statistically salient terms in the context of a pivot-term.[44] Still for the same purpose, but using a very different paradigm to textometry, distributional similarity models could be created, and the user could query the model for the most-similar corpus-terms to their terms interest. E.g. the use of word2vec models [Mikolov et al., 2013] could be tested, or other models that have been shown to perform similarly [Levy et al., 2015].

[44] E.g. with their *Cooccurrence* modules: [this link] for Trameur, [this link, p. 47ff] for TXM.

6.4.4 *Number of nodes in the network*

The domain-expert found that the 150-node maps act as a "summary" of the content of the 250-node maps. He stated preferring the more detailed map, arguing that, for a historian, having as much data as possible would be desirable.

6.4.5 *Concept-mentions vs. Keyphrases*

When asked explicitly, the expert's comment about the different possible uses of the networks based on each type of terms was that both types of networks could be used in tandem. Also, that he would use the concept-mention based ones as a didactic device for Bentham non-experts. For instance, for an undergraduate assignment on punishment in Bentham, students could use the network to see terms related to this notion in the manuscripts before starting their work. However, for a Bentham scholar, he finds the networks based on keyphrase extraction more useful, since they contain more Bentham-specific technical terms, which can be particularly useful as mentioned above in order to find contexts containing alternative formulations for core Bentham notions.

Evidence on the usefulness of each type of network based on other user comments (rather than based on the answer to an explicit question about this), was the following: Looking at the area of the 250 concept-mention network[40] shown in Figure 11, the expert mentioned that he appreciated that the terms refer to "general concepts" in Bentham's thought (e.g that when he wrote about *happiness*, he had in mind notions like *interest*, the *government* and the *people*, all of which are located in the vicinity of *happiness* in the network, with no more than three nodes mediating between any of those terms). Another one of his comments was that the network provides an integrated view of what Bentham is thinking, and he interpreted that Bentham's democratic program is present in the network in the sense that nodes close in the network refer to both problems identified by Bentham (like *corruption* or *bribery*) and some of the remedies he proposed (like *community*, or the *Constitution*).

6.4.6 *Interpretability of the task's results*

The task should be considered as preliminary validation of the potential of the methods and products developed, rather than exhaustive evidence, for several reasons. First, the small number of users consulted. More users could be approached in future work. Second, as Khovanskaya et al. [2015] report, users tend to cooperate with what they perceive as the goal of the experiment, and tend to provide evidence agreeing with their perceived goal. For instance, if they interpret that the task intends to assess whether the tool is useful, they may choose to provide a positive message about the tool, avoiding critical feedback. I asked the expert to provide negative feedback if relevant, with a view to limiting this bias, but the effectiveness of asking for this explicitly is open to question. Finally, the visual nature of the product under evaluation (concept networks) poses an additional difficulty. As Rieder and Röhle [2012] suggest, following Heintz [2007, esp. p. 78], it is easy for someone looking at an image to overestimate the value and reliability of the visual evidence, and to assume that an external reality must be recoverable from the image via interpretation, rather than inquire about (and potentially question) the methods and possible biases involved in the production of the image. For this reason, it is important to make domain experts aware of the steps involved in producing concept networks and the biases they can introduce. We tried to do this in our evaluation by explaining such methods briefly to the experts. A possible source of information to assess whether the interpretation problems just described are taking place would be to show experts different versions of the networks and compare their comments on each. We did some work along these lines by showing

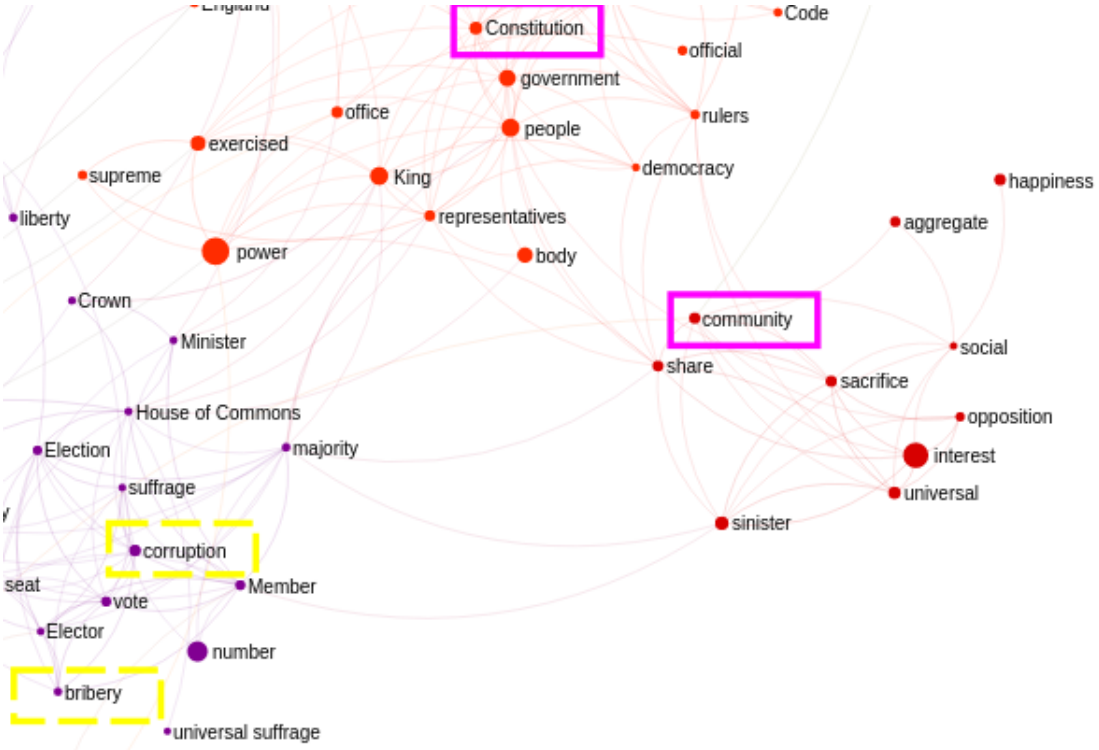


Figure 11: Area focused on by domain-expert as representing general Bentham concepts and the relation between them: He describes that the terms surrounded in full pink squares (e.g. *community*) are part of the remedy Bentham proposes for problems like those in the terms with dashed yellow squares (e.g. *corruption*). The map corresponds to the 250 concept-mention network, rendered with the Gephi Sigma JS exporter.

experts different maps, based on two types of lexical extraction. Finding more systematic ways to control for possible interpretation problems would be interesting future work.

VII CONCLUSIONS AND OUTLOOK

An application was presented to navigate the manuscripts of Jeremy Bentham, a 18th–19th century corpus in political philosophy, ethics and related topics. The manuscripts were provided by University College London, whose Bentham Project is transcribing the material. Entity Linking to DBpedia and keyphrase extraction were used to find important concepts in the corpus. Based on both concept sources, navigable corpus networks were created, besides offering full-text search on the corpus. Reference resolution against DBpedia presented some problems, as Entity Linking to DBpedia resulted in some anachronistic concepts, i.e. modern terms not applicable to Bentham's writings. For this reason, in order to create corpus networks based on Entity Linking, rather than using the DBpedia concept labels assigned to corpus mentions, labels chosen among the corpus mentions themselves were used. Keyphrase extraction, which took place in an entirely corpus-driven manner without relying on external knowledge resources, gave appropriate results.

Several results emerged from the domain-expert evaluation. The corpus overview provided by the networks corresponds to the expert's knowledge of the corpus in the sense that no obvious misrepresentations were observed. Networks based on keyphrases were more informative for the expert than the term mentions found by Entity Linking. Keyphrases can express precise notions in Bentham's thought. The DBpedia term-mentions were found to represent basic

elements of meaning underlying those characteristic Bentham notions, but not the precise expressions used by Bentham. The expert considered keyphrase networks useful for finding alternative phrasings for a term. Such alternative formulations are useful for editorial work on the corpus. They can be used to look for new evidence on how Bentham discusses certain notions, including passages where he used alternative expressions to refer to them. The expert saw the networks as a source of terms that can help find such passages.

Based on this expert feedback, the most relevant future work would be creating distributional semantics models for keyphrases in the corpus, and creating a corpus navigation application that allows a domain-expert to query the model with his or her terms of interest, in order to retrieve the most similar terms. Besides, the application should show the terms' context of occurrence, so that the expert can assess the results in context.

References

Sophie Aubin and Thierry Hamon. Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, pages 380–387. Springer, 2006.

Sren Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. *DBpedia: A nucleus for a web of open data*. Springer, 2007. URL http://link.springer.com/chapter/10.1007/978-3-540-76298-0_52.

Jeremy Bentham. *The Collected Works of Jeremy Bentham*. 1968 – ongoing.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

Tim Causer and Melissa Terras. Crowdsourcing Bentham: beyond the traditional boundaries of academic history. *International Journal of Humanities and Arts Computing*, 8(1):46–64, 2014a.

Tim Causer and Melissa Terras. Many hands make light work. Many hands together make merry work: Transcribe Bentham and crowdsourcing manuscript collections. *Crowdsourcing Our Cultural Heritage*, pages 57–88, 2014b.

Tim Causer, Justin Tonra, and Valerie Wallace. Transcription maximized; expense minimized? Crowdsourcing and editing The Collected Works of Jeremy Bentham. *Literary and Linguistic Computing*, 27(2):119–137, June 2012. ISSN 0268-1145, 1477-4615. doi: 10.1093/llc/fqs004. URL http://llc.oxfordjournals.org/cgi/doi/10.1093/llc/fqs004.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 121–124. ACM, 2013. URL http://dl.acm.org/citation.cfm?id=2506198.

Serge Fleury and Maria Zimina. Trameur: A Framework for Annotated Text Corpora Exploration. In *COLING (Demos)*, pages 57–61, 2014. URL http://anthology.aclweb.org/C/C14/C14-2.pdf#page=69.

Serge Heiden. The txm platform: Building open-source textual analysis software compatible with the tei encoding scheme. In *24th Pacific Asia Conference on Language, Information and Computation*, pages 389–398. Institute for Digital Enhancement of Cognitive Development, Waseda University, 2010.

Serge Heiden, Jean-Philippe Magué, and Bénédicte Pincemin. Txm: Une plateforme logicielle open-source pour la textométrie-conception et développement. In *10th International Conference on the Statistical Analysis of Textual Data-JADT 2010*, volume 2, pages 1021–1032. Edizioni Universitarie di Lettere Economia Diritto, 2010.

Bettina Heintz. Zahlen, Wissen, Objektivitt: Wissenschaftssoziologische Perspektiven. In Andrea Mennicken and Hendrik Vollmer, editors, *Zahlenwerk*, pages 65–85. VS Verlag fr Sozialwissenschaften, Wiesbaden, 2007. ISBN 978-3-531-15167-0. URL http://link.springer.com/10.1007/978-3-531-90449-8_4.

Vera Khovanskaya, Eric PS Baumer, and Phoebe Sengers. Double binds and double blinds: evaluation tactics in critically oriented HCI. In *Proceedings of The Fifth Decennial Aarhus Conference on Critical Alternatives*, pages 53–64. Aarhus University Press, 2015. URL http://dl.acm.org/citation.cfm?id=2882863.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*,

pages 21–26. Association for Computational Linguistics, 2010. URL http://dl.acm.org/citation.cfm?id=1859668.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

Frédérique Mélanie, Johan Ferguth, Katherine Gruel, and Thierry Poibeau. Archaeology in the digital age: From paper to databases. In *Digital Humanities 2015*, 2015.

Pablo N. Mendes, Max Jakob, Andrs Garca-Silva, and Christian Bizer. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM, 2011. URL http://dl.acm.org/citation.cfm?id=2063519.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. Digging in the Dirt: Extracting Keyphrases from Texts with KD. In *Second Italian Conference on Computational Linguistics CLIC-It 2015*, Italy, 2015.

Paul Rayson. From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4): 519–549, 2008. doi: http://dx.doi.org/10.1075/ijcl.13.4.06ray. URL http://www.jbe-platform.com/content/journals/10.1075/ijcl.13.4.06ray.

Bernhard Rieder and Theo Röhle. Digital methods: Five challenges. In David Berry, editor, *Understanding digital humanities*, pages 67–84. Palgrave, 2012.

Pablo Ruiz Fabo, Clément Plancq, and Thierry Poibeau. More than word cooccurrence: Exploring support and opposition in international climate negotiations with semantic parsing. In *LREC: The 10th Language Resources and Evaluation Conference*, pages 1902 – 1907, 2016.

Alix Rule, Jean-Philippe Cointet, and Peter S. Bearman. Lexical shifts, substantive changes, and continuity in State of the Union discourse, 17902014. *Proceedings of the National Academy of Sciences*, 112(35):10837–10844, September 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1512221112. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1512221112.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing,*, Manchester, UK, 1994.

A.H. Toselli and E. Vidal. Handwritten text recognition results on the Bentham collection with improved classical n-gram-HMM methods. In *International Workshop on Historical Document Imaging and Processing (HIP)*. ACM, August 2015.

Peter D Turney. Learning algorithms for keyphrase extraction. *Information retrieval*, 2(4):303–336, 2000.

Julie Weeds and David Weir. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–475, 2005. URL http://www.mitpressjournals.org/doi/abs/10.1162/089120105775299122.