THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à l'École normale supérieure

Concept-Based and Relation-Based Corpus Navigation: Applications of Natural Language Processing in Digital Humanities

Ecole doctorale n°540

TRANSDISCIPLINAIRE LETTRES / SCIENCES

Spécialité SCIENCES DU LANGAGE

COMPOSITION DU JURY :

Mme. BEAUDOUIN Valérie
Télécom ParisTech, Rapporteur

Mme. SPORLEDER Caroline
Universität Göttingen, Rapporteur

M. GANASCIA Jean-Gabriel
Université Paris 6, Membre du jury

Mme. GONZÁLEZ-BLANCO Elena
UNED Madrid, Membre du jury

Mme. TELLIER Isabelle
Université Paris 3, Membre du jury

Mme. TERRAS Melissa
University College London, Membre du jury

Soutenue par Pablo **RUIZ FABO**
le 23 juin 2017

Dirigée par **Thierry POIBEAU**

ENS

PSL RESEARCH UNIVERSITY PARIS

PSL Research University
École normale supérieure

Doctoral Thesis

Concept-Based and Relation-Based Corpus Navigation: Applications of Natural Language Processing in Digital Humanities

Author:
Pablo Ruiz Fabo

Supervisor:
Thierry Poibeau

Research Unit: Laboratoire LATTICE
École doctorale 540 – Transdisciplinaire Lettres / Sciences

Defended on June 23, 2017

Thesis committee:

Valérie Beaudouin	Télécom ParisTech	Rapporteur
Jean-Gabriel Ganascia	Université Paris 6	Examinateur
Elena González-Blanco	UNED Madrid	Examinateur
Caroline Sporleder	Universität Göttingen	Rapporteur
Isabelle Tellier	Université Paris 3	Examinateur
Melissa Terras	University College London	Examinateur

Abstract

Social sciences and Humanities research is often based on large textual corpora, that it would be unfeasible to read in detail. Natural Language Processing (NLP) can identify important concepts and actors mentioned in a corpus, as well as the relations between them. Such information can provide an overview of the corpus useful for domain-experts, and help identify corpus areas relevant for a given research question.

To automatically annotate corpora relevant for Digital Humanities (DH), the NLP technologies we applied are, first, Entity Linking, to identify corpus actors and concepts. Second, the relations between actors and concepts were determined based on an NLP pipeline which provides semantic role labeling and syntactic dependencies among other information. Part I outlines the state of the art, paying attention to how the technologies have been applied in DH.

Generic NLP tools were used. As the efficacy of NLP methods depends on the corpus, some technological development was undertaken, described in Part II, in order to better adapt to the corpora in our case studies. Part II also shows an intrinsic evaluation of the technology developed, with satisfactory results.

The technologies were applied to three very different corpora, as described in Part III. First, the manuscripts of Jeremy Bentham. This is a 18th–19th century corpus in political philosophy. Second, the PoliInformatics corpus, with heterogeneous materials about the American financial crisis of 2007–2008. Finally, the *Earth Negotiations Bulletin* (ENB), which covers international climate summits since 1995, where treaties like the Kyoto Protocol or the Paris Agreements get negotiated.

For each corpus, navigation interfaces were developed. These user interfaces (UI) combine networks, full-text search and structured search based on NLP annotations. As an example, in the ENB corpus interface, which covers climate policy negotiations, searches can be performed based on relational information identified in the corpus: The negotiation actors having discussed a given issue using verbs indicating support or opposition can be searched, as well as all statements where a given actor has expressed support or opposition. Relation information is employed, beyond simple co-occurrence between corpus terms.

The UIs were evaluated qualitatively with domain-experts, to assess their potential usefulness for research in the experts' domains. First, we payed attention to whether the corpus representations we created correspond to experts' knowledge of the corpus, as an indication of the sanity of the outputs we produced. Second, we tried to determine whether experts could gain new insight on the corpus by using the applications, e.g. if they found evidence unknown to them or new research ideas. Examples of insight gain were attested with the ENB interface; this constitutes a good validation of the work carried out in the thesis. Overall, the applications' strengths and weaknesses were pointed out, outlining possible improvements as future work.

Keywords: Entity Linking, Wikification, Relation Extraction, Proposition Extraction, Corpus Visualization, Natural Language Processing, Digital Humanities

Résumé

Note : Le résumé étendu en français commence à la p. 263.

La recherche en Sciences humaines et sociales repose souvent sur de grandes masses de données textuelles, qu'il serait impossible de lire en détail. Le Traitement automatique des langues (TAL) peut identifier des concepts et des acteurs importants mentionnés dans un corpus, ainsi que les relations entre eux. Ces informations peuvent fournir un aperçu du corpus qui peut être utile pour les experts d'un domaine et les aider à identifier les zones du corpus pertinentes pour leurs questions de recherche.

Pour annoter automatiquement des corpus d'intérêt en Humanités numériques, les technologies TAL que nous avons appliquées sont, en premier lieu, le liage d'entités (plus connu sous le nom de Entity Linking), pour identifier les acteurs et concepts du corpus ; deuxièmement, les relations entre les acteurs et les concepts ont été déterminées sur la base d'une chaîne de traitements TAL, qui effectue un étiquetage des rôles sémantiques et des dépendances syntaxiques, entre autres analyses linguistiques. La partie I de la thèse décrit l'état de l'art sur ces technologies, en soulignant en même temps leur emploi en Humanités numériques.

Des outils TAL génériques ont été utilisés. Comme l'efficacité des méthodes de TAL dépend du corpus d'application, des développements ont été effectués, décrits dans la partie II, afin de mieux adapter les méthodes d'analyse aux corpus dans nos études de cas. La partie II montre également une évaluation intrinsèque de la technologie développée, avec des résultats satisfaisants.

Les technologies ont été appliquées à trois corpus très différents, comme décrit dans la partie III. Tout d'abord, les manuscrits de Jeremy Bentham, un corpus de philosophie politique des 18^{e} et 19^{e} siècles. Deuxièmement, le corpus PoliInformatics, qui contient des matériaux hétérogènes sur la crise financière américaine de 2007–2008. Enfin, le *Bulletin des Négociations de la Terre* (*ENB* dans son acronyme anglais), qui couvre des sommets internationaux sur la politique climatique depuis 1995, où des traités comme le Protocole de Kyoto ou les Accords de Paris ont été négociés.

Pour chaque corpus, des interfaces de navigation ont été développées. Ces interfaces utilisateur combinent les réseaux, la recherche en texte intégral et la recherche structurée basée sur des annotations TAL. À titre d'exemple, dans l'interface pour le corpus ENB, qui couvre des négociations en politique climatique, des recherches peuvent être effectuées sur la base d'informations relationnelles identifiées dans le corpus : les acteurs de la négociation ayant abordé un sujet concret en exprimant leur soutien ou leur opposition peuvent être recherchés. Le type de la relation entre acteurs et concepts est exploité, au-delà de la simple co-occurrence entre les termes du corpus.

Les interfaces ont été évaluées qualitativement avec des experts de domaine, afin d'estimer leur utilité potentielle pour la recherche dans leurs domaines respectifs. Tout d'abord, on a vérifié que les représentations générées pour le contenu des corpus sont

en accord avec les connaissances des experts du domaine, pour déceler des erreurs d'annotation. Ensuite, nous avons essayé de déterminer si les experts pouvaient être en mesure d'avoir une meilleure compréhension du corpus grâce à l'utilisation des applications développées, par exemple, si celles-ci permettent de renouveler leurs questions de recherche existantes. On a pu mettre au jour des exemples où un gain de compréhension sur le corpus est observé grâce à l'interface dédiée au *Bulletin des Négociations de la Terre*, ce qui constitue une bonne validation du travail effectué dans la thèse. En conclusion, les points forts et faiblesses des applications développées ont été soulignés, en indiquant de possibles pistes d'amélioration en tant que travail futur.

Mots Clés : Liage d'entité, Entity Linking, Wikification, extraction de relations, extraction de propositions, visualisation de corpus, Traitement automatique des langues, Humanités numériques

Acknowledgements

I would like to thank my supervisor, Thierry Poibeau, for everything. I would also like to thank the other colleagues I did research with. The domain-experts who provided feedback about the applications in the thesis also need to be thanked. The thesis was carried out at the Lattice lab, which is a place to recommend for Linguistics, NLP, and Digital Humanities, and whose community I am thanking too. I had the chance to teach at some courses on corpus analysis tools and NLP applications, that's an experience I'm grateful for and the people who gave me the chance to do so need to be thanked, as well as the very dedicated co-workers I met there and the students for the experience. The people who had feedback at talks, conferences or schools also helped me develop the work in the thesis and thanks are due to them. Finally, I'd like to thank my former colleagues, the fine people at V2 who let me go to do this thesis, and also Queen St. people and others, with whom I also learned some of the things that were useful for the work here. The thesis is dedicated to my family who were always very supportive.

Contents

Abstract **iii**

Résumé **v**

INTRODUCTION **1**
Scientific Context 1
Contributions 3
Digital and Computational Humanities Orientation 5
Thesis Structure 6

I STATE OF THE ART **9**

Introduction **11**

1 Entity Linking in Digital Humanities **15**
1.1 Entity Linking 15
1.2 Related Technologies: Entity Linking, Wikification, NERC, NED and Word Sense Disambiguation 16
1.3 A Generic End-to-End Entity Linking Pipeline 18
1.4 Intrinsic Evaluation in Entity Linking 20
1.4.1 Evaluation Measures 20
1.4.2 Evaluating against Ever-Evolving KBs 21
1.4.3 Reference corpora 22
1.4.4 Example Results 22
1.5 Entity Linking and Related Technologies in Digital Humanities 23
1.5.1 Special applications of EL and NERC in DH 23
1.5.2 Generic-domain EL application in DH and its challenges 24
1.6 Challenges and Implications for our Work 26

2 Extracting Relational Information in Digital Humanities **29**
2.1 Introduction 29
2.1.1 The Information Extraction field 29
2.1.2 Technologies reviewed 30
2.2 Syntactic and Semantic Dependency Parsing 31
2.2.1 Syntactic Dependency Parsing 31

2.2.2 Semantic Role Labeling . . . 32
2.2.3 Parser examples . . . 33
2.2.4 Parser evaluation and example results . . . 34
2.3 Relation Extraction . . . 35
2.3.1 Traditional Relation Extraction . . . 36
2.3.2 Open Relation Extraction . . . 37
2.3.3 Evaluation in relation extraction and example results 39
2.3.4 Traditional vs. open relation extraction for DH . . . 42
2.4 Event Extraction . . . 43
2.4.1 Task description . . . 43
2.4.2 Approaches . . . 44
2.4.3 Evaluation and example results . . . 45
2.5 Applications in Digital Humanities . . . 45
2.5.1 Syntactic parsing applications . . . 46
2.5.2 Relation extraction applications . . . 47
2.5.3 Event extraction applications . . . 48
2.6 Summary and Implications for our Work . . . 49
2.6.1 Summary . . . 49
2.6.2 Implications for our work . . . 51

II NLP TECHNOLOGY SUPPORT 53

Introduction 55

3 Entity Linking System Combination 59
3.1 Introduction . . . 59
3.2 Related Work . . . 60
3.3 Annotation Combination Method . . . 60
3.3.1 Systems combined . . . 61
3.3.2 Obtaining individual annotator outputs . . . 62
3.3.3 Pre-ranking annotators . . . 63
3.3.4 Annotation voting scheme . . . 64
3.4 Intrinsic Evaluation Method . . . 65
3.5 Results and Discussion . . . 66
3.5.1 Results . . . 66
3.5.2 Discussion: Implications for DH research . . . 68
3.6 Summary and Outlook . . . 70

4 Extracting Relations between Actors and Statements 73
4.1 Introduction . . . 73
4.2 Proposition Extraction Task . . . 74
4.2.1 Proposition definition . . . 74

4.2.2 Corpus of application . . . 74
4.2.3 Proposition representation . . . 76
4.3 Related Work . . . 76
4.4 System Description . . . 78
4.4.1 NLP pipeline . . . 78
4.4.2 Domain model . . . 79
4.4.3 Proposition extraction rules . . . 81
4.4.4 Proposition confidence scoring . . . 84
4.4.5 Discussion about the approach . . . 85
4.5 Intrinsic Evaluation, Results and Discussion . . . 86
4.5.1 NLP pipeline evaluation . . . 86
4.5.2 Proposition extraction evaluation . . . 87
4.5.3 Discussion . . . 90
4.6 Summary and Outlook . . . 90

III APPLICATION CASES **93**

Introduction **95**

5 Concept-based Corpus Navigation: Bentham's Manuscripts and PoliInformatics **99**
5.1 Introduction . . . 99
5.2 Bentham's Manuscripts . . . 100
5.2.1 Corpus Description . . . 100
5.2.1.1 Structure of the corpus and TEI encoding . . . 101
5.2.1.2 Corpus sample in our study and preprocessing 102
5.2.2 Prior Analyses of the Corpus . . . 106
5.2.3 Corpus Cartography based on Entity Linking and Keyphrase Extraction . . . 109
5.2.3.1 Lexical Extraction . . . 109
5.2.3.2 Lexical Clustering and Network Creation . . . 113
5.2.3.3 Network Visualization . . . 116
5.2.4 User Interface: Corpus Navigation via Concept Networks . . . 117
5.2.4.1 User Interface Structure . . . 117
5.2.4.2 Search Interface . . . 117
5.2.4.3 Navigable Corpus Maps . . . 118
5.2.5 User Interface Evaluation with Experts . . . 124
5.2.5.1 Introduction and basic evaluation data . . . 124
5.2.5.2 Expected outcomes . . . 124
5.2.5.3 Evaluation task . . . 125

5.2.5.4 Results, discussion, and possible UI improvements . . . 126
5.2.5.5 Summary of the UI evaluation . . . 132
5.2.6 Summary and Outlook . . . 132
5.3 PoliInformatics . . . 135
5.3.1 Corpus Description . . . 135
5.3.1.1 Corpus sample in our study and preprocessing 136
5.3.2 Related Work . . . 137
5.3.2.1 Prior work on the corpus . . . 137
5.3.2.2 Prior tools related to our user interface . . . 138
5.3.3 Entity Linking Backend . . . 139
5.3.3.1 DBpedia annotations: acquisition, combination and classification . . . 139
5.3.3.2 Annotation quality assessment: confidence and coherence . . . 141
5.3.4 User Interface: Corpus Navigation with DBpedia Facets 145
5.3.4.1 Visual representation of annotation quality indicators . . . 145
5.3.4.2 Search and filtering functions . . . 147
5.3.4.3 Automatic annotation selection . . . 148
5.3.4.4 Result sorting . . . 150
5.3.5 User Interface Example Uses and Evaluation . . . 150
5.3.5.1 Using confidence scores . . . 151
5.3.5.2 Using coherence scores . . . 151
5.3.5.3 Examples of automatic annotation selection 153
5.3.5.4 Validating a corpus network . . . 154
5.3.5.5 A limitation: Actors unavailable in the knowledge base . . . 158
5.3.6 Summary and Outlook . . . 159

6 Relation-based Corpus Navigation: The *Earth Negotiations Bulletin* 163
6.1 Introduction . . . 163
6.2 Corpus Description . . . 164
6.2.1 The *Earth Negotiations Bulletin* . . . 164
6.2.2 Corpus sample in our study and preprocessing . . . 165
6.3 Prior Approaches to the Corpus . . . 166
6.3.1 Corpus cartography . . . 166
6.3.2 Grammar induction . . . 167
6.3.3 Corpus navigation . . . 167
6.4 NLP Backend: Proposition Extraction and Enrichment . . . 169
6.4.1 Proposition extraction . . . 170

6.4.2 Enriching proposition messages with metadata 171
6.5 User Interface: Corpus Navigation via Enriched Propositions 174
6.5.1 Search Workflows: Propositions, sentences, documents 175
6.5.2 Browsing for agreement and disagreement 182
6.5.3 UI Implementation 183
6.6 User Interface Evaluation with Domain-experts 184
6.6.1 Scope and approach 184
6.6.2 Hypotheses 185
6.6.3 Evaluation Task 186
6.6.4 Results and discussion 189
6.7 Summary and Outlook 195

CONCLUSION **199**
Expert Evaluation: Reproducing Knowledge and Gain of Insight . 199
Generic and Corpus-specific NLP Developments 203
Lessons Learned regarding Implementation 205
Final Remarks 206

Appendices **208**

A Term Lists for Concept-based Navigation **209**

B Domain Model for Relation-based Navigation **227**

C Test-Sets for Intrinsic Evaluation **235**

D Domain-Expert Evaluation Reports **237**

E List of Publications Related to the Thesis **261**

Résumé de la thèse en français **263**

Bibliography **303**

List of Figures

3.1 Entity Linking: Annotation voting scheme for system combination . . . 65

4.1 Proposition Extraction: Example sentences in the ENB corpus 75
4.2 Proposition Extraction: Generic rule . . . 82
4.3 Proposition Extraction: Rule for opposing actors . . . 82

5.1 UCL Transcribe Bentham Interface, with an example document 103
5.2 Bentham Corpus Sample: Distribution of pages per decade . 105
5.3 Bentham Corpus Sample: Distribution of pages across main content categories . . . 105
5.4 UCL Bentham Papers Database: Metadata-based Search . . . 108
5.5 UCL Libraries Digital Collections: Bentham Corpus Search . 108
5.6 Our Bentham User Interface Structure . . . 118
5.7 Bentham UI: Navigable concept map. Results for search query *power* . . . 120
5.8 Bentham UI: Network navigation by sequentially selecting neighbours . . . 121
5.9 Bentham UI: Heatmaps ~ Corpus areas salient in the 1810s and 1820s . . . 123
5.10 Bentham UI Evaluation: Example of nodes connecting two clusters in the 150 concept-mention map . . . 127
5.11 Bentham UI Evaluation: Searching the index to verify contexts of connected network-nodes (e.g. *vote* and *bribery*) . . . 127
5.12 Bentham UI Evaluation: Nodes matching query *power* in the 250 concept-mention map . . . 128
5.13 Bentham UI Evaluation: Terms matching *interest* in the 250 keyphrase map ~ Synonyms and antonyms for *sinister interest* 129
5.14 Bentham UI Evaluation: Area focused on by domain-expert as representing general Bentham concepts and the relation between them . . . 131
5.15 PoliInformatics UI: Results for query *credit ratings*, restricted to Organizations . . . 146
5.16 PoliInformatics UI: Description of functions . . . 149
5.17 PoliInformatics UI: Original vs. automatically selected results 154

5.18 PoliInformatics Organizations Network: Original vs. manually corrected using information on UI 155
5.19 PoliInformatics UI: Annotation quality measures suggesting errors 157

6.1 Sciences Po médialab's interface for the ENB corpus 168
6.2 Relation-based Corpus Navigation: System Architecture . . 170
6.3 Our UI for the *Earth Negotiations Bulletin* (ENB) corpus: Main View 175
6.4 ENB UI: Overview of actors making statements about *gender*, and of the content of their messages. 179
6.5 ENB UI: Comparing two actors' statements on *energy* via keyphrases and thesaurus terms extracted from their messages 181
6.6 ENB UI: Agree-Disagree View for the European Union vs. the Group of 77 183

List of Tables

1.1 Entity Linking example results for four public systems and datasets (Weak Annotation Match measure) 23
1.2 Varying performance of Entity Linking systems across corpora 25
1.3 Correlations between Entity Linking system performance and named-entity types in corpus 25

2.1 Comparison of Open Relation Extraction results 42

3.1 Entity Linking Results: Strong Annotation Match 67
3.2 Entity Linking Results: Entity Match 67
3.3 Keyphrase extraction results for the top three systems at SemEval 2010, Task 5. 69

4.1 Proposition Extraction: Confidence scoring features 84
4.2 Proposition confidence score examples 84
4.3 Proposition Extraction: NLP pipeline evaluation 86
4.4 Proposition Extraction Results: Exact Match 89
4.5 Proposition Extraction Results: Error types 89

6.1 Proposition-based Navigation: Basic data about domain-expert evaluation sessions 188

List of Abbreviations

ACL	Association For Computational Linguistics.
ADHO	Alliance Of Digital Humanities Organizations.
AoC	*Anatomy of a Financial Collapse* Congressional Report.
API	Application Programming Interface.
COP	Conference Of The Parties.
CSV	Comma-Separated Values.
DH	Digital Humanities.
EL	Entity Linking.
ENB	*Earth Negotiations Bulletin*.
FCIC	Federal Crisis Inquiry Commission.
GEXF	Graph Exchange XML Format.
HTML	HyperText Markup Language.
IPCC	Intergovernmental Panel On Climate Change.
JSON	JavaScript Object Notation.
KB	Knowledge Base.
NED	Named Entity Disambiguation.
NERC	Named Entity Recognition And Classification.
NLP	Natural Language Processing.
POS	Part Of Speech.
ROVER	Recognizer Output Voting Error Reduction.
SRL	Semantic Role Labeling.
TEI	Text Encoding Initiative.
UI	User Interface.
WSD	Word Sense Disambiguation.
XML	Extensible Markup Language.

Introduction

Scientific Context

Data relevant for social sciences and humanities research often takes the shape of large masses of unstructured text, which it would be unfeasible to analyze manually. Discussing the use of textual evidence in political science, Grimmer et al. (2013) list a variety of relevant text types, like regulations issued by different organizations, international negotiation documents, and news reports. They conclude that *"[t]he primary problem is volume: there are simply too many political texts"*. In the case of literary studies, scholars need to address the complete text of thousands of works spanning a literary period (Clement et al., 2008; Moretti, 2005, pp. 3–4). Such amounts of text are beyond a scholar's reading capacity, and researchers turn to automated text analyses that may facilitate understanding of relevant aspects of those textual corpora.

Some types of information that are generally useful to understand a corpus are actors mentioned in it (e.g. people, organizations, characters), core concepts or notions of specific relevance for the corpus domain, as well as the relations between those actors and those concepts. A widespread approach to gain an overview of a corpus relies on network graphs called concept networks, social networks or socio-technical networks depending on their content (see Diesner, 2012, esp. pp. 5, 84). In such graphs, nodes represent terms relevant in the corpus (actors and concepts), and the edges represent either a relation between the terms (like support or opposition), or a notion of proximity between them, based on overlap between their contexts. Creating networks requires then a method to identify nodes, as well as a way to extract relations between nodes or to define node proximity, such as different clustering methods.

Networks have yielded very useful results for social sciences and humanities research. To cite an example based on one of the corpora studied in this thesis, Baya-Laffite et al. (2016) and Venturini et al. (2014) created concept networks to describe key issues in 30 years of international climate negotiations described in the *Earth Negotiations Bulletin* (ENB) corpus, providing new insight regarding the evolution of negotiation topics.

Established techniques to extract networks from text exist, and networks offer useful corpus navigation possibilities. However, Natural Language Processing (Jurafsky et al., 2009) can complement widespread methods for network creation. Sequence labeling and disambiguation techniques like Entity Linking can be exploited to identify the network's nodes: actors and concepts. The automatic definition of network edges is usually based on node co-occurrence, while more detailed information about the relation between actors and concepts is not usually automatically identified for defining edges. Nonetheless, such information can also be obtained via Natural Language Processing (NLP) methods. As for corpus navigation, networks do not in themselves provide access to the corpus fragments that were used as evidence to create the networks. But they can be complemented with search workflows that allow a researcher to access the contexts for network nodes and the textual evidence for the relation between them.

Applying NLP for text analysis in social sciences and humanities poses some specific challenges. First of all, researchers in these domains work on texts displaying a large thematic and formal variety, whereas NLP tools have been trained on a small range of text-types, e.g. newswire (Plank, 2016). Second, the experts' research questions are formulated using constructs relevant to their fields, whereas core tools in an NLP pipeline (e.g. part-of-speech tagging or syntactic parsing) provide information expressed in *linguistic* terms. Researchers in social sciences, for example, are not interested in automatic syntactic analyses per se, but insofar as they provide evidence relevant for their research questions: e.g. *Which* actors *interact with each other in this corpus?*, or *What* concepts *does an actor mention, and showing what* attitudes *towards those concepts?* Adapting tools to deal with a large variety or corpora, and exploiting their outputs to make them relevant for the questions of experts in different fields is a challenge.

In the same way that exploiting NLP technologies to make them useful to experts in social sciences and humanities is challenging, evaluating the application of NLP tools to those fields also poses difficulties. A vast literature exists about evaluating NLP technologies using NLP-specific measures. However, these NLP measures do not directly answer questions about the usefulness for a domain expert of a tool that applies NLP technologies. Even less do they answer questions about potential biases induced by the technologies (e.g. focusing only on items with certain corpus frequencies), and how these biases affect potential conclusions to draw from the data (see examples in Rieder et al. (2012, p. 77), or discussions in Marciniak (2016)). As Meeks et al. (2012) state, research is needed with "as much of a focus on what the computational techniques obscure as reveal".

In summary, researchers in social sciences and humanities need ways to gain relevant access to large corpora. Natural Language Processing can help provide an overview of a corpus, by automatically extracting actors, concepts, and even the relation between them. However, NLP tools do not perform equally well with all texts and may require adaptation. Besides, the connection between these tools' outputs and research questions in a domain-expert's field need not be immediate. Finally, evaluating the usefulness of an NLP-based tool for a domain-expert is not trivial. The contributions of the thesis in view of these challenges are outlined in following.

Contributions

Bearing in mind the challenges above, this thesis presents ways to find, via NLP, relevant actors and core concepts in a corpus, and their exploitation for corpus navigation, both via network extraction, and via corpus search functions targeting corpus elements (paragraphs, sentences) that provide evidence for those actors and concepts.

Corpus navigation workflows

As a contribution towards obtaining useful overviews of corpora, two types of corpus navigation workflows are presented.

- First, **concept-based navigation**, where (full-text) search and networks are combined, and where the extraction of terms to model the corpus relies on a technology called *Entity Linking* (Rao et al., 2013). This technology finds mentions to terms from a knowledge repository (like Wikipedia) in a corpus, annotating the mentions with the term they refer to. Other sequence extraction technologies like Named Entity Recognition (p. 17) or keyphrase extraction (p. 112) have been used more commonly than Entity Linking for network creation. The contribution here is assessing the viability of this technology, which has been used comparatively infrequently to create networks, as a means to detect concepts and actors in a corpus.
- Second, **relation-based navigation**. We formalize relations within *propositions*. A proposition is defined as a triple containing a subject, an object and a predicate relating both. Depending on the type of predicate, the nature of the subject and object will differ, e.g. if the predicate is a reporting verb, the subject will be a speaker, and the object will be the speaker's statement. Relation-based navigation allows for structured searches on the corpus based on proposition elements: actors, concepts and the relations between both, identifying the sentences that are evidence for such relations. The relations mediating between two terms

(e.g. support or opposition) are identified automatically, allowing for the creation of networks where edges encode an explicitly identified type of relation, rather than encoding a general notion of co-occurrence.

From the network creation point of view, the contribution here is integrating an additional source of evidence (relations expressed in the text) in the network creation process, so that the networks can encode a more precise relation between nodes than proximity.

From the corpus navigation point of view, the contribution is an easier access to information about actors and concepts than when not using propositions to guide navigation: A search interface was created, where users can navigate the corpus according to all proposition elements, quickly arriving at sentences containing given concepts or actors, or showing a relation between them.

Relations automatically extracted from text have been incorporated in network creation in Van Atteveldt (2008), Van Atteveldt et al. (2017), besides Diesner (2012) and references reviewed therein. However, I use a different source of relation information to those works, focusing equally on nominal and verbal predicates, besides providing a user interface (UI) to navigate results.

NLP output adaptation

As a second contribution, the thesis provides examples of ways to exploit NLP tools and their outputs for corpora of different characteristics, and for specific user needs.

- As regards **Entity Linking**, the quality of results provided by this technology varies a lot depending on the corpus (see Cornolti et al., 2013 for results comparison). In the thesis, several entity linking tools are combined in order to adapt to different corpora, maintaining a more uniform quality in spite of corpus variety.
- Regarding the **extraction of relation information**, actors, their messages, and the predicates relating both were identified in a corpus of international climate negotiations, with certain non-standard linguistic traits (e.g. personal pronouns *he/she* can refer to countries, and the subjects of reporting verbs tend to be countries, rather than people). NLP outputs were adapted to deal with such corpus-specific usage features. Moreover, the NLP technology used to identify propositions in the corpus, called Semantic Role Labeling (SRL) (Carreras et al., 2005), provides outputs that make sense to a linguist (they represent fine-grained semantic distinctions in verb and noun meaning), but can be opaque to researchers in

other domains. Outputs of SRL were mapped to categories such as *Actor*, *Action* and *Message*, relevant to social scientists examining *who said what* in diplomatic negotiations.

It can also be considered that another way in which the outputs of NLP tools were adapted to domain experts' needs is the mere fact of providing user interfaces (UIs) displaying the NLP-based extractions as searchable elements and as navigation elements (e.g. as facets for filtering results), so that experts can have a structured access to the corpus based on those NLP outputs.

Domain-relevant Evaluation

The solutions developed in this thesis are intended to help social sciences and humanities researchers analyze their corpora, providing new quantitative and qualitative data for them to assess. Extensive evaluation of a tool by domain-experts, attending to aspects like the actual usefulness of the tool for their research questions, tool-induced biases, and their impact on the research, is rare (Traub et al., 2015, p. 1).

The contribution in this respect is offering an example of qualitative evaluation of a tool with domain-experts, based on one-hour interviews with the experts while they used the tool. This can be seen as an original initiative, given the rarity of such evaluations, which is spurring emergent domains like *tool criticism* (Traub et al., 2015).

Digital and Computational Humanities Orientation

An informal definition of the scope of Digital Humanities (DH) was given by Fitzpatrick (2010), in a well-cited blog, as *"a nexus of fields within which scholars use computing technologies to investigate the kinds of questions that are traditional to the humanities [. . .] or who ask traditional kinds of humanities-oriented questions about computing technologies"*. Though informal, this broad characterization agrees with the variety of work described as Digital Humanities in overviews of the field like (Berry, 2012, pp. 1–20; Schreibman et al., 2004).[1]

More recently, some authors (see Biemann et al., 2014, particularly pp. 87–91) discuss that they see two types of research in the work described as DH in the overviews just cited. First, what these authors (i.e. Biemann et al., 2014) call Digital Humanities "proper", which in their characterization focuses on digital resource creation and access. Second, research which these authors call *Computational Humanities*, and which analyzes digital materials with advanced computational techniques, while trying to assess the value of those

[1] This is again a broad characterization, for critical commentary and debate on the concept of Digital Humanities, a historical overview of how the term came about, and related disciplines, see Terras et al. (2013).

computational means for addressing humanities questions. They see work in what they term Computational Humanities as situated in a continuum between the Humanities or the Digital Humanities (in the sense they use the latter term) and Computer Science. This thesis applies NLP technologies, adapting them to specific use cases, integrating them in user interfaces to make the technology more easily usable by domain-experts from humanities and social sciences. Besides, a critical reflection on the computational tools and methods developed is provided, based on an evaluation by domain-experts who are expected to benefit from those technological means. As such, should we want to adopt the *Digital vs. Computational Humanities* terminology sometimes proposed, the work here can be considered within the Computational Humanities.

Thesis Structure

The rest of the thesis is organized as follows. The main technologies applied in the thesis are Entity Linking (EL) and several technologies that allow extracting relation information, especially Semantic Role Labeling and syntactic dependency parsing. **Part I** covers the related state of the art, paying attention to how the technologies are applied in Digital Humanities. Chapter 1 addresses Entity Linking and Chapter 2 examines methods for extracting relational information.

Part II describes the approaches developed in the thesis to apply those technologies, Chatper 3 for Entity Linking and Chapter 4 for extracting relations between speakers and their messages in a political negotiation corpus, bearing in mind the need to adapt standard NLP technologies to corpus characteristics and user needs.

Part III discusses application cases of the technologies just described. Chapter 5 presents the idea of **concept-based corpus navigation**, where the lexical items used to model the corpus have been identified using entity linking. Two corpora were used as a case-study.
The first corpus is the **unedited manuscripts of Jeremy Bentham** (Causer et al., 2014a), an 18^{th}–19^{th} century English philosopher and social reformer. The corpus consists of ca. 4,7 million words. Different types of concept networks, static and dynamic across time, were created. A UI was developed to navigate the corpus, via full-text search or via networks. A domain-expert provided feedback on the system, confirming that the networks produced cover the conceptual areas of Bentham's thought.
The second corpus studied is a subset of ca. 400,000 words from the **Poli-Informatics corpus** (Smith et al., 2014), about the 2008 American financial

crisis. The corpus contains heterogeneous material like transcripts for hearings carried out by a government-appointed commission to investigate the causes of the crisis, or official reports produced by Congress about that same topic. The corpus was annotated with a combination of Entity Linking systems, and a UI was developed to allow experts to select the best annotations to model the corpus with, based on extraction quality criteria also present on the UI (e.g. confidence scores). Networks were created for the corpus based on the annotations selected. Experts can also navigate the corpus using those annotations as facets, or using full-text search. Examples are shown that suggest the benefits of the system proposed for a domain-expert: e.g. noisy entities can be removed from the analysis based on metrics like low confidence scores.

Chapter 6 presents an application of relation-based navigation in order to examine support and opposition in a corpus of international climate negotiations, the *Earth Negotiations Bulletin* (Vol. 12).[2] The corpus comprises ca. 500,000 words. A domain model including actors and reporting predicates (verbs and nouns) was applied on the output of an NLP pipeline (Agerri et al., 2014) offering Semantic Role Labeling (Carreras et al., 2005) and syntactic dependency parsing (Buchholz et al., 2006), besides pronominal anaphora resolution (Pradhan et al., 2011). Based on the output of the NLP pipeline, combined with the domain-model, it was possible to identify relations between actors and their messages, extracting propositions. Propositions are defined as $\langle actor, predicate, message \rangle$ triples. They capture *who said what* in the negotiation, and via what type of predicate: a *support* predicate or an *opposition* one. Additionally, the propositions' messages were enriched with automatic keyphrase extraction, generic-domain entity linking to DBpedia (Auer et al., 2007), and domain-specific linking to a climate-policy thesaurus (Bauer et al., 2011). This allows to relate keyphrases and entities to the actors who emitted the messages containing them, via the relevant relation (support or opposition). Evaluation interviews, of over one hour each, were performed with three domain-experts. A report on the evaluation sessions as well as a critical discussion of the findings is provided. The evaluations suggested that the UI helps experts gain an overview of the behaviour of actors in the negotiations, of the treatment of negotiation issues, and can also help gain new insight on certain actors and issues.

Publications Related to the Thesis

The technology, user interfaces developed, or the expertise acquired through the thesis, contributed to the following publications or presentations:

[2] http://www.iisd.ca/vol12

CONCEPT-BASED NAVIGATION

Technology for Entity Linking:

– **SEM 2015, International Joint Conference for Computational and Lexical Semantics* (Ruiz Fabo et al., 2015a)

– *SemEval 2015, International Workshop on Semantic Evaluation* (Ruiz Fabo et al., 2015b)

Applications to corpus navigation:

Bentham corpus: *DH 2016, Digital Humanities Conference*. (Tieberghien et al., 2016)

PoliInformatics corpus:

– *NAACL 2015, North American Association for Computational Linguistics, Demo Track.* (Ruiz Fabo et al., 2015c).

– *DH 2015, Digital Humanities Conference.* (Poibeau et al., 2015).

Improving topic models with entity-based labeled LDA: *IJCoL, Italian Journal of Computational Linguistics, Special Issue on DH and NLP* (Lauscher et al., 2016)

RELATION-BASED NAVIGATION: Both the backend **technology** and the **application** (user interface for the **Earth Negotiations Bulletin** Corpus) were presented at:

– *LREC 2016, International Language Resources and Evaluation Conference.* (Ruiz Fabo et al., 2016b)

– *DH 2016, Digital Humanities Conference.* (Ruiz Fabo et al., 2016a)

A publication list giving the complete references grouped by publication type can be found in Appendix E.

Part I

State of the Art

State of the Art: Introduction

In the course of this thesis we had the opportunity to work with a diverse range of corpora, relevant for social and political science and for the humanities. The volume of these corpora is large enough (0.5 to 5 million words) for text analysis technologies to be a useful help for experts wishing to study the corpora.

Our first corpus comes from the 2014 PoliInformatics NLP challenge, an international workshop hosted at the Conference of the Association for Computational Linguistics. This challenge sought to examine how Natural Language Processing (NLP) can help analyze a social and political phenomenon like the 2007-8 American Financial Crisis, based on heterogeneous written sources like Congress Hearing transcripts and Congress reports. The open-ended questions posed by the challenge were *Who was the Financial Crisis?* and *What was the Financial Crisis?*

A technology that immediately comes to mind regarding these *Who* and *What* questions is Entity Linking (EL), which finds mentions to terms from a knowledge repository in a corpus, and tags those mentions with the relevant term. For instance, it spots mentions to Wikipedia terms like person and organization names, or technical terms in economic policy. This allows us to relate documents or paragraphs discussing the same issues, to gain an overview of how they are being discussed in the corpus.

The second corpus we had access to consists in ca. 5 million words from the unedited manuscripts of Jeremy Bentham (1748–1832), the British philosopher and social reformer. These manuscripts are currently being transcribed by volunteers via crowdsourcing, in an effort led by University College London, who owns most of the manuscripts. We had a collaboration with UCL Digital Humanities, to perform text mining on the corpus. Here again, we saw Entity Linking as a way to get a first overview of this large volume of textual content, which had not previously been analyzed with automatic means, identifying core notions in it.

The third corpus we had the occasion to work with is the *Earth Negotiations Bulletin (ENB)*, which consists in daily reports on international climate negotiations, detailing each party's statements in the negotiation. The 21st UN Climate Change Conference took place in Paris in 2015, and, besides the

corpus, we had access to political science experts working on those issues, as we were collaborating with Sciences Po on automatic text analysis of related materials.

The ENB corpus reports on negotiation processes. It is then important to know not only who emitted a message in the negotiations, and what issues were dealt with, but also, *who* addressed *what* issue and *how* (i.e. in an opposing or supporting manner). In other words, besides a notion of concepts and actors, to analyze this negotiation corpus in more depth, we needed to find relations between those concepts and actors. NLP has long worked on relation extraction, and we applied this technology to the ENB corpus.

In short, analysis needs for the corpora we had the opportunity to work with, based on collaborations (Bentham and ENB), or on an international challenge (PoliInformatics), led us to focus on two NLP technologies: Entity Linking and Relation Extraction. Part I in the thesis surveys the state of the art in these technologies, particularly as relevant to Digital Humanities (DH) application cases.

Chapter 1

Entity Linking in Digital Humanities

1.1 Entity Linking

Entity Linking (EL) (Rao et al., 2013) looks in a text for mentions to a knowledge base's concepts, and annotates those textual mentions with the relevant concept from the knowledge base (KB). A main use of the technology is relating to each other documents or textual spans referring to the same KB concept, abstracting away from variability in the ways the concept is expressed. For instance, a scientist who won the 1911 Nobel Prize in Chemistry can be mentioned in a text as *Curie*, *Marie Skłodowska Curie*, *Mrs. Curie* etc. Entity Linking will annotate any of those sequences with the relevant term in the knowledge base, i.e. the DBpedia term *Marie_Curie* [1], assuming a system that links against the DBpedia KB. Besides dealing with variability in the way a KB-term is expressed in texts, Entity Linking systems also need to assign the correct KB-term to textual mentions ambiguous across several terms. E.g. the mention *Curie* could refer to both *Pierre_Curie*[2] and *Marie_Curie*, among other terms.

The knowledge bases linked to are usually general ones like DBpedia (Auer et al., 2007), Freebase (Bollacker et al., 2008), Yago (Suchanek et al., 2007) or BabelNet (Navigli et al., 2012). However, domain-specific repositories can also be targeted (e.g. Frontini et al., 2015, where the KB contains specialized resources for French literature). The KBs linked to are usually Linked Open Data repositories (i.e. public repositories that contain structured machine readable information accessible through query protocols like SPARQL, as part of data resources in the Semantic Web).[3] As such, enriching a corpus via Entity Linking can serve as an initial step to publish the corpus annotated with entities in a linked data format. This is another source of interest for the technology.

[1] http://dbpedia.org/page/Marie_Curie
[2] http://dbpedia.org/page/Pierre_Curie
[3] https://www.w3.org/standards/semanticweb/

The focus in this thesis is on applying general-domain entity linking to DH corpora. The rest of the chapter is organized thus. In 1.2, EL and some related technologies are described, and the definition of EL adopted here is presented. In 1.3, the steps in a generic EL workflow are introduced: mention detection and disambiguation. Evaluation methods in EL are discussed in 1.4. After that, some applications of EL and related technologies in DH are presented (1.5), looking at both the generic domain EL tools that I focus on, and domain-specific applications. Finally, 1.6 outlines how the thesis relates to the technology reviewed in the chapter.

1.2 Related Technologies: Entity Linking, Wikification, NERC, NED and Word Sense Disambiguation

Some authors (e.g. Chang et al., 2016; Hachey et al., 2014) distinguish between two tasks: First, **Entity Linking**, where only mentions corresponding to named entities are considered, a *named entity* (Nadeau et al., 2007) being a lexical sequence from a given inventory of types, like persons, organizations, locations, products, etc. Second, **Wikification**, where mentions to any term present in a knowledge base like Wikipedia (or its semantic web version, DBpedia) are considered, without restricting to a series of categories the set of terms to be linked.[4]

In this thesis the term ***Entity Linking*** is used to refer to both Entity Linking "proper" and Wikification, for several reasons. First, the literature does not uniformly distinguish between both terms; several classic articles that describe systems linking to any Wikipedia page refer to their contribution as annotating Wikipedia *entities* in a corpus (Cornolti et al., 2013; Ferragina et al., 2010; Kulkarni et al., 2009; Mendes et al., 2011).[5] Second, the set of sequence types considered as named entities has broadened since this term's first definition (Grishman et al., 1996), which only included person names, organizations, locations, time, percentage, and currency expressions. For instance, the Extended Named Entity Hierarchy presented in (Sekine et al., 2004) contains around 200 types, including categories like *religion* or *colour*.[6] Finally, the focus of the thesis is assessing to what an extent annotating text with DBpedia terms (of all types) is helpful to domain experts in several corpus navigation applications, and a nuanced distinction between Entity Linking and Wikification are not central to this end.

For reasons related to those in the preceding paragraph, I will speak indistinctly of linking text to a KB's *concepts*, *entities* or, more neutrally, *terms*. This

[4]The set of terms to link to does exclude Wikipedia pages like lists or disambiguation pages.
[5]More precisely, most of these authors refer to Wikipedia's terms as *Wikipedia entities* or *concepts* synonymously. Kulkarni et al. (2009) use the word *entity* only.
[6]See `http://nlp.cs.nyu.edu/ene/version6_1_0eng.html` for the type definitions.

is in line with the way this terminology is used in the literature (Cornolti et al., 2013; Ferragina et al., 2010; Mendes et al., 2011).

Two technologies related to EL are **Named Entity Recognition and Classification (NERC or NER)** and **Named Entity Disambiguation (NED)**. NERC consists in detecting sequences called *Named Entities*, just described. The classification part consists in assigning them a type from a type inventory. NERC is often the first step in an Entity Linking pipeline (this applies however to detecting mentions to entity-like KB terms only, not to any type of KB term). NED refers to a later step in an EL pipeline: Once potential mentions to KB-terms have been spotted in a text, the NED step chooses the most likely KB-term for each mention. This step involves disambiguation, since, as pointed out above, a given mention in a text (e.g. *Curie*) can refer to several KB-terms (e.g. *Pierre_Curie*,[2] *Marie_Curie*[1] and the radioactivity unit *Curie*[7]).

As a final terminology remark, the term *Entity Linking* is in fact sometimes used to describe systems performing NED only. These systems take as their input text where the mentions that need to be linked to the KB have already been identified, and assign a KB-term to them, if appropriate KB-candidates are found.

A final technology related to EL to be mentioned here is called **Word Sense Disambiguation (WSD)** (Agirre et al., 2007; Navigli, 2009; 2012). Both in EL and WSD, the task assigns to textual mentions the correct item from a reference inventory. In WSD, lexical items are disambiguated against an inventory of word-senses, like the senses assigned to each lemma in a dictionary, and, in EL, disambiguation takes place against an encyclopedic inventory (like Wikipedia and similar knowledge-bases). A difference, mentioned by Moro et al., 2014, is that EL, unlike WSD, can attempt to disambiguate partial mentions (e.g. a person's last name like *Byron*, without the first name) to the relevant KB-term (e.g. *Ada Byron* or *Lord Byron*). Moro et al., 2014 propose a joint EL/WSD approach, linking to a knowledge-base integrating both lexicographic and encyclopedic knowledge (Navigli et al., 2012), showing how disambiguating word senses can help Entity Linking and vice-versa. WSD is not applied in this thesis,[8] but it is a useful technology to help gain automatic understanding of textual content, and the graph-based and classification methods used in WSD are related to methods employed in EL. These are reasons to mention the WSD technology here.

[7] http://dbpedia.org/page/Curie

[8] We used the Babelfy tool, which implements the approach in Moro et al., 2014, but we did not exploit word-senses systematically. We only used the subset of its results that has a corresponding DBpedia entities or concepts.

1.3 A Generic End-to-End Entity Linking Pipeline

The thesis focuses on combining the results of end-to-end entity linking systems, which take a text as input and annotate KB entities in it. Examples of such systems are early tools like (Bunescu et al., 2006), (Cucerzan, 2007) and (Mihalcea et al., 2007), or newer systems like the ones I have combined in this thesis: TagMe2 (Ferragina et al., 2010; Cornolti et al., 2013), DBpedia Spotlight (Daiber et al., 2013; Mendes et al., 2011), Wikipedia Miner (Milne et al., 2008a), AIDA (Hoffart et al., 2011) and Babelfy (Moro et al., 2014).

And end-to-end EL system performs three steps:

1. **Mention detection or spotting**: Textual sequences that can potentially be linked to the KB are identified.
2. **Candidate generation**: This consists in mapping mentions, detected in the previous step, to term-labels in the KB that can be good matches for the mention.
3. **Mention disambiguation**: The optimal KB-term is selected among the candidates provided by the previous step. If no candidate matches the requirements (e.g. passing an adequacy threshold), the mention remains unlinked.

A brief discussion of these steps follows (see Ji et al., 2014 for more detailed descriptions of different methods to implement the workflow).

Spotting can be dictionary based (e.g. based on a dictionary with the anchor-text for all Wikipedia links, as a representation of textual mentions that can refer to Wikipedia pages), or can be based on Named Entity Recognition and Classification (NERC). A spotting dictionary can be enriched with the probability that a mention refers to each of the KB-terms it links to, and this in turn can be exploited in the later step of mention disambiguation.

Candidate generation can be based on a variety of techniques. The goal is retrieving a set of KB-term labels that are likely matches for a textual mention. To this end, simple string equality and string-similarity approaches can be applied, but also acronym generation, or other string transformations, like reducing a person name to its initials (see Rao et al., 2013, p. 6). Wikipedia's link structure (i.e. redirects and disambiguation pages) can also be used for candidate generation. It is useful for acronym expansion or for nicknames, e.g., in Wikipedia, the term *the Mile High City*[9] redirects to *Denver* (in Colorado),[10] which makes KB-term *Denver* a candidate for the textual mention *the Mile High City*. In systems where spotting is dictionary-based

[9] https://en.wikipedia.org/w/index.php?title=The_Mile_High_City&redirect=no
[10] https://en.wikipedia.org/wiki/Denver

(not based on NERC), variants for textual mentions may be included directly in the dictionary, rather than generated on the fly.

Candidate disambiguation usually considers the proportion of times a textual mention links to each KB-term. This acts as a prior probability that a KB-term is the correct link for the mention. Besides, disambiguation compares (a) tokens in the context of a textual mention and (b) tokens (words or link-anchors) in the KB's definition for the term or the term's page overall. *Overlap* between those two sets of tokens is another one of the factors defining the strength of each candidate for each mention. Evidence from context vector overlap is sometimes referred to as a *local features* (Ratinov et al., 2011). Besides overlap between mention context and KB text, most systems also implement a measure of *coherence* among the KB candidate terms proposed for mentions in a subset of the corpus (e.g. in the same document, or in a window of paragraphs or sentences inside a document). Such measures are sometimes called *global features*.

Coherence between KB candidates is defined differently depending on the system. The measure in (Strube et al., 2014) is based on Wikipedia category overlap. Milne et al. (2008b) use a graph-based notion of coherence, relying on common inlinks to two pages from a third Wikipedia page as the basis of relatedness between those two pages (see Equation 5.2 for the formal definition). Other systems have also adopted this or similar graph-based measures (Ferragina et al., 2010; Hoffart et al., 2011). A new disambiguation method was presented by Moro et al. (2014), where coherence takes into account the proportion of a mention's occurrences covered by each KB candidate term, besides a graph-based component whereby candidates lesser connected to other candidates (via links in the BabelNet KB) are pruned, so that winning candidates come from a densest subgraph of the graph for the candidates considered.

A system that does not use a coherence measure is DBpedia Spotlight. It chooses the KB-candidate whose context vector in Wikipedia pages is most similar (using cosine similarity) to a textual mention's context vector, weighting tokens in the vectors with a measure of their discriminative power to tease candidates apart, based on how many KB-candidates have that token in their context vector, and how frequently—they call this weight *Term-Frequency – Inverse Candidate Frequency* (Mendes et al., 2011, p. 3).[8]

Confidence scores: Many EL systems provide a confidence score for their outputs. This represents the disambiguation algorithm's estimate of the quality of the outputs proposed. These scores are useful to filter out outputs which are likely to be of low quality. Factors defining this score can be the candidate's prior probability for its mention and the candidate's coherence

scores with respect to other candidates proposed for mentions in the context of that candidate's mention (see the preceding paragraphs). Spotlight does not use coherence scores (see previous paragraph). Its confidence scores rely on candidates' context-similarity score, and on the similarity-score difference between the first and second-ranked candidate, as an indication of the candidate's ambiguity in the context.

NIL clustering: This refers to clustering coreferential mentions for terms that are not part of the target Knowledge Base (KB). Mentions to such terms are called *NIL* mentions (Ji et al., 2014). Not all EL systems carry out this step. The tools we applied in the thesis do not perform NIL clustering. However, this is useful for cases where actors (like people or organizations) that are important in a corpus are not covered by the KB (see p. 158).

1.4 Intrinsic Evaluation in Entity Linking

As the literature has emphasized, evaluation and result reporting practices in EL have been inconsistent, leading to systems' results not being comparable to each other. This spurred metaanalyses of the literature like (Cornolti, 2012) and (Cornolti et al., 2013), and the creation of evaluation tools like Cornolti's BAT Framework,[11] GERBIL (Usbeck et al., 2015),[12] and *neleval* (Hachey et al., 2014)[13]. This section discusses current EL evaluation methods, remaining evaluation challenges, and provides example results for EL systems.

1.4.1 Evaluation Measures

To systematize EL evaluation, the studies just cited defined several evaluation modes or evaluation measures, each with different criteria regarding what counts as a correct result. For each evaluation mode, the metrics relevant for this thesis[14] are *Precision (P)*, *Recall (R)* and $F1$, with standard definitions.[15] The three criteria to define a correct result in current EL systems, and the related evaluation measures, are the following:

1. EL step evaluated: The *Mention Match* measure evaluates to what an extent entity mentions have been correctly identified—this is equivalent to evaluating the NERC step. The measure can take into account entity types or ignore them. The *Annotation Match* measure evaluates both mention detection and mention disambiguation, i.e. it evaluates both the NERC and the NED step.

[11] https://github.com/marcocor/bat-framework
[12] http://gerbil.aksw.org/gerbil/overview
[13] https://github.com/wikilinks/neleval/wiki
[14] Other metrics are used, like clustering metrics to evaluate the quality of grouping *NIL* mentions, i.e. grouping mentions that likely refer to the same term, but where the term is not part of the KB. However, these aspects are not evaluated in this thesis.
[15] $P = \frac{true\ positives}{true\ positives + false\ positives}$; $R = \frac{true\ positives}{true\ positives + false\ negatives}$; $F1 = 2\frac{P \cdot R}{P+R}$

For cases where it is only relevant to assess whether a KB-term has been assigned to a document, without taking into account whether all mentions linkable to that term are correctly linked or not. The *Entity Match* measure reflects this.

2. Match strictness: *Mention Match* and *Annotation Match* have a *weak* and a *strong* version. The strong version requires an exact mention match between system and reference results, whereas the weak one requires an overlap[16] between the mention proposed by the system and the reference mention.

3. Level of aggregation: In *macro-averaged* results, metrics (P, R, $F1$ in our case) are computed for each document and aggregated across documents. This gives equal importance to each document irrespective of the number of evaluation items in each. In *micro-averaged* results, metrics are aggregated at corpus level, which makes longer documents contribute more strongly to the overall result.

1.4.2 Evaluating against Ever-Evolving KBs

Another challenge for the comparability of EL results is due to the fact that the knowledge-bases linked to by EL systems (e.g. DBpedia) undergo constant modifications. New terms are added to the KBs, the preferred variant for a term (i.e. a Wikipedia page title) gets demoted as a *redirect* page forwarding to the preferred title, and so on.

A way to make results more easily comparable is to map reference sets and system results to the same version of the KB. E.g. mapping to the same version of Wikipedia, in order to ignore added or deleted terms or redirection vs. preferred term differences.[17]

Note that, even after mapping results across KB versions, a direct comparison of systems linking to different versions of Wikipedia may be unfair, since the difficulty of the task need not be equivalent across versions, as Milne et al. (2008b) discuss. As pages are added to Wikipedia, the inventory to disambiguate against increases, so the task may be more difficult for newer Wikipedia versions. However, the task may actually be getting easier: In newer versions of Wikipedia, the most-common sense baseline has improved, indicating that common senses are increasingly dominant. Milne et al. (2008b) argue that this can improve the results for systems trained and tested on newer Wikipedia versions.

[16] See Chapter 3 for a definition of *overlap* between mentions.
[17] E.g. with the `fetch_map` function in https://github.com/wikilinks/conll03_nel_eval/blob/master/conll03_nel_eval/fetch_map.py

1.4.3 Reference corpora

Besides evaluation campaign corpora, that are released to participants only (Ji et al., 2014; 2015), several public test-sets exist.[18] Corpora differ along several dimensions: First, in terms of entity types annotated (variety of entity types and presence or absence of common nouns tagged as concepts to link). Second, in terms of text-types and topics (newswire, vs. scientific texts, news or blog format, or even non-canonical varieties like microtext, i.e. tweets). Finally, regarding the KB against which the corpus was annotated (DBpedia, Yago, Freebase). A recent review, providing details about those differences, is (Van Erp et al., 2016). I describe the reference sets used to evaluate this thesis' combination of EL systems on p. 65 (Chapter 3).

Given the variability among reference sets, the literature recommends that an EL system be evaluated against various test sets, and using several evaluation measures (like those in 1.4.1 above).

1.4.4 Example Results

Table 1.1 shows some example results, taken from (Cornolti et al., 2013)[19] for public end-to-end generic domain EL systems, as tested on four public corpora, as an indication of the technology's performance. The evaluation measure is weak annotation match: the KB-term proposed as a disambiguation of each mention must match the reference, but mentions only need to overlap with mention positions in the reference. For a different overview, frequently updated results for several systems, on a wider range of corpora, and more evaluation measures are available on the GERBIL platform.[12] The datasets on the table are described on p. 65 below, since I evaluated my system combination approach on these same datasets.

Some remarks about these example results. If the results seem low, consider the following: According to the descriptions, available from the literature, of the creation of the datasets on Table 1.1, human annotators disagree in about 20% of cases whether an entity annotation should be provided or not, or what the correct KB-term should be (see details on p. 69 in Chapter 3). In other words, this is a difficult task over which humans disagree, so it is not surprising for automatic results to not be very high. Moreover, for us to understand the value of the technology for a given application in DH, these quantitative results need to be complemented with domain-experts' assessment of the results' helpfulness for their intended use of the

[18] The DBpedia Spotlight project hosts some of them at https://github.com/dbpedia-spotlight/evaluation-datasets/tree/master/data, as does the BAT Framework for EL evaluation: https://github.com/marcocor/bat-framework/tree/master/src/main/resources/datasets

[19] Even if the publication is from 2013, I obtained equivalent results in late 2015, using Cornolti's BAT Framework[11] to access the EL systems.

Corpus	AIDA/CoNLL B				IITB				MSNBC				AQUAINT			
System	*t*	P	R	F1	*t*	P	R	F1	*t*	P	R	F1	*t*	P	R	F1
Tagme	0.258	61.4	**55.5**	**58.3**	0.102	45.2	42.0	43.6	0.188	48.5	**55.0**	**51.6**	0.198	**45.9**	56.7	**50.7**
Spotlight	0.086	31.2	40.4	35.2	0.023	46.2	**50.0**	48.0	0.070	31.8	35.2	33.5	0.078	20.1	44.0	27.6
W Miner	0.57	46.9	57.3	51.6	0.219	56.8	48.2	**52.2**	0.758	54.9	39.5	46.0	0.57	37.8	**62.9**	47.2
AIDA	0.0	**74.1**	34.0	46.7	0.0	**65.7**	4.1	7.6	0.0	**74.6**	34.8	47.4	0.0	35.4	15.1	21.2

TABLE 1.1 – Entity linking example results for four public systems and datasets: Weak Annotation Match. Optimal confidence thresholds *(t)*, Micro-averaged Precision, Recall, F1. The best-result for each of P, R, F1 is bold. Data from (Cornolti et al., 2013).

annotations. Low scores may have a negative impact or not. These issues are discussed further on p. 69 in Chapter 3.

As a final remark on these results (or those available on the GERBIL platform): They are obtained with publicly hosted APIs, using default settings at the time the services were accessed. This limits of course the reproducibility of the results. Reproducibility issues in EL were examined in (Hasibi et al., 2016).

1.5 Entity Linking and Related Technologies in Digital Humanities

In this thesis, the focus is on generic domain EL, end-to-end (NERC + NED), targeting many entity and concept types. However, in this section I would like to first discuss domain-specific EL and NERC work that has been important in DH. Most of the domain-specific EL work described in this section uses rule-based disambiguation heuristics, which are unrelated to the generic pipeline described above (section 1.3). However, I wish to discuss such EL work since in my perception it is representative of DH work involving (automatically) linking text to knowledge repositories. As regards the NERC work described in this section, it uses creatively a wide variety of rule-based and statistical approaches, and even if it is unrelated to the generic EL pipeline in section 1.3 and to the work in this thesis generally, I consider it once again representative of what the NERC technology can offer to specific DH needs.

1.5.1 Special applications of EL and NERC in DH

A very active area of EL research in DH is annotating geographical locations in text. On the one hand, geographical knowledge-bases for Humanities interests have been developed (e.g. Pleiades[20] of the Getty Thesaurus[21]). As regards linking tools, several studies have adapted the Edinburgh Geoparser[22] to find and disambiguate locations in historical text. The adaptation usually involves custom rule-based NERC and disambiguation heuristics

[20] http://pleiades.stoa.org/
[21] https://www.getty.edu/research/tools/vocabularies/tgn/
[22] https://www.ltg.ed.ac.uk/software/geoparser/

like fuzzy matching against gazetteers (i.e. placename dictionaries), and exploiting document metadata (dates) or entity-candidate metadata like geographical coordinates, preferring candidates geographically close to other candidates in the same text (Alex et al., 2015; Grover et al., 2010). The Spatial Humanities project[23] has also used such methods (Gregory et al., 2014). A different geoparsing approach is described in Frontini et al. (2016), who use their REDEN entity linking system,[24] exploiting generic KBs rather than gazetteer-based heuristics. Many other projects that identify locations in literary or historical texts exist, like Pelagios (Isaksen et al., 2014) and SyMoGIH (Beretta, 2015)—see Frontini et al. (2016) for a review of other projects, although automatic tagging is often not the focus.[25] Finally, a specificity of location linking in literary texts is the treatment of fictional places, one of the issues which the Literaturatlas project is addressing (Piatti et al., 2009; Reuschel et al., 2011).

Speaking not of EL strictly, but of NERC, the challenges in DH for this technology have been discussed in overviews like (Sporleder, 2010) or (Ehrmann et al., 2016). Besides producing NER systems for historical language varieties (e.g. Borin et al., 2007; Volk et al., 2010) or classical languages (e.g. Erdmann et al., 2016), a type of NERC specific to the Humanities, that the community has produced important work on, is the recognition of fictional characters. First examples of character detection are provided in (Coll Ardanuy et al., 2015; Elson et al., 2010), where string-matching word-based and context-based rules are applied to the output of a standard NERC in order to detect person names and create coreference chains for them. Besides similar heuristics, Bamman et al. (2014) exploited syntactic dependency features and, for character mention clustering, also used pronominal anaphora. Vala et al. (2015) detect, in addition to proper-noun characters, generic noun-phrase characters like *the governor*, thanks to bootstrapping contexts strongly predictive of character-like subjects. Brooke et al. (2016) proposed an entirely different approach: an unsupervised system, without gazetteers, that trains character classifiers with Brown clusters representing entity types.

1.5.2 Generic-domain EL application in DH and its challenges

Going back to general-domain EL, its application in DH has been explored to a lesser extent than domain-specific EL, e.g. in Nanni et al. (2016) or Lauscher et al. (2016). A source of difficulty for applying EL to the variety of corpora dealt with by Humanities and Social Sciences researchers is that the quality of EL systems' results varies according to the corpus. A second

[23] http://www.lancaster.ac.uk/fass/projects/spatialhum.wordpress/
[24] https://github.com/cvbrandoe/REDEN
[25] The ADHO GeoHumanities Special Interest Group is a resource for information on other projects, see http://geohumanities.org/.

System	*Corpus*					
	AIDA/CoNLL B (news, sports)			IITB (web, various topics)		
	P	R	F1	P	R	F1
Spotlight	31.2	40.4	**35.2**	46.2	50.0	**48.0**
TagMe	61.4	55.5	**58.3**	45.2	42.0	**43.6**
Wikipedia Miner	46.9	52.8	49.7	56.8	48.2	52.2
AIDA	63.3	29.1	39.8	65.7	4.1	7.6

TABLE 1.2 – Entity Linking results for different systems on two corpora, highlighting their varying performance across corpora. The measure is Weak Annotation Match (Cornolti et al., 2013—the data come from that same reference)

System	*Correlations*			
	Nbr. *PER*	Nbr. *ORG*	Nbr. *LOC*	Nbr. *OTHER*
Babelfy	**0.769**	-0.376	0.254	**-0.431**
Spotlight	0.217	-0.480	-0.461	0.26
TagMe	0.257	-0.272	-0.194	0.036
Wikipedia Miner	0.082	**-0.679**	-0.632	**0.497**

TABLE 1.3 – Correlations between Entity Linking system performance and named-entity types in corpus (Person, Location, Organization, Miscellaneous). Data from the GERBIL platform[26] on 11/20/2015, for end-to-end EL (mention recognition + linking), measured with Strong Annotation Match (Cornolti et al., 2013)

difficulty is that the generic knowledge bases (KBs) commonly targeted by EL systems (e.g. DBpedia) do not cover important specialized terms in a researcher's domain. This section discusses such challenges.

1.5.2.1 Corpus-dependent performance of EL systems

Table 1.2 shows how the best system on the AIDA/CoNLL corpus (Hoffart et al., 2011) is the worst one on the IITB corpus (Kulkarni et al., 2009), and Table 1.3 shows how the performance of EL systems correlates with corpus characteristics like the types of named entities they contain. In view of this variability in results, and taking into account that DH researchers may need to annotate a wide variety of entity types (besides KB-terms that do not fit a clear entity type) it is not easy for a DH researcher to choose an EL tool.

As outlined in 1.6 below, the thesis describes ways to combine EL outputs so that the combined results improve over each system's individual outputs.

1.5.2.2 Generic vs. domain-specific EL

Regarding the appropriateness of generic-domain EL to very specialized fields, for some applications generic EL is sufficient. Raganato et al. (2016) performed multilingual EL and Word Sense Disambiguation (see p. 17) on the Bible, against an open-domain KB (BabelNet by Navigli et al., 2012),

using for EL the Babelfy system (Moro et al., 2014). The results were satisfactory for their intended use (cross-language retrieval of passages matching a KB-term). However, disambiguation precision was below Babelfy's performance with news corpora.

For other applications, the use of a generic KB for EL will result in important domain-specific terms not being annotated by the system. As Lauscher et al. (2016, section 5.3) argue, missing annotations are more problematic for the scholar than obviously wrong annotations. A domain-expert can easily recognize off-domain disambiguations and remove them from their analysis, while spotting missing tags is more time-consuming.

To help mitigate the problem of domain-specific entity/concept annotations not being covered in the generic knowledge-bases linked to by standard EL systems, some tools have been developed to link to domain-specific resources. An example is REDEN (Brando et al., 2015; Frontini et al., 2015),[27] which disambiguates against domain-specific linked-data sources (optionally combined with generic ones), choosing KB-candidates based on graph centrality measures. In their case study, they annotated person and location names against the Bibliothèque nationale de France (BnF) vocabularies[28] (besides DBpedia), thus adapting the tool to a diachronic corpus of French literary criticism, and improving results over a generic linking tool (DBpedia Spotlight by Daiber et al., 2013; Mendes et al., 2011).

The way domain-specific resources complement generic ones in EL is an important issue and further research on this will be very useful. However, as outlined below, this thesis focuses on combining different generic tools to get an improvement in EL results, rather than mixing specialized and general knowledge sources.

Domain-specific knowledge-bases are not always available. For such situations, methods have been proposed in the literature to cluster coreferential mentions that cannot be linked to a knowledge-base; this is sometimes called *NIL clustering* (Ji et al., 2014). This allows annotating entities or concepts that may be important in a corpus, in spite of not being part of a knowledge base. NIL clustering was not used in this thesis, but it would be useful to apply it for person-names in our PoliInformatics case-study. Approaches to do so (e.g. Coll Ardanuy et al., 2016a; Rao et al., 2010) are outlined there (p. 158).

1.6 Challenges and Implications for our Work

As mentioned above, general-domain Entity Linking (EL) has not commonly been used in Digital Humanities (DH). The thesis assesses this technology's

[27] https://github.com/cvbrandoe/REDEN
[28] http://data.bnf.fr/

contribution to corpus exploration and analysis in DH, through applying it to a historical specialized corpus in philosophy (5.2 in Chapter 5) and a current-day corpus which covers economic policy and social issues (5.3 in Chapter 5). In each case, the technology's applicability and limitations were assessed. In the case of the Bentham corpus, a domain-expert provided feedback on the use of having EL annotations on the corpus.

Another challenge mentioned above was variable results for EL systems depending on the corpus, which makes it difficult for DH researchers to choose an EL system. In the thesis, a method to combine EL systems using a weighted voting scheme is presented, in order to obtain more uniform results across corpora, as explained in Chapter 3.

Chapter 2

Extracting Relational Information in Digital Humanities

2.1 Introduction

Chapter 1 discussed Entity Linking, which helps identify relevant actors and concepts in a corpus. The technology was applied in this thesis to identify important notions in Jeremy Bentham's manuscripts (Chapter 5.2), and to annotate major actors in a corpus documenting the American Financial Crisis of 2007 (Chapter 5.3).

Identifying such elements is not trivial and is a useful step towards gaining an overview of a large corpus with the help of automatic means. However, more information can be obtained automatically. Consider the third case-study in this thesis, a corpus describing international political negotiations about climate change (Chapter 6). For such material, it is relevant to identify not only who is acting in the negotiation, and what concepts are being discussed, but also each negotiating party's position regarding those concepts and other parties: What do they support or oppose? Which parties agree with each other? NLP provides several technologies that allow us to identify how entities are related to each other. This chapter provides a brief overview of these technologies.

As an initial terminology clarification, Chapter 1 described a notion of semantic relatedness within a Knowledge Base (like Wikipedia) based on links between those elements. The relations we're focusing on in this chapter are of other types. For instance, they express attributes for an entity (e.g. a person's job, or a company's location), or the role an entity plays in a situation, e.g. who opposes or supports whom in a conflict.

2.1.1 The Information Extraction field

Besides Named Entity Recognition and Entity Linking (Chapter 1), extracting relation information is part of Information Extraction, which seeks to

turn unstructured text into structured data (Jurafsky et al., 2009, 725ff.). For instance, it can be used to populate a relational database reflecting knowledge about a corpus such as the entities mentioned in it, attributes for those entities, and how the entities are related.

Historically, information extraction approaches have been assessed at a series of challenges or evaluation campaigns, starting with the MUC or Message Understanding Conference (1987–1997).[1] The Automatic Content Extraction program (ACE)[2] followed the MUC between 1999 and 2008. Since 2009, the Knowledge Base Population (KBP) tracks of the Text Analysis Conference (TAC)[3] have performed ACE-type evaluations. The tasks in those campaigns have often been domain-specific, with later tasks covering a broader range of domains. The MUC tasks concentrated on extracting information from military or news reports for the purposes of business intelligence or military intelligence (Grishman et al., 1996). ACE covered more domains and text types than MUC (Cunningham, 2005, p. 674), and KBP has focused on newswire and web data like discussion forums and blogs (Ji et al., 2014, 5.2.3).[4]

Those evaluation campaigns have assessed tasks and technologies like the ones described in this chapter. As just mentioned, the technologies have mostly been developed for contemporary texts, either generic ones like news, or containing specialized vocabulary (e.g. financial or military texts). Information extraction for other specialized fields (biology and medicine) is also a large field in itself, as (Jurafsky et al., 2009, 757ff.) review. For Humanities and Social Science applications, additional difficulties for these technologies can be posed by historical language varieties or complex style, and adapting these tools to make up for this increased challenge is part of the effort in Digital and Computational Humanities.

2.1.2 Technologies reviewed

We start by introducing syntactic and semantic dependency parsing (2.2). These are the basis of some of the techniques to extract relational information discussed later in the chapter. Whereas all technologies in this chapter help finding how entities are related in a text, the term *relation extraction* is usually reserved to a subset of them, discussed in 2.3. In 2.4, event

[1] http://www-nlpir.nist.gov/related_projects/muc/
[2] http://itl.nist.gov/iad/mig/tests/ace/
[3] http://tac.nist.gov/tracks/index.html
[4] Since 2012, the DEFT program (Deep Exploration and Filtering of Text) funded by the US Defense Department agency DARPA has also supported Information Extraction research. The reason to cite this program is that some of the work performed under it has influenced evaluation task definition at KBP. See http://www.darpa.mil/program/deep-exploration-and-filtering-of-text. Aguilar et al. (2014), i.a. discuss DEFT annotation standards compared to those in other evaluation campaigns.

extraction is introduced. This aims at determining which entities participate in a given activity and their roles within it. Examples of application of the above technologies in Digital Humanities (DH) are provided in 2.5. Finally, 2.6 summarizes the chapter and outlines how the thesis relates to the technologies reviewed.

The areas reviewed constitute a vast field each, and the overview cannot be but partial. The focus is on giving an indication of the value of the technologies for DH, rather than on algorithms. Besides, only part of the technologies surveyed here are applied in the thesis (Chapters 4 and 6). I am citing other technologies because they are valuable for DH, they can be used as alternatives to implement the applications I describe in the thesis, and, for somewhat different applications or corpora, they may work better than the choices I made for my use case.

2.2 Syntactic and Semantic Dependency Parsing

This section describes technologies to automatically annotate syntactic and semantic functions in sentences. These technologies are not usually seen as a means in themselves to detect relations for information extraction tasks, i.e. tasks the general goal of which can be assimilated to automatically filling a database with information about entities in texts, their attributes, and relations. However, consider some applications in Digital Humanities, like establishing who is mentioned as playing an active role (vs. being in a passive position) in reports about a conflict by different media outlets, in order to assess possible media bias. Or quantifying subjects and objects for verbs expressing support or opposition in debates, as an indication of actors' attitudes towards an issue. For these applications, the automatic analysis of syntactic functions and semantic roles provides in itself a lot of the information required. In the thesis, these technologies provided the basis for analyzing issues over which actors agree or disagree in climate negotiations. For these reasons, I am reviewing such technologies as a source of relational information.

2.2.1 Syntactic Dependency Parsing

Syntactic dependency parsing determines which words fulfill a given syntactic function (e.g. *subject* or *object*) in a sentence. In a dependency parse tree, edges represent syntactic relations and terminal nodes are the sentence's words. Representing a sentence as a hierarchy of dependency relations was introduced by Tesnière (1959). In his approach, the subject and object depend on the verb, which is their *head*. This has been kept in later approaches, as Covington (1990, 8ff.) reviews. More generally, modifiers depend on the

modified element (e.g. an adjective depends on the noun it modifies, which is its head).

Several annotation schemes for syntactic functions exist. CoNLL dependencies were the basis of a series of evaluation campaigns between 2006 and 2009.[5] Surdeanu et al. (2008, 167ff) provide an overview of the annotation scheme.[6] In Stanford dependencies (De Marneffe et al., 2008), the goal was to provide a formalism that simplifies relation extraction tasks, at the cost of abstracting away from some linguistic detail not essential to those tasks. The more recent Universal Dependencies format (Nivre et al., 2016) aims to ease multilingual parsing and is based on a cross-linguistically applicable annotation format.

2.2.2 Semantic Role Labeling

Semantic dependency parsing, also known as Semantic Role Labeling (SRL), relies on an inventory of semantic frames, which represent sets of semantic roles, fulfilled by *arguments*, and the relation expression that those arguments depend on, called the *predicate*. SRL determines which frames from the inventory are instantiated in a sentence, as well as the lexical sequences corresponding to the predicate and arguments

Two widely used semantic role inventories are PropBank (Palmer et al., 2005) and FrameNet (Fillmore et al., 2003; Baker et al., 1998). For nominal predicates, NomBank is available (Meyers et al., 2004).

In PropBank, semantic roles are defined for individual verb-senses. There are two very general roles, *Arg0* and *Arg1*. *Arg0* corresponds to a prototypical agent, and *Arg1* to a prototypical patient (Palmer et al., 2005, p. 75). In essence, the prototypical agent causes an event or state-change in another participant, and the prototypical patient undergoes a change of state or is causally affected by another participant (see Dowty (1991, p. 572) for more details). Other arguments are numbered sequentially (*Arg2*, *Arg3* etc.). The meaning of these other arguments needs not generalize across predicates. There are also modifier roles, for adjuncts, like *ARGM-TMP* for time expressions. The PropBank annotation scheme was conceived in order to facilitate training a statistical SRL parser based on manual PropBank annotations. As such, it is based on the Penn treebank, a large syntactically annotated corpus used to train syntactic parsers (Marcus et al., 1993). PropBank semantic frame annotations were added to the treebank's nodes (Palmer et al., 2005,

[5] http://www.conll.org/previous-tasks

[6] The earliest CoNLL task for syntactic dependencies is Buchholz et al. (2006). I am citing the CoNLL 2008 overview paper by Surdeanu et al. because it gives a clear description of the set of functions annotated.

88ff.). The original predicates in PropBank are verbs, but nouns were added later.

As regards FrameNet, a specificity in this knowledge base is that its frames group together semantically similar predicates, i.e. those that can be described as taking similar arguments. This type of grouping is not directly available in PropBank. For instance, the *Convey_importance* frame applies to the following verbs: *emphasize*, *stress*, *underline*, *underscore*.[7] Arguments are defined at frame level, not at verb-sense level as in PropBank. The verbs just mentioned take as arguments a *Speaker* or *Medium* role and a *Message* role. These roles also appear in other communication-related frames, e.g. *Statement*.[8] Noun predicates are available in FrameNet from its outset; e.g. the *Statement* frame lists several noun predicates.[8]

Regarding NomBank (Meyers et al., 2004), it provides argument annotations for the most frequent nouns in the Penn treebank. It uses PropBank-like arguments and is designed for compatibility with PropBank. For instance, the PropBank Arg0 and Arg1 arguments of *appoint* match the NomBank Arg0 and Arg1 arguments for *appointment*.

There have been initiatives to integrate the knowledge from most of the above repositories. For instance, SemLink (Palmer, 2009) and the more recent Predicate Matrix (López de Lacalle et al., 2014, 2016), which integrate information from PropBank, FrameNet and some other knowledge bases; the 2016 version of the Predicate Matrix integrates NomBank too. The advantage of integrating the repositories lies in a richer corpus annotation. The SRL tool used in this thesis (Chapters 4 and 6) provides Predicate Matrix links.

2.2.3 Parser examples

Many parsers for syntactic and semantic dependencies exist, and a thorough review is out of scope here, as is a review of parsing algorithms. As a practical remark, three examples of widely used statistical parsers, retrainable but with pre-trained models in several languages, are the following. First, Mate Tools (Bohnet, 2010; Björkelund et al., 2010), with pre-trained models for CoNLL-format syntactic dependencies in several languages, and with models for SRL to PropBank/NomBank.[9] Second, the Stanford parser (Manning et al., 2014 i. a.) for Stanford dependencies.[10] Third, Malt parser, where the syntactic dependency format (CoNLL or Stanford) for pre-trained models

[7] https://framenet2.icsi.berkeley.edu/fnReports/data/frame/Convey_importance.xml
[8] https://framenet2.icsi.berkeley.edu/fnReports/data/frame/Statement.xml
[9] https://code.google.com/archive/p/mate-tools/wikis/ParserAndModels.wiki
[10] http://nlp.stanford.edu/software/stanford-dependencies.shtml

varies depending on the language.[11] For Universal Dependencies parsers, see Straka et al. (2016).[12] Other parsers are also in wide use, Choi et al. (2015) compared some of them (e.g. ClearNLP and TurboParser).

For SRL with FrameNet, an evaluation task was organized by Baker et al. (2007) at the SemEval workshop. Current FrameNet parsers I am aware of are SEMAFOR (Das et al., 2010),[13] the Lund University FrameNet parser (Johansson et al., 2007),[14] and Shalmaneser (Erk et al., 2006).[15]

2.2.4 Parser evaluation and example results

As mentioned above (p. 32), several evaluation campaigns have taken place, besides individual studies comparing parsers (p. 35).

Details differ per campaign, but all in all the basic measures in dependency parsing evaluation are *Labeled Attachment Score (LAS)* and *Unlabeled Attachment Scores (UAS)*. In LAS, a correct result requires correctly identifying the head and dependent, besides the type of relation between them. In UAS, the type of relation is not considered.

In syntactic dependencies, for quantifying LAS and UAS an accuracy is computed, i.e. the number of correctly annotated tokens divided by the total number of tokens in the test-set. For SRL, precision, recall and F1[16] over the predicate and arguments are computed.

As regards the scores attained by parsers, scores reported independently of an application context need not indicate the usefulness of the technology for a researcher in DH. An evaluation of error types in the researcher's corpus and their impact for research results should be assessed for a concrete application. I am providing some example results only as an indication of the performance of these technologies, for news texts in most cases. For other genres performance may decrease.

The CoNLL 2008 and 2009 tasks involved joint syntactic and semantic dependency parsing. Their results improve over results obtained at earlier CoNLL tasks for each of those technologies separately.[5] The genre of the test corpora was news. For such material, at the 2009 task (Hajič et al., 2009) syntactic LAS accuracy was 89.88% for the best system, with 12 out of 13 participants scoring above 80.00%. Still for English, the best system's semantic LAS F1

[11] http://www.maltparser.org/
[12] https://ufal.mff.cuni.cz/udpipe
[13] https://github.com/Noahs-ARK/semafor
[14] http://nlp.cs.lth.se/software/semantic-parsing-framenet-frames/ Note that Lund also created an SRL system for PropBank/NomBank (Johansson et al., 2008), the latest versions of which are integrated in Mate Tools.
[15] http://www.coli.uni-saarland.de/projects/salsa/shal/
[16] These terms are defined in footnote 15 on p. 20. Note that F1 values can be reported as a 0 to 1 range or as a 0 to 100 range, without a difference in meaning.

was 86.15, with 9 out of 20 participants scoring above 80. Across languages, for 12 out of the 13 participants syntactic LAS accuracy was between 72.54% and 85.77%, and the average semantic LAS F1 for 11 out of the 20 participants was between 70.31 and 80.47.[17] The English corpus was taken from the Penn Treebank III (Marcus et al., 1993).

Note that at the CoNLL 2009 task, for semantic dependencies, the predicates had already been identified in the test-set (Hajič et al., 2009, p. 3). It was not required to identify the tokens that correspond to predicates. In the CoNLL 2008 task (Surdeanu et al., 2008) and in Björkelund et al. (2010), we find results on the same corpus as the 2009 task, but with automatic predicate identification performed by the systems. Under these conditions, the best semantic LAS F1 is approx. 81. This is approx. 5 points down from an evaluation where predicate tokens are already given to the system. The evaluation more relevant for this thesis is when predicates need to be automatically identified, as this corresponds to the use case in the thesis, which requires automatically analyzing unannotated text.

A newer evaluation for syntactic dependency parsers was performed by Choi et al. (2015). They used a broader domain range than CoNLL, as they included web texts and phone conversations besides news. LAS accuracy ranged between 85.93% and 90.09%, depending on the exact testing conditions (Choi et al., 2015, p. 391). The reference set was taken from the OntoNotes 5 corpus (Weischedel et al., 2011; Pradhan et al., 2013).

An event to look forward to regarding improvements in syntactic dependency parsing is the upcoming 2017 CoNLL task, which focuses on Universal Dependencies. Our lab is participating at this task.[18]

2.3 Relation Extraction

As mentioned above, whereas all technologies in this chapter can be used to identify relations between entities, *relation extraction* (RE), used as a technical term, usually refers to a subset of the approaches used to that end. A brief overview of RE in that sense is provided in following. The first RE approaches were domain-specific, and are sometimes termed *traditional relation extraction*. A later approach, starting in the mid 2000s and called *open relation extraction*, seeks to provide open domain systems. Both approaches co-exist.

As a simple illustration of what RE does, consider the following example. In (1), we have the second sentence in the Penn Treebank (Marcus et al.,

[17] http://ufal.mff.cuni.cz/conll2009-st/results/results.php#cjsynt
http://ufal.mff.cuni.cz/conll2009-st/results/results.php#csemf1
[18] http://universaldependencies.org/conll17/

1993). In (2), we see entities annotated in it with Named Entity Recognition and Classification (p. 17):

(1) ***Mr. Vinken*** *is* ***chairman*** *of* ***Elsevier N.V.****, the Dutch publishing group.*

(2) *[**Mr. Vinken*** $_{PER}$*] is [**chairman*** $_{JobTitle}$*] of [**Elsevier N.V.*** $_{ORG}$*], the Dutch publishing group.*

It could be useful to extract relations among the entities in this sentence. For instance, if we're following news about publishing companies, for business intelligence, we would be interested in determining that a person (Mr. Vinken) is related to the Elsevier N.V. organization. And that Mr. Vinken's job title at Elsevier is *chairman*. Relation extraction would identify such information.

2.3.1 Traditional Relation Extraction

The first relation extraction systems were domain-specific. Poibeau (2002, 202ff.) reviews participant systems at the Message Understanding Conferences (MUC) in the 1990s. Automata-based methods were widely used, relying on a dictionary of domain-relevant entities and relation triggers, and (hand-crafted) patterns formalizing the relations between them. One of the best performing systems was FASTUS (Hobbs et al., 1993), where a cascade of finite state automata is used to approximate the syntactic analysis of a sentence (which allows skipping tokens unlikely to be relation participants) and to perform pattern matching. Towards the end of the MUC conferences, researchers started to systematically examine the application of machine learning for semi-automatic relation pattern acquisition.

Relation extraction systems have been trained using supervised machine learning, based on hand-tagged examples. Jurafsky et al. (2009, 735ff) provide a summary of the machine learning models used in the literature (e.g. naive Bayes and Support Vector Machines) and the types of features exploited. Features include lexical features in the named entities to relate (like their form and type), surface features from the span of tokens between the entities (e.g. word-forms or number of tokens), and features from a syntactic parse of the sentence.

Relation Extraction methods relying on hand-crafted dictionaries and rules or learned on the basis of hand-tagged examples are costly to develop given the manual effort required. For that reason, methods for reducing the required amount of manual tagging have been developed, which Grishman (2015, 12ff.) and Jurafsky et al. (2009, 738ff) summarize. One such approach is **bootstrapping**, i.e. creating a larger training set on the basis of a small number of hand-tagged examples, called *seeds*. Reliable entity tuples from the seed set, if found within the same sentence in a large corpus, can be

used to find new relation expressions not available in the seed set. Similarly, sentences containing reliable relation expressions in the seed set can be used to find training examples with new entity-tuples in a large corpus. Relevant work on how to define "reliable" tuples is (Agichtein et al., 2000), who created a system called *Snowball* applying bootstrapping methods. Also, Riloff et al. (1999), who worked on assessing the reliability for bootstrapping of patterns indicating relations. Another approach for learning relation expressions is **distant supervision**: A large database of previously available tuples expressing the relation (rather than a small seed-set) is searched in a large corpus, to find new relation expressions (Grishman, 2015).

For an overview of RE tasks at the evaluation campaigns mentioned on p. 29 (MUC, ACE and KBP), see p. 39, Diesner (2012, 33ff.) or Ezzat (2014, 15ff).

2.3.2 Open Relation Extraction

The relation extraction (RE) systems described above (2.3.1) require hand-tagged data for each relation type to annotate, and it is costly to train them for a wide range of relations, even if applying automatic methods for training set creation, as outlined on p. 36, can help decrease that effort. Wishing to extend RE to a large variety of domains, and thinking of computationally efficient methods applicable to web-scale corpora, Open Relation Extraction (ORE), or Open Information Extraction (OIE) was proposed.[19] ORE methods learn generic, cross-domain relation patterns, in a self-supervised manner, requiring no hand-tagged data or a small amount of it. Most of the systems I survey were created at the University of Washington's KnowItAll project since the mid 2000s.[20] This project has produced the main ORE tools, but other systems exist.

The first such system known to me is *TextRunner*, by Banko et al. (2007). The system trains itself by obtaining tuples from a corpus of several thousand sentences tagged for syntactic dependencies. From this corpus, it extracts tuples containing noun phrases as the relation arguments, and other expressions that are likely to be relation phrases as the predicate, according to syntactic constraints. The system then learns a classifier to identify relations based on these initial tuples, using shallow features like part-of-speech (POS) tags and their sequences. Such shallow features were used in order to ensure that the system scales to web-size corpora, as they are not costly to compute. The relation extractor thus learned recognizes a variety of relations, not restricted to a pre-defined set, in other corpora. For instance, ⟨*proper noun, acquired, proper noun*⟩, ⟨*proper noun, is based in, proper noun*⟩, ⟨*proper noun,*

[19] *ORE* and *OIE* can be used synonymously.
[20] That project's systems are available at https://knowitall.github.io/openie/

worked with, proper noun⟩. These three examples show that the relations extracted are varied, describing for instance a business transaction via *acquired*, a locative relation via *is based in*, or employment information like *worked with*.

The next ORE system by the same team, *ReVerb* by Etzioni et al. (2011), learns an argument extractor rather than applying TextRunner's noun-phrase based heuristic, to overcome errors in TextRunner's argument identification. ReVerb also enforces additional shallow constraints based on part-of-speech sequences to eliminate some uninterpretable relations. Another improvement in *ReVerb* is that it handles light verb constructions, where the lexical meaning of the verb is carried by a noun following the verb, rather than by the verb itself. For instance, it is able to extract *took place in* as the predicate, instead of only *took*, which would be uninformative as a predicate in the same context.

The next ORE system, *OLLIE* by Mausam et al. (2012) provides improved results over the two preceding ones. First, whereas the two previous systems used shallow features (largely based on part-of-speech) to learn open relation patterns, in OLLIE dependency paths are part of the patterns. The method still scales to large corpora thanks to using a fast parser, *Malt parser* by Nivre et al. (2004). Patterns can be purely syntactic, have lexical constraints, or have semantic constraints, e.g. "the argument must be a person", see Mausam et al. (2012, section 3.2). Using dependency features rather than shallow part-of-speech sequence features improves OLLIE's recall with respect to predecessors. A second improvement in OLLIE is that it extracts relations mediated by nouns or adjectives, not only by verbs like the two previous systems. Finally, OLLIE analyzes the context around the tuples extracted, and determines if there is an *attribution*, i.e. when the relation is not asserted as factual, but attributed to a source. It also annotates *clausal modifiers*, when the relation is presented as the hypothetical outcome of a condition, instead of asserted as true.

One feature in common to all of the systems mentioned is that they provide a confidence score for each tuple extracted. This can be used to filter weaker results.

Two new ORE systems were introduced by Mesquita et al. in 2013 and 2015. Mesquita et al. (2013) presents *Exemplar*.[21] This is a rule-based ORE system, featuring very generic rules to identify relation predicates and their arguments; the predicates can be verbal or nominal. Its rules are based on part-of-speech tags, and syntactic dependency relations between candidate relation-triggers and candidate arguments. Parsing is done with either Malt

[21] https://github.com/U-Alberta/exemplar/

(*Exemplar[M]*) or Stanford parser (*Exemplar[S]*).[22] This system improves over the ones cited above, as assessed on a heterogeneous corpus (p. 41).

Mesquita (2015) presents the *Efficiens* system. He notes that deep ORE, based on syntactic dependencies or SRL, improves over shallow ORE (based on part-of-speech patterns), but at the cost of a considerable increase in processing time. To find a balance between result quality and processing time, the Efficiens system determines, on a sentence-per-sentence basis, to what an extent relation extraction for the sentence would improve if using dependencies or SRL instead of a shallow analysis. The user provides parameters according to their time-constraints, indicating the proportion of corpus sentences that he or she can accept to be processed with the deeper, slower methods. Respecting the proportions chosen by the user, the system decides which sentences to apply each method to, based on expected improvement for the sentence if using deeper methods. The expected improvement is based on a variety of features indicating complexity of the relation(s) in the sentence.

2.3.3 Evaluation in relation extraction and example results

As I keep mentioning when reporting evaluation results from standard tasks, such results can only give an indication of how the technology performs. Relation extraction tasks as defined in evaluation campaigns need not correspond well to the challenges posed by a specific application in Digital Humanities (DH), which may pose additional difficulties, or, on the contrary, be more constrained, limited to a small range of domain-specific expressions, and easier to solve than some of the standard tasks described below.

Evaluation in traditional relation extraction is discussed first (2.3.3.1), followed by evaluation in open relation extraction (2.3.3.2).

2.3.3.1 Traditional relation extraction evaluation

Evaluation measures have depended on the campaign. Sometimes argument identification and predicate identification are evaluated separately. When several relation types are assessed, detection can be evaluated separately from classification.

The 7th Message Understanding Conference MUC 7 had a relation extraction task called *Template Relation*, where three relation types were evaluated: *employee of*, *location of*, *product of* (Chinchor et al., 1998). Best systems reached an F1 of 75 (Chinchor, 1998). The Automatic Content Extraction (ACE) campaign, which followed the MUC, asked for more complex and more varied relations. The list can be found in the task's evaluation plan by NIST-ACE

[22] These syntactic dependency parsers were introduced on p. 33

(2005, p. 3) and includes finding an artifact's users or inventors, people's citizenship or residence, organizations' founders or members, family relations, and locations. F1 reported by Wang et al. (2006) is ca. 71.5 for relation detection and 56.78 for fine-grained classification. A coarse-grained classification yielded an F1 of 65.2.[23]

As regards the TAC-KBP campaign, it had a relation extraction task until 2014, called *Slot Filling* (Surdeanu et al., 2014).[24] This task provides participants with a series of person names and organizations, and a large document collection to extract information from (> 1 million news stories and forum posts). Systems have to discover entity attributes present in the document collection. For instance, a person's cause of death or criminal charges against that person. For organizations, their date of foundation or their headquarters have to be found out. The test set contains 50 entities of each type. The best systems achieved an F1 of ca. 35 (median F1 of 19.8).

If these values seem low, note that many of the entities are ambiguous across several referents, which poses an additional difficulty, i.e. the task benefits from applying entity disambiguation, a not trivial process (p. 17). Also, solving some of the task items involves identifying complex coreferences, and some of the relations to extract are implicit. Humans achieved an F1 of approx. 70 at the same task (Surdeanu et al., 2014, Table 6), which indicates that the task is not trivial for humans either. This task is not directly comparable to ACE, since, in KBP, the evaluation is not mention-based but based on filling a Knowledge Base (KB). In other words, the KBP task does not consist in annotating a relation for every mention of an entity in a test-set, as was done at ACE. Rather, the KBP task requires annotating a relation once for each entity in a KB, if evidence for the relation is found in a large document collection. See Aguilar et al. (2014) for a detailed comparison of ACE and KBP relations.

It is hard to assess the relevance of these evaluation results for DH applications, other than to be aware of the factors that can pose difficulties, and to be aware that, as a way to compare approaches to a given corpus, a corpus-specific evaluation may be advisable, rather than relying exclusively on the results of evaluation campaigns. Surdeanu et al. (2014) provide detailed discussion of the approaches that worked fine at the last KBP Slot Filling

[23] ACE reported results at `http://www.itl.nist.gov/iad/mig/tests/ace/2005/doc/ace05eval_official_results_20060110.html`. The evaluation scheme involves a "model of the application value of system output" instead of F-scores, as explained in (NIST-ACE, 2005, p. 4). System errors have an associated weighted cost, which is deduced from the maximum possible score of 100. The study cited above by Wang et al. (2006) does not use the ACE scores, but the more usual F1 metric.

[24] From 2015 onwards, *Slot Filling* is part of a larger task called *Cold Start*, where entities have to be discovered prior to discovering the relations between them.

tasks, and the most challenging aspects in the task's items, quantifying error sources.

2.3.3.2 Open relation extraction evaluation

Banko et al. (2008) compared ORE performed with *TextRunner* (p. 37) to traditional RE performed with a statistical system trained via supervised learning, evaluating both on a small number of relation types. They report that, for a small number of relations, the traditional system achieves much higher recall than the open one, but it takes hundreds of hand-tagged examples for the traditional system to match the open system's precision. They also combine both systems creating a hybrid extractor whose precision is higher than the traditional one, with comparable recall (Banko et al., 2008, section 5.2).

Mausam et al. (2012) compare their ORE system OLLIE (p. 38) with Semantic Role Labeling (SRL) as a source of open relation extraction, using the Lund PropBank/NomBank parser for SRL (Johansson et al., 2008).[25] They find that neither tool dominates the other, with the union of both parsers' results improving over each parser's individual ones. Similarly, Christensen et al. (2011) had combined the TextRunner ORE system (p. 37) with Lund SRL, also obtaining improvements in the combination.

The clearest evaluations of ORE systems I have found are in Mesquita et al. (2013, 2015). To overcome comparability problems across previous studies, they created their own benchmark of 1,100 sentences, which includes informal, imperfectly written web text (the Web-500 corpus),[26] news stories, and Penn Treebank sentences. They annotated the corpus for relations in a domain-independent manner, and compared several ORE systems (including the ones mentioned above), besides Lund SRL. Their results for the systems above are reproduced in Table 2.1.

Looking at Table 2.1, the *Exemplar* system in Mesquita et al. (2013) obtains better result quality than the rest. Similarly to arguments I made above (pp. 39, 40), the F1 figures on that table themselves need not be meaningful to evaluate a tool for a specific DH application; the task as defined in that paper may not reflect the specific application need, and in that case it would provide more telling results to evaluate the technologies under conditions comparable to the intended application. For instance, in this thesis, SRL was used as the basis of extracting statements in a negotiation corpus (see p. 89), and F1 was 0.69 (in a 0 to 1 range), higher than SRL's results on the ORE test-sets on Table 2.1.

[25] Lund University also created a FrameNet SRL parser, see p. 33. But we're talking about their PropBank/NomBank one here.

[26] Web-500 sentences are commonly used in ORE evaluation since the papers describing the TextRunner system Banko et al., 2007

Corpus	NYT-500				WEB-500				PENN-100			
System	Time (s)	P	R	F1	Time (s)	P	R	F1	Time (s)	P	R	F1
ReVerb	**0.02**	0.70	0.11	0.18	**0.01**	0.92	0.29	0.44	**0.02**	0.78	0.14	0.23
OLLIE	0.05	0.62	0.27	0.38	0.04	0.81	0.29	0.43	0.14	0.81	0.43	0.56
Exemplar[M]	0.08	0.70	**0.39**	0.50	0.06	0.95	0.44	0.61	0.16	0.83	0.49	**0.62**
Exemplar[S]	1.03	0.73	**0.39**	**0.51**	0.47	0.96	**0.46**	**0.62**	0.62	0.79	0.51	**0.62**
Lund SRL	11.40	**0.78**	0.24	0.37	2.69	0.91	0.37	0.52	5.21	0.86	0.35	0.50

TABLE 2.1 – Comparison of Open Relation Extraction (ORE) results for the systems introduced in 2.3.2 on a corpus of 1,100 sentences. Time per sentence in seconds, Precision, Recall and F1 (as a 0 to 1 range). *Lund* is an SRL system, that is used here as a source of open, cross-domain, relations. The *NYT-500* corpus contains 500 news sentences. *WEB-500*: Web corpus. *PENN-100*: 100 sentences from the Penn Treebank. Best results in bold. Data and corpus from Mesquita et al. (2013).

Another factor that ORE evaluations focus on is processing time. Shallow, part-of-speech based systems like TextRunner or ReVerb are the fastest, and full SRL ones like Lund are the slowest. Systems that rely on dependency parsing for relation extraction lie in the middle (OLLIE, Exemplar). Depending on corpus size, this would be a practical aspect to consider in developing DH applications. Table 2.1 also lists seconds per sentence for each tool.

2.3.4 Traditional vs. open relation extraction for DH

Open Relation Extraction can be useful when we don't know beforehand what relations are important in a corpus. It could be used to identify important types of relations in a corpus, and then, if the results obtained that way are not sufficient, a specific "traditional RE" extractor could be developed to obtain optimal results for those types of relations.

In the study by Banko et al. (2008), ORE provided good precision with much less recall than relation-specific extractors. However, they only evaluated a small number of relation types. Later studies by Mesquita et al. (2013, 2015) show higher F1 values for newer ORE extractors. But the comparability between the studies by Banko et al. and Mesquita et al. is limited, since they did not evaluate on the same types of relations. More generally, as argued above (p. 41), F1 values reported in a study need not be indicative of the methods' performance at a DH-relevant task.

A relation-specific extractor may require tagging data by hand for supervised learning, which may be time consuming. At the same time, if we are only interested in a small set of relations, ORE results would need to be post-processed to identify the relevant subset of its results, which also involves some effort.

As I mentioned above, the choice will depend on the specific application, and evaluating on a corpus that reflects the application's challenges could be more informative than results reported in the studies just reviewed. These can be relevant insofar as their evaluation method is pertinent for our use case.

2.4 Event Extraction

This section briefly discusses Event Extraction (EE), a task similar to Relation Extraction.

2.4.1 Task description

For this task, an event is "something that happens", borrowing from the definition in the Automatic Content Extraction (ACE) campaign annotation guidelines for 2005 (LDC, 2005, p. 5). It is a "specific occurrence involving participants". Events frequently describe a change of state, and most events can be assigned a time and location of occurrence (LDC, 2005, pp. 5, 49). A simple example for Event Extraction can be shown based on the following Penn Treebank sentence:

(3) *In 1979, Hearst hired editor James Bellows, who brightened the editorial product considerably.*

In (3), the expression indicating an event is the verb *hired*. The participants are Hearst (a publishing house), James Bellows, and the position of editor. Using the ACE 2005 annotation scheme (LDC, 2005, p. 65), this is a *Personnel/Start-Position* event, and its participants can be tagged as in (4).

(4) *[In 1979 $_{Time}$], [Hearst $_{Entity}$] [hired $_{trigger}$] [editor $_{Position}$] [James Bellows $_{Person}$], who brightened the editorial product considerably.*

The task is related to Relation Extraction (RE). In RE, an entity's attributes are extracted. In (3), *editor* would be extracted as the *job title* for James Bellows, and Hearst would be his *employer*. In Event Extraction, the emphasis is on identifying the event whereby entities enter a relation, rather than the entities' attributes only. Event Extraction overlaps with Open Relation Extraction (ORE), in the sense that ORE systems can output tuples that constitute events, like some of the examples mentioned above (p. 37), e.g. ⟨*proper noun*, *acquired*, *proper noun*⟩.

Event Extraction has been present in information extraction campaigns since the MUC conference. At MUC there was a task called *Scenario Template* (Grishman et al., 1996, p. 468), where event indicators and participants had to be identified in a corpus. In MUC-3, the scenario templates covered topics like attacks or terrorism. In MUC-5, corporate events like the creation of joint ventures were covered.

The ACE campaign had an event task in 2005. It focused on several broad event categories, divided into several subtypes. For instance, the LIFE category contains event subtypes *Be-Born*, *Marry*, *Divorce*, *Injure*, *Die*, and the CONFLICT category contains *Attack* and *Demonstrate*. The PERSONNEL

category, mentioned above, describes "human resources events", like *Start-Position*, *End-Position* etc. The JUSTICE category contains event subtypes like *Arrest-Jail*, *Sue*, *Pardon* among others. The task's evaluation plan (NIST-ACE, 2005, p. 3) shows the complete typology, which the task annotation guidelines (LDC, 2005) describe in detail.

ACE events are also annotated for modality, i.e. whether the event is asserted, or not. Events that are not asserted can be expressed as beliefs, hypotheses, requests, threats, promises etc. (LDC, 2005, p. 20). Later EE tasks, like those in the KBP Event track, which started in 2014, have adopted an ACE-like event typology and also assess modality.[27]

2.4.2 Approaches

Event Extraction (EE) involves identifying an event trigger, which indicates the presence of the event, as well as the event's participants, according to a typology, e.g. the one in ACE 2005 mentioned in 2.4.1. Both rule-based and machine learning approaches have been developed for EE.

A study by Chen et al. (2009) gives a detailed account of features relevant for EE using machine learning. Their discussion refers to training with the ACE event corpus (Walker et al., 2005) in Chinese, but their features (below) are not presented as language-specific.

For **trigger identification**, lexical, syntactic, semantic and entity neighbour features are used. Lexical features include word, part-of-speech (POS) tags, and their bigrams. Syntactic features consist in paths involving the candidate-trigger in the parse tree and the nature of the nodes connected to the candidate. Semantic features refer to presence of the candidate in predicate dictionaries. A study by Ahn (2006) also discusses exploiting information from the WordNet lexical database (Fellbaum, 1998) for trigger-identification features. Entity neighbours are considered both at token level and in terms of the parse tree.

As regards **arguments**, they need to be identified and classified as the correct role within the event. The features in Chen et al. (2009) presuppose that the trigger has already been identified, and include lexical features of the trigger and candidate arguments and their neighbours, besides entity type. Syntactic features relating the trigger and candidate arguments are also considered.

Other descriptions of features for learning an event extractor (including the **modality** attributes) can be found in Ahn (2006) and Jurafsky et al. (2009, p. 749).

[27] From KBP 2015 onwards (Ellis et al., 2015), the annotation scheme draws from the DARPA DEFT program (p. 30) *ERE* guidelines for *E*ntities, *R*elations, and *E*vents (Song et al., 2015).

2.4.3 Evaluation and example results

In EE, aspects evaluated are event type and subtype, trigger detection, argument detection and classification, and event modality (i.e. asserted or not). In evaluation campaigns, different subsets of those aspects have been assessed. For instance, KBP has two event-related tasks, called *Event Nugget Detection* and *Event Argument Linking*. In Event Nugget Detection (Mitamura et al., 2015), the span of text for the event and its type or subtype need to be detected, besides event modality. In Event Argument Linking, the arguments need to be detected and classified (Ellis et al., 2015).

Event coreference is also an important EE task that has regularly been evaluated in campaigns. It consists in identifying the different mentions (i.e. spans of text) referring to the same event.

KBP results for event nuggets (i.e. identifying the text-span for the event, its trigger and type, and event modality) have reached F1 scores between ca. 45 for all attributes and 65 for identification only (Mitamura et al., 2015). KBP results for argument identification have been lower, as reported in system descriptions by task participants.[28] My interpretation is that these lower scores for argument identification are due to a task definition with very detailed requirements, which increase difficulty.

Regarding results on the ACE corpus for English (Walker et al., 2005), event trigger identification and classification reaches an F1 of ca. 65,[29] and argument identification and classification attains an F1 of up to 45 (Li et al., 2013; 2014).

As I pointed out above when discussing public evaluation results for other technologies, these results refer to specific test corpora, which use a varied, but limited, set of event types. For a specific DH application, results may differ. The papers just cited discuss some of the difficulties faced by event extraction systems, which may be more informative for choosing how to implement an event extractor than the evaluation scores at the tasks above.

2.5 Applications in Digital Humanities

The chapter so far has reviewed a broad range of technologies, each of which is a large area in itself. We now turn our attention to how these technologies have been used in Digital Humanities. Given the breadth of the material, a thorough review would be unrealistic. We attempt to provide a broad picture of how these technologies are used in DH in general based on several examples.

[28] https://tac.nist.gov/publications/2015/results.html

[29] These studies do not use ACE's own scoring method (footnote 23), but the more usual F1 (see footnote 15 on p. 20).

2.5.1 Syntactic parsing applications

I am covering two aspects: First, applications of syntactic parsing in contemporary languages, with pre-existing models for parsers like the ones cited on p. 33. Second, the creation of parser models for historical languages.

2.5.1.1 Contemporary languages

Van Atteveldt (2017) has exploited syntactic dependency parsing with the Stanford CoreNLP parser (p. 33) to extract clauses, defined as ⟨*subject*, *predicate*⟩ tuples, as well as the sources the clauses are attributed to. This was used to examine the portrayal of Israeli vs. Palestinian actors in war, in US vs. Chinese media. For instance, who is presented as the aggressor or victim? His earlier work (2008) had also used syntactic dependencies to analyze the portrayal of political actors in Dutch news, depending on ideological affinities between media outlets and politicians, among other factors.

Schrodt and collaborators have developed the PETRARCH system (Schrodt et al., 2014),[30] which also uses the Stanford parser (and the related CoreNLP pipeline) to identify political events with the help of dictionaries of actors and pattern-sets. A universal dependencies (p. 32) version is also under development.[31] This work is an evolution of their earlier TABARI system, which used dictionaries and surface patterns, without full syntactic parsing (Schrodt, 2014), and which has been applied to many automatic event coding projects, e.g. for analyzing news on international conflicts (Schrodt et al., 2013).

Hulden (2016) used syntactic analysis to examine academic articles' portrayal of agency in labour relations. She analyzed subjects and objects representing different groups of actors, and the verbs attaching to them. She found agency patterns that illustrate power relations as portrayed in the corpus. To extract subjects and objects, she applied the Tregex tool (Levy et al., 2006) on the output of the Stanford parser. This tool allows identifying patterns in syntactic trees, using regular-expression-like rules that describe paths and nodes in the tree.

The applications just discussed have a similar goal to the work on international climate negotiations in this thesis (Chapter 6), where actors in the negotiations and their statements were identified. However, I used Semantic Role Labeling parses (p. 32) besides syntactic dependencies.

[30] https://github.com/openeventdata/petrarch2
[31] https://github.com/openeventdata/UniversalPetrarch

2.5.1.2 Historical languages

Developing parsers involves the creation of syntactically annotated corpora (treebanks) so that the parsers can be trained, and the development of parser models themselves based on the training sets. Examples of treebank development for historical varieties are, for Latin, the Latin Dependency Treebank by Bamman et al. (2007) and the Index Thomisticus Treebank by McGillivray et al. (2009). For other romance languages, the SRCMF project[32] (Prévost et al., 2013) created dependency treebanks for medieval French. These have then been used to train parsing models (Stein, 2014; 2016 and Guibon et al., 2014, 2015a, 2015b). Scrivner et al. (2012) developed an Old Occitan annotated corpus and parser via cross-linguistic transfer from models in related languages which already had available treebanks, like Catalan. Moving to English, Taylor (2007) and Taylor et al. (1994) created tagged corpora for Old and Middle English.[33] For other historical Indo-European languages, the PROIEL project has created several treebanks (Haug et al., 2008).[34] As for other language families and even more demanding cases, given a yet more pronounced scarcity of pre-existing linguistic resources, Inglese (2015) describes work towards creating a Hittite treebank using Universal Dependencies (p. 32), discussing philological issues as well as linguistic annotation challenges.

2.5.2 Relation extraction applications

Diesner (2012, 19ff.) has studied in detail the application of relation extraction to network creation. She analyzed the impact of automatic annotation errors in network creation, as well as the differences between networks created with data obtained by NLP methods vs. created by experts. The application corpora consist in news (covering political conflict), scientific funding proposals (relevant for analyzing policy), besides the Enron corpus, with e-mails at Enron before its bankruptcy.

Several studies have used relation extraction, often in order to automatically create ontologies from a body of text, i.e. sets of related concepts to describe a knowledge domain. The ontologies are then used to enrich document metadata, which can help provide better navigation of the document collection. Van Erp et al. (2009) extracted relations from Wikipedia text in order to build an ontology relevant for describing cultural heritage collections. Sanabila et al. (2014) performed relation extraction to find attributes of characters in Indonesian mythology, automatically inducing an ontology.

[32] The acronym stands for *Syntactic Reference Corpus of Medieval French*, http://srcmf.org/

[33] https://www.ling.upenn.edu/hist-corpora/citing-corpora.html. The page also points to historical corpora in other languages, e.g. Icelandic.

[34] Pragmatic Resources for Old Indo-European Languages, http://www.hf.uio.no/ifikk/english/research/projects/proiel/

Relation extraction has also been used for other purposes than supporting the creation of ontologies. Szpektor et al. (2007) and Généreux (2007) applied relation extraction to improve (cross-language) retrieval in cultural heritage corpora. Klein et al. (2014) extracted relations between commodities and locations in a historical corpus about global trading in the British empire.

Relation Extraction, and other Information Extraction technologies like Named Entity Recognition (p. 17), have also been put to use to automatically annotate excavation reports and other archaeological grey literature. Vlachidis (2012) created an Information Extraction system (including RE) to annotate such material against the CIDOC-CRM ontology for the cultural heritage domain (Crofts et al., 2011), including an extension specialized in modeling archaeological objects and processes called CRM-EH, by Cripps et al. (2004).[35] Relations extracted tie together an archeological find and its location or its period as expressed in the corpus. The system is rule-based and implemented with the JAPE pattern matching engine in the GATE platform (Cunningham et al., 2002).[36]

Regarding **Open Relation Extraction** (ORE), Bamman et al. (2016) present an unsupervised model to infer the latent political positions of speakers emitting a statement like "Obama is a socialist" in blogs. The emphasis and innovation lies in the inference models rather than on the relation extraction aspect. However, I am citing this work since it does use ORE to obtain the basic units on which the model operates. Taking propositions obtained by ORE as $\langle subject, predicate \rangle$ tuples, they propose a method to determine the subjects most likely to be discussed by communities of liberal vs. conservative orientation, as well as the most likely elements predicated on those subjects by each community. In a political blog context, unlike in more factual contexts like an encyclopedia article, a declarative statement like "Obama is a socialist" need not be a description of reality, but the speaker's opinion. Instead of extracting declarative propositions that are expected to represent facts, the study provides a method to discover the *political import* or the likely ideological position behind the statements.

2.5.3 Event extraction applications

Cybulska et al. (2011) performed historical event extraction on Dutch news covering part of the Bosnian War in the 1990s. The system exploits information from an NLP pipeline and is otherwise knowledge-based, relying on knowledge repositories like WordNet and on extraction patterns created on the basis of human corpus analysis. They defined an event as a

[35] For instance, the following URL shows RE results for a medieval castle find: `http://andronikos.co.uk/Anno_CRM-EH.php?id=3226&view=abstract`

[36] GATE stands for *General Architecture for Text Engineering*

tuple containing an event type, action, participants, time and location. This event definition is intended to allow comparing how historical events are represented in different sources. They differentiate *historical* events and *non-historical* ones. One of the factors that permits identifying non-historical events is *modality*: actions reported as potential, or expressed with subjectivity indicators are not considered historical events. Both verbal and nominal action triggers are handled. Some of the event types covered are *deportation*, *murder*, *offensive*, or *signing a treaty*. Their NLP pipeline includes Word-Sense Disambiguation (p. 17) to WordNet, and WordNet sense annotations are one of the elements exploited in the system's event detection rules. For instance, candidate participants are restricted to *human* participants using WordNet-derived semantic classes. The rules are pattern-based, formalizing part-of-speech and semantic constraints. The results were evaluated on a manually annotated reference set, with overall event detection F1 being at 53. This is not a low result taking into account the level of detail in their event tuples, which require time and location besides participants. Looking at individual event elements, event trigger identification F1 was around 60 and reached 70 for time expressions.

Finally, I would like to mention the recent EU-funded Newsreader project. This project has developed software for different aspects of Event Extraction, including modality detection and event coreference, in several European languages (Vossen et al., 2014 i.a.).[37] They analyzed large corpora as a use case, covering topics relevant for political analysis, like business, finance and economy news. The Semantic Role Labeling tool which I applied for statement extraction in chapters 4 and 6 was developed by that project.

2.6 Summary and Implications for our Work

Whereas Chapter 1 focused on detecting core actors and concepts in a corpus, this chapter discussed methods to automatically detect relations between them. Both tasks are part of Information Extraction, which seeks to derive structured information from unstructured text (2.1.1). Each of the technologies that can help identify relations is a vast field in itself, and only a brief overview was provided here. A summary of the chapter and some comments on the relevance of the technologies for this thesis follow.

2.6.1 Summary

Several sources can be used to identify relational information. If the relations can be modeled satisfactorily with **syntactic** functions, a **dependency parse** of the sentence (2.2.1) may provide the relevant relations. In semantic

[37] http://www.newsreader-project.eu/results/software/. I have no affiliation with this project, but I am familiar with some of the tools they created.

dependency parsing, also called ***Semantic Role Labeling*** (SRL), the relations belong to a predefined typology of semantic roles (2.2.2). Semantic roles can be very general like *agent* or *patient/theme* (used in the PropBank annotation scheme), or tied to a set of activities or situations involving similar actions or states and similar participants, like the *frames* in the FrameNet scheme.

The term ***relation extraction*** (RE) is generally used as a technical term for the task of finding entity attributes in text (2.3). In "Ginni Rometty is the CEO of IBM", Relation Extraction would determine that *IBM* corresponds to the *employer* attribute of entity *Ginni Rometty*, and that *CEO* is the *job title* attribute. RE was initially applied to a small number of attributes in specific domains, with first systems being automata-based (2.3.1). Later systems have employed both hand-crafted rules and statistical learning. The automatic creation of training sets, with methods like bootstrapping and distant supervision (p. 36), has also been an important area in relation extraction.

Another approach to RE is **Open Relation Extraction** (ORE), where the set of relations to extract is not predefined and not restricted to a domain (2.3.2). Shallow methods (based on part-of-speech patterns), as well as deeper methods requiring full syntactic dependency parsing and SRL have been considered. The focus of ORE is providing efficient tools that scale to web-corpora, even in cases where methods requiring deep syntactic and semantic analysis are applied.

RE extracts entity attributes. In **Event Extraction**, we are also interested in annotating the events that entities participate in. In "Ginni Rometty became CEO of IBM in 2012", there is an event with participants *Ginni Rometty* and *IBM*, whereby the attribute *CEO* becomes associated with *Ginni Rometty* at the time specified in the sentence (2012). Event Extraction requires identifying event triggers (actions that indicate the event), as well as the participants. Depending on task definition, it may also be required to identify time, location, and modality (i.e. whether the event is asserted or expressed in a non-factual mode like a hypothesis).

Evaluation campaigns have taken place for all the technologies above (p. 30). Syntactic dependency parsing and SRL reach F1 scores above 85 and 80 respectively (pp. 34–35). Relation Extraction has attained lower results, up to F1 scores around 60 to 65 if relation classification is required besides relation detection, and depending on evaluation method (p. 39). For Event Extraction, results vary depending on the event elements assessed. Event trigger detection and classification reaches an F1 of about 65, and F1 for event argument detection and classification reaches approx. 45 (p. 45). I provided example results reported in the evaluation campaign literature. These

numbers need not be a good predictor of the technology's performance for a specific application in new corpora. Nor do low scores entail lesser value of the technology for a DH application. Task requirements at evaluation campaigns can be very detailed, which increases difficulty in a way that a given specific application would not pose. And vice-versa, a DH use case may pose challenges that were not considered at standard evaluations.

Examples of the relevance of **these technologies in DH** were provided (2.5). Parsing has supported the analysis of agentivity portrayal in news, for political analysis. An example of Event Extraction to annotate historical events in war was also reviewed, which is intended to facilitate the comparison of event portrayal in different sources (2.5.3). Relation extraction has been used in cultural heritage domains (2.5.2), either to help build an ontology for the domain, or to apply existing ones, like CIDOC-CRM in archaeological reports. A common goal of several of the applications reviewed for relation extraction was enriching a document collection with automatically created metadata, which would enhance findability of information in the collection, or corpus navigability. This matches the overall objective in information extraction to create structured information from a mass of unstructured text. It is also an objective in this thesis: With a view to comparing participants' positions, a body of negotiation reports are structured according to the participants mentioned in them, their statements, and the expressions relating both, as will be developed in Chapters 4 and 6.

2.6.2 Implications for our work

As will be seen in Chapters 4 and 6, the application of relational information extraction in the thesis involves extracting statements emitted by actors in political negotiations about climate change. Technologies providing adequate information to this end are syntactic and semantic dependency parsing; the task is not unlike the "clause extraction" application mentioned on p. 46.

Several of the DH-relevant Relation Extraction and Event Extraction applications reviewed in 2.5.2 and 2.5.3 are specific to a domain and knowledge-based, using rules created on the basis of human corpus analysis. This was also the method for the application in the thesis, given a manageable domain with a constrained set of knowledge to model. The thesis uses only a small subset of the technologies reviewed in this chapter. A large part of the material was reviewed in order to offer a more complete picture. Also, because for corpora of different characteristics to the one in the thesis, such other technologies could be better alternatives for obtaining the same types of results. For instance, for a corpus where the relevant relations to extract are not known beforehand, Open Relation Extraction methods (2.3.2) could be a better way to start.

Part II

NLP Technology Support

NLP Technology Support: Introduction

An objective in this thesis is creating applications that provide an overview of a corpus and allow navigating it based on actors and concepts mentioned in it, as well as on the relations between those actors and concepts.

Part I surveyed technologies relevant to that end. Chapter 1 introduced a technology that helps find actors and concepts, called Entity Linking. This detects mentions in a corpus to terms from an encyclopedic knowledge repository like DBpedia. Chapter 2 presented an overview of technologies that identify relations between sentence elements, such as syntactic dependency parsing, semantic role labeling, relation extraction and event extraction.

As was mentioned in Part I, applying Natural Language Processing (NLP) technologies to the variety of corpora in Humanities and Social Sciences poses several challenges. In Part II, we discuss the implementations we undertook involving Entity Linking (Chapter 3) and technologies to extract relation information (Chapter 4), in order to apply them to specific corpora relevant for Digital Humanities. These implementations are then exploited in the applications discussed in Part III.

The quality of Entity Linking results varies according to the corpus. Taking this into account, Chapter 3 describes our implementation of a weighted voting procedure to combine the outputs of Entity Linking tools, in order to obtain combined results that improve over the ones provided by individual systems. The procedure was evaluated on four publicly available test-sets. The result combination procedure was then used in an application to navigate a heterogeneous corpus about the American Financial Crisis of 2007–8, known as the *PoliInformatics* corpus. This application will be presented in Part III (Chapter 5).

Chapter 4 describes our system to identify propositions for speech events, defined as triples containing an actor, a message emitted by that actor, and the predicate (a verb or a noun) relating both. The predicate gives an indication of the actor's position regarding the issues mentioned in the message: whether there's an attitude of support, opposition, or a neutral

tone. The system also identifies which actors support and oppose each other and about what statements. The system is based on an NLP pipeline, which provides syntactic dependency parsing and semantic role labeling. The system also relies on a domain model with actors and predicates. Rules are applied to the NLP output in order to extract propositions. The system was used to analyze a corpus of reports on international climate policy negotiations, from a publication called the *Earth Negotiations Bulletin* (ENB).

Some of the proposition extraction rules are generic, and some rules were created to adapt to the way information is presented in the ENB corpus, which pre-existing tools would not treat optimally. As regards evaluation, the system was assessed against a manually annotated reference corpus.

The proposition extraction system in Chapter 4 is at the basis of an application, presented in Part III (Chapter 6), to navigate the ENB corpus using propositions, enriched with keyphrases and entities. The application intends to help examine actors' positions in the negotiation. A qualitative evaluation by domain-experts (also in Chapter 6 in Part III) suggests the usefulness of our proposition extraction workflow.

Chapter 3

Entity Linking System Combination

3.1 Introduction

Entity Linking (EL) (Rao et al., 2013) maps mentions (i.e. sequences of words in a text) to terms from a knowledge base (KB), be it generic KBs like DBpedia, YAGO, Freebase or BabelNet, or domain-specific ones. Mentions can be ambiguous across several KB-terms, and the challenge for an EL system is to choose the KB-term that best reflects the sense of the mention in its context. Consider the phrase *Clinton Sanders debate*. In it, *Clinton* more likely refers to DBpedia term *Hillary_Clinton* than to *Bill_Clinton*. Conversely, in the phrase *Clinton Bush debate*, *Clinton* more likely refers to *Bill_Clinton*.

This thesis focuses on EL systems which disambiguate to generic KBs, and their applicability to Digital Humanities (DH). The definition of Entity Linking adopted here also involves what some authors call *Wikification*, i.e. linking not only named-entity mentions to the KB, but also other types of mentions susceptible to refer to a KB-term—a review of these notions was provided in Chapter 1.2.

The EL literature has shown that the quality of generic EL systems' results varies widely depending on characteristic of the corpora they are applied to, or on the types of entities we need to link (Cornolti et al., 2013; Usbeck et al., 2015). This poses challenges for a DH researcher wishing to find the best EL system for their corpus. A way to make up for the uneven performance of entity linking methods across corpora would be mixing different annotators' results, so that the annotators' strengths complement each other. This chapter presents an EL system combination approach that improves results over the systems taken individually. The method involves a weighted voting scheme that had not been previously applied to EL, and is introduced in 3.3. A quantitative intrinsic evaluation on four EL datasets is presented in 3.4, and the results and remaining challenges are discussed in 3.5. A detailed chapter summary can be found in 3.6.

3.2 Related Work

The goal of combining different NLP systems is obtaining combined results that are better than the results of each individual system. Fiscus created the ROVER method (Fiscus, 1997),[1] with weighted voting to improve speech recognition outputs. ROVER was found to improve parsing results by (De La Clergerie et al., 2008). In Named Entity Recognition (NER), Rizzo et al. improved results combining systems via different machine learning algorithms in their NERD-ML tool (Rizzo et al., 2014).

In entity linking, the potential benefits of combining annotations have been explored before. Rizzo et al. (2012) describe the NERD system,[2] which combines the results of several entity linkers. However, we are not aware of a system that, like ours, makes an automatic choice among the systems' conflicting annotations, based on an estimate of each annotation's quality. Our approach to choose among conflicting annotations is inspired by the ROVER method, which had not been previously attempted for EL to our knowledge. A further difference in our system is that, whereas in NERD (or NERD-ML) several commercial annotators had been used, we selected the set of linkers to combine among publicly accessible systems. Besides, all of the systems we selected are open source, with the exception of Babelfy—In the case of Babelfy, the KB it links to (BabelNet) is open source, and the tool's disambiguation algorithm was described in (Moro et al., 2014), but the code was not made public.

The intent in this thesis was to apply a simple method to combine systems' outputs and assess its viability. For more complex methods of result combination, the machine learning literature describes stacked generalization or stacking (Ting et al., 1997; Wolpert, 1992), where the output of individual classifiers acts as the input to a higher-level classifier, to improve predictions. Stacking has been applied to NER (e.g. Wang et al., 2008) and it could be applied to the NED step in Entity Linking as well. This would be interesting future work.

3.3 Annotation Combination Method

This section describes the EL systems combined, the workflow to obtain each system's annotations and their combination with a weighted voted scheme.

[1] *ROVER* stands for *Recognizer Output Voting Error Reduction*

[2] This is not the same as the NERD-ML system mentioned in the preceding paragraph. NERD-ML applies machine learning to select among entities *in the NER step*, but does not select among annotators' outputs in the *disambiguation* step.

3.3.1 Systems combined

Our workflow performs English EL to Wikipedia, combining the outputs of the following EL systems: Wikipedia Miner,[3], TagMe 2 ,[4] DBpedia Spotlight,[5] AIDA[6] and Babelfy. The systems rely on a variety on algorithms and it can be expected that their results will complement each other. A description of the systems, listed in chronological order of appearance, follows.

Wikipedia Miner (Milne et al., 2008a): This system was the first to use a graph-based notion of coherence relying on common inlinks from a third Wikipedia page to the pages of two Wikipedia terms as the basis of relatedness between those two terms. The measure is known as *Milne-Witten coherence* in the literature, and it is defined in Equation 5.2. For disambiguation, the system computes the Milne-Witten coherence between candidate KB-terms for ambiguous mentions and the KB-term for *unambiguous mentions* in the context (i.e. mentions that can only be linked to a single KB-term). The system balances the resulting coherence score with candidates' prior probability, depending on an estimation of context quality: in contexts where KB-candidates for ambiguous mentions show low coherence with the KB-terms for unambiguous mentions (low-quality contexts), more importance is given to prior probability than in high-coherence contexts.

This system was accessible as a publicly hosted service, which was stopped in January 2016. Local setup is not trivial. This makes it difficult to reproduce results obtained with this system.[7]

TagMe 2 (Ferragina et al., 2010; Cornolti et al., 2013): Spotting is based on a surface-form dictionary, which maps Wikipedia link-anchor texts to the title of their target page. Besides anchors, page titles, redirects, and cleaned up page titles (stripped of class specifiers in parenthesis, like *(politician)*) are also considered possible surface forms. Disambiguation is performed thus: For each mention, each candidate KB-term votes for all the other candidates. The vote is the average relatedness between each candidate term and all the others, computed with the Milne-Witten coherence by Milne et al. (2008a), defined below in Equation 5.2. Unlike in Wikipedia Miner, both ambiguous and unambiguous mentions are used to compute a candidate's coherence, to account for short texts that may have no unambiguous mentions. The

[3] https://github.com/dnmilne/wikipediaminer

[4] https://tagme.d4science.org/tagme/ The system is a reworked version of TagMe (Ferragina et al., 2010)

[5] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki

[6] https://github.com/yago-naga/aida

[7] A. Conde created a fork at https://github.com/Neuw84/wikipediaminer, and reports having set the system up for his pipeline in NLP for educational applications (Conde et al., 2016), but the instance is not publicly available, and the configuration is likely not the same as in Wikipedia Miner's original public web service.

relatedness vote is balanced with a candidate's prior probability for the mention, to arrive at the final candidate score.

DBpedia Spotlight (Mendes et al., 2011; Daiber et al., 2013): Spotting is dictionary based, in a similar way to TagMe. Disambiguation uses cosine similarity between tokens in a context window around a mention, and tokens in the context vector for candidate DBpedia terms. The context vector for a candidate DBpedia term is the concatenation of all paragraphs mentioning the term in Wikipedia. Vector tokens are weighted with a measure encoding their discriminative ability to tease candidates apart, called *Term Frequency – Inverse Candidate Frequency*.[8] This takes into account how many candidates have the token in their context vector, and with what frequency.

AIDA (Hoffart et al., 2011): This system links to the YAGO knowledge-base (Suchanek et al., 2007), but its outputs also contain Wikipedia URLs, which were used to compare its annotations with the other systems'. For mention spotting, AIDA uses Stanford NER (Finkel et al., 2005). For disambiguation, both context-vector similarity (with Mutual Information weights) and Milne-Witten type graphical coherence (p. 142) are used. These factors are balanced with candidates' prior probability. In fact, coherence is used to solve conflicts when there is a strong disagreement (above a threshold) between the candidate ranking proposed by prior probability and the one proposed by context similarity.

Babelfy (Moro et al., 2014): This system links to the BabelNet knowledge base (Navigli et al., 2012). However, it also outputs DBpedia URLs, which were used to compare its outputs to the other systems', which link to DBpedia or Wikipedia. Spotting in Babelfy is based on a dictionary of BabelNet lexicalizations (i.e. sequences that may express a BabelNet term). Disambiguation proceeds thus: First, given the candidate KB-terms for each textual mention, a semantic signature for each candidate is built, i.e. a directed graph created via random walks with restart (Tong et al., 2006), where edges in more densely connected areas bear more weight. Then, a candidate graph is created for candidate terms included in each other's semantic signature. The output is a ranked list of KB-candidates, which come from a densest subgraph in the candidate graph.

3.3.2 Obtaining individual annotator outputs

This subsection describes how annotations are obtained from the five entity linking systems just discussed.

[8] By analogy with the *Term frequency – Inverse document frequency* measure (Salton et al., 1983, Section 3B) common for term weighting in Information Retrieval based on the Vector Space Model (Salton et al., 1975). This measure assigns higher weights to terms that are frequent in a document, but infrequent in the collection overall, thus being able to discriminate a subset of documents from the rest of the corpus.

Clients were created to request the annotations for a text from each linker's web-service,[9] using the services' default settings except for the confidence threshold, which is configured in our workflow.

We obtained optimal thresholds for each system (Each column *t* in Table 3.1 and Table 3.2) with the BAT Framework (Cornolti et al., 2013).[10] The BAT Framework allows calling several entity linking tools and compares their results using different annotation confidence thresholds, with a view to finding the thresholds that yield best results according to several evaluation measures.

Annotations are filtered out if their confidence is below the thresholds obtained in the way just described. The remaining annotations proceed to the annotation-voting step.

3.3.3 Pre-ranking annotators

Our annotation voting exploits annotators' precision on an annotated reference set in order to weight the annotations produced by each annotator. To obtain these precision scores, it is not viable to create a reference set for each new corpus that we need to perform entity linking on. To help overcome this issue, we adopt the following approach: We have ranked the annotators for precision on two reference sets: AIDA/CONLL Test B (Hoffart et al., 2011), and IITB (Kulkarni et al., 2009). The IITB dataset contains a large proportion of annotations for category *Others*, i.e. entities that are not a person, organization or location (see Waitelonis et al., 2016, Table 1), whereas in AIDA/CONLL B such annotations amount to 14% only (Tjong Kim Sang et al., 2003, Table 2). The proportion of annotations in a corpus that fall into the *Others* category is a strong predictor of annotators' performance on that corpus, according to a study on how different dataset features correlate with annotators' results, available on the GERBIL platform[11] (Usbeck et al., 2015). Taking this into account, in order to annotate a new corpus, if annotations for the *Others* category are needed for that new corpus, the annotator ranking for the IITB corpus will be used in order to weight the new corpus' annotations, since IITB is the only one among our two reference sets that contains a large proportion of annotations for *Others*, and an annotator performing well on IITB is likely to perform well when annotations for *Others* are needed. If, conversely, annotations for the *Others* category are not needed, the annotator

[9] The public deployments were used, but for AIDA, which was set up locally: Source v2.1.1, Data 2010-08-17v7. In AIDA, the `tech=GRAPH` option was used (non-default, but recommended by AIDA's authors for benchmarking). For Babelfy, the REST API version used was 0.9, which is now deprecated.

[10] https://github.com/marcocor/bat-framework

[11] See *Annotator - Dataset* feature correlations at http://gerbil.aksw.org/gerbil/overview

ranking for the AIDA/CONLL B reference corpus is used in order to weight the new corpus' annotations.

Note that, rather than on characteristics of the corpus, the ranking that will be used to combine EL systems depends on a choice by the user, i.e. whether they want to annotate a small set of entity types, or a wider set of mentions, including those for KB-terms expressed by common nouns. This does not seem a flexible choice and in this sense an annotation combination that relies on characteristics of the user's corpus to decide on a system ranking would be desirable. As reported in (Ruiz Fabo et al., 2015c), I had tested corpus features like lexical cohesion (inspired by Hearst, 1997) or document length as predictors of annotator performance. However, my results were uneven and I do not use such features in the system described in this chapter.

3.3.4 Annotation voting scheme

The voting scheme is in Figure 3.1. Each annotation is formalized as a pairing between a mention m (a span of characters in the text) and a Wikipedia entity e. For each annotation $\langle m, e \rangle$, the set Ω_m contains the annotations whose mentions overlap with m. Two mentions (p_1, e_1) and (p_2, e_2) overlap iff with p_1 and p_2 as the mentions' first character indices, and e_1 and e_2 as the mentions' last character indices $((p_1 = p_2) \wedge (e_1 = e_2)) \vee ((p_1 = p_2) \wedge (e_1 < e_2)) \vee ((p_1 = p_2) \wedge (e_2 < e_1)) \vee ((e_1 = e_2) \wedge (p_1 < p_2)) \vee ((e_1 = e_2) \wedge (p_2 < p_1)) \vee ((p_1 < p_2) \wedge (p_2 < e_1)) \vee ((p_2 < p_1) \wedge (p_1 < e_2))$. The set Ω_m is divided into disjoint subsets, each of which contains annotations linking to a different entity. Each subset L is voted by $vote(L)$: For each annotation o in L, N is the number of annotators we combine (i.e. 5), $r_{o,anr}$, is the rank of annotator anr, which produced annotation o, and P_{anr} is anr's precision on the ranking reference corpus (see 3.3.3). Finally, parameter α influences the distance between the annotations' votes based on their annotators' rank, and was set at 1.75 based on the best results on both ranking reference corpora. The entity for the subset L which obtains the highest vote among Ω_m's subsets is selected if its vote is higher than P_{max}, i.e. the maximum precision for all annotators in the ranking corpus.[12]

Once an entity has been selected for a set of overlapping mentions, the mention itself needs to be selected. Best results were obtained when the most common mention in the set was selected. In case of ties, the longest mention among the most common ones was selected (e.g. if two mentions occur twice each in the set, select the longer one).

Note that we had tested an alternative voting scheme in (Ruiz Fabo et al., 2015b), which takes into account annotation confidence scores (provided

[12] See Table 3.1 and Table 3.2 below for P_{max} values in the ranking reference corpora: P_{max} is the maximum (excluding row *Combined*) in the P columns for *AIDA/CONLL B* and *IITB*

for each set Ω_m of overlapping annotations :
for $L \in \Omega_m$:

$$vote(L) = \frac{\sum_{o \in L} \left(N - (r_{o,anr} - \alpha)\right) \cdot P_{o,anr}}{N}$$

$$\text{if } \max_{L \in \Omega_m} \left(vote(L)\right) > P_{max} : \text{select } \operatorname*{argmax}_{L \in \Omega_m} \left(vote(L)\right)$$

FIGURE 3.1 – Annotation voting scheme

by most EL systems). At the SemEval 2015 evaluation campaign, Task 13 (Multilingual All-Words Sense Disambiguation and Entity Linking, Moro et al., 2015), this alternative voting scheme obtained good results at the English Entity Linking subtask, ranking third of the 10 participant systems. However, the task's dataset[13] only contained 86 items for Entity Linking, so the task's results need not be a solid indication of a system's performance on other corpora. Besides, the fact that EL systems do not obligatorily provide confidence scores was a reason for me to prefer the voting scheme presented in Figure 3.1 here over the alternative one, and I did not test the alternative one on the datasets in 3.4 below.

3.4 Intrinsic Evaluation Method

This section describes the evaluation approach: the datasets, evaluation measures and tools are presented.

Datasets: The workflow was tested on four golden sets. First, the two datasets that had also been used as reference sets in order to obtain the weights to vote annotations with (3.3.3): AIDA/CONLL B (231 documents with 4485 annotations; 1039 characters avg., news and sports topics) and IITB (103 documents with 11245 annotations; 3879 characters avg., topics from news, science and others). In order to test whether the annotator weights obtained from those two corpora can improve results when applied to annotator combination on other corpora, we tested on two additional datasets: MSNBC (Cucerzan, 2007), with 20 documents and 658 annotations (3316 characters avg., news topics) and AQUAINT (Milne et al., 2008b), with 50 documents and 727 annotations (1415 characters avg., news topics).

The AQUAINT dataset contains annotations for common-noun entities (besides annotations for proper nouns referring to entities of type *Person*, *Location*, or *Organization*). For this reason, according to the procedure described in 3.3.3 above, its annotations were weighted according to annotators' ranking on the IITB corpus, which also contains common-noun annotations.

[13] http://alt.qcri.org/semeval2015/task13/index.php?id=data-and-tools

The MSNBC dataset does not contain common-noun annotations, so the annotator ranking for the AIDA/CONLL test-set was used in order to combine annotations in MSNBC.

Measures: The EL literature has stressed the importance of evaluating systems on more than one measure. We tested the workflow on strong annotation match (SAM) and entity match (ENT) (Cornolti et al., 2013). SAM requires an annotation's position to exactly match the reference, besides requiring the entity annotated to match the reference entity. ENT ignores positions and only evaluates whether the entity proposed by the system matches the reference.

Mapping files: Evaluating EL to Wikipedia requires making sure that we consider the same set of target entities for each EL system, since the versions of Wikipedia deployed within each system may differ, and the same happens with the versions of Wikipedia the golden sets had been annotated against. Before evaluation, annotations in both the golden sets and the system results were mapped to Wikipedia-page titles as of March 2015 (e.g. titles that redirected to a newer preferred variant for the title in the March 2015 version were mapped to the newer preferred variant).[14]

Tools: Evaluation was carried out with the neleval tool[15] from the TAC-KBP Entity Discovery and Linking task (Ji et al., 2014). The tool implements several EL-relevant metrics, accepting a common delimited format for golden sets and results across corpora. The tool's significance testing function via randomized permutation/bootstrap methods was also applied to our results.

3.5 Results and Discussion

This section presents the results and provides a reflection on the relevance of the results for DH research.

3.5.1 Results

The annotator rankings and weights with which annotations were weighted in our voting scheme (Figure 3.1) can be read off the P column for the ranking reference corpora (AIDA/CONLL or IITB) in Table 3.1 and Table 3.2. For instance, results for MSNBC were combined using the ranking from AIDA/CONLL. Looking at Table 3.1, this means that MSNBC annotations (for the SAM measure) were weighted with the following values, in format (Annotator, Rank, Weight): (AIDA, 0, 0.767), (Tagme, 1, 0.548), (Wikipedia

[14] The mapping was created based on fetch_map from the conll03_nel_eval tool by Hachey et al. (2013), https://github.com/wikilinks/conll03_nel_eval

[15] https://github.com/wikilinks/neleval/wiki

Corpus	AIDA/CoNLL B				IITB				MSNBC				AQUAINT			
System	*t*	P	R	F1	*t*	P	R	F1	*t*	P	R	F1	*t*	P	R	F1
Tagme	0.219	54.8	53.9	54.4	0.086	41.1	42.6	41.8	0.188	44.7	42.4	*43.5*	0.188	39.9	46.5	43.0
Spotlight	0.086	28.1	38.8	32.6	0.016	41.0	48.2	44.3	0.063	21.8	28.1	24.6	0.055	15.6	45.3	23.2
W Miner	0.57	45.3	50.3	47.7	0.25	55.2	44.4	*49.2*	0.664	42.3	38.2	40.2	0.57	34.8	57.6	*43.4*
AIDA	0.0	76.7	46.7	*58.1*	0.0	50.2	5.6	10.0	0.0	63.6	23.8	34.7	0.0	50.3	27.7	35.7
Babelfy	dna	34.7	34.0	34.3	dna	46.8	14.9	22.7	dna	31.8	28.8	31.1	dna	22.6	31.5	26.3
Combined	dna	64.8	61.7	***61.9**	dna	59.3	44.7	***50.0**	dna	54.3	43.4	***48.2**	dna	34.1	64.1	**44.5**

TABLE 3.1 – Strong annotation match (SAM). Optimal confidence thresholds *(t)*, Micro-averaged Precision, Recall, F1 for each annotator and combined system. The best-result is bold and the second-best italicized. Statistically significant differences between the best and second-best scores are starred. The combined system, and Babelfy in the version we accessed, output no confidence thresholds (dna).

Corpus	AIDA/CoNLL B				IITB				MSNBC				AQUAINT			
System	*t*	P	R	F1	*t*	P	R	F1	*t*	P	R	F1	*t*	P	R	F1
Tagme	0.234	58.2	67.9	62.7	0.102	47.6	45.7	46.7	0.328	66.8	49.9	57.1	0.198	63.8	55.4	59.3
Spotlight	0.094	30.8	40.1	34.8	0.008	36.6	51.8	42.9	0.063	21.6	27.5	24.2	0.055	26.2	49.8	34.3
W Miner	0.477	46.9	57.3	51.6	0.195	61.3	43.3	*50.6*	0.664	50.1	52.8	51.4	0.523	59.9	62.5	*61.1*
AIDA	0.0	79.7	79.7	***79.7**	0.0	61.4	11.72	19.7	0.0	74.6	56.3	*64.2*	0.0	67.8	37.3	48.1
Babelfy	dna	35.6	37.9	36.7	dna	48.4	16.3	24.4	dna	36.5	37.5	37.0	dna	39.1	37.8	38.3
Combined	dna	65.0	78.5	*71.1*	dna	60.7	44.6	***51.4**	dna	66.7	62.3	**64.4**	dna	58.4	67.3	***62.5**

TABLE 3.2 – Entity Match (ENT). Optimal confidence thresholds *(t)*, Micro-averaged Precision, Recall, F1 for each annotator and combined system. The best-result is bold and the second-best italicized. Statistically significant differences between the best and second-best scores are starred. The combined system, and Babelfy in the version we accessed here, do not output any confidence thresholds (dna).

Miner, 2, 0.453), (Babelfy, 3, 0.347), (Spotlight, 4, 0.281). The P_{max} value that each annotation's vote is compared to in MSNBC is 0.767.

In Tables 3.1 and 3.2, the best F1 score in each corpus is marked in bold, and the second-best F1 is in italics. The combined workflow obtains the best score in all cases, except ENT scores on AIDA/CONLL B. For the SAM measure, the improvements range between 0.8 points and 4.7 points of F1. For the ENT measure, improvements range between 0.2 and 1.4 points of F1. The differences are statistically significant in the majority of cases (scores with a star). Significance ($p < 0.05$) was assessed with the random permutation method in the neleval tool.

The combined workflow was able to improve over the best individual system regardless of which this system was: Tagme, Wikipedia Miner or AIDA. In some cases, the improvements over the best individual system's F1 take place because of markedly increased recall in the combined system compared to the best individual system's recall, without a major decrease in precision in the combined system (see AQUAINT results for ENT). The opposite pattern of improvement is also attested: In the MSNBC results for SAM, it is the increased precision of the combined workflow that makes its F1 improve over the best individual system's F1.

Regarding the significant drop in F1 in the combined system vs. the best individual system (AIDA) in the ENT results for the AIDA/CONLL B corpus, note that, in this case, the difference between AIDA's individual results and the results for the second-best individual system was much higher (17.2 points of F1) than anywhere else in the rest of tests performed. When such a large difference exists between the best individual system and the rest, an alternative type of voting may be needed in order to improve results over the best individual system.

3.5.2 Discussion: Implications for DH research

A first comment to make about the results is that one of the Entity Linking systems they were obtained with (Wikipedia Miner) is no longer publicly available, and a local setup is not trivial.[7] This makes it difficult to use the system combination as described here. To ensure continued availability of an application, it is preferable to work with tools that can be deployed locally. This has the additional advantage that the systems' configuration can be controlled by whoever deploys the systems, unlike publicly hosted services that do not expose configuration settings. Access to the configuration facilitates the reproducibility of results—see (Hasibi et al., 2016) for detailed discussion of EL reproducibility.

Speaking of the results now, Strong Annotation Match F1 scores range between 44.5% and 61.9%, the range covering somewhat higher values for Entity Match (51.4% to 79.7% of F1). These figures may seem low to a DH researcher, who might pose the question whether these results are good enough for scholarly work. Some reflection about this is in order.

First, agreement between human annotators as to what a relevant mention to link (and, to a lesser extent, what the correct link should be), is limited. The human annotation process (including a quantification of agreement between annotators) has been described for three of the four datasets I used as test-sets.[16] For the AIDA/CoNLL dataset, (Hoffart et al., 2011, p. 789, Table 1) report a disagreement between the two annotators for 21.1% of the mentions.[17] For IITB, (Kulkarni et al., 2009, p. 462, Figure 7) report that there was disagreement for 20% of the mentions proposed by their two annotators. For the AQUAINT reference-set, (Milne et al., 2008b) computed agreement between three annotators, so disagreement is of course higher and not comparable to what the other references here report (see section 5.2 in their article for a detailed account). Finally, for MSNBC, (Cucerzan, 2007)

[16] The papers describe ratios of agreement vs. disagreement, rather than using an Inter-annotator agreement metric as those described in (Artstein et al., 2008).

[17] This is over the 1393 articles including Set B (the 231-document test-set, which I used here), Set A (the development-set, unused here), and the training set. It is not reported how disagreement cases were distributed across corpus subsets.

		Top 5 candidates			Top 10 candidates			Top 15 candidates		
System	Rank	P	R	F1	P	R	F1	P	R	F1
HUMB	1	39.0	13.3	19.8	32.0	21.8	26.0	27.2	27.8	**27.5**
WINGNUS	2	40.2	13.7	20.5	30.5	20.8	24.7	24.9	25.5	25.2
KP-Miner	3	36.0	12.3	18.3	28.6	19.5	23.2	24.9	25.5	25.2

TABLE 3.3 – Micro-averaged F1 for the top 5, 10 and 15 candidates proposed by the best three keyphrase extraction systems participating at SemEval 2010, Task 5, "Automatic Keyphrase Extraction from Scientific Articles". These results are mentioned here as an example of a technology which is considered useful in DH, yet achieves low F1 scores at a quantitative evaluation task.

does not describe annotations by several humans. Most disagreement cases take place over whether a mention should be considered or not; the choice of the best KB-term for each mention causes much less disagreement. Since there is 20% disagreement among human annotators, the upper bound for humans at the task is not at 100%: In this sense, this is a difficult task, and automatic results need to be assessed accordingly.

A second question would be whether there is a better alternative to these EL systems. Another technology that annotates raw text to identify key concepts is *keyphrase extraction*, which extracts sequences of words that are both frequent in the corpus and discriminant within it, i.e. meaningful to isolate a subset of the corpus. Keyphrase extraction is a basic text mining technology routinely used in information retrieval or search engines (Rose et al., 2010, pp. 3–4). It is relevant to DH research (e.g. in the ALCIDE text analysis platform (G. Moretti et al., 2016), just to cite a system deploying the newest keyphrase extractor known to me, *KD* by G. Moretti et al., 2015). Industrial applications include automatic bibliographic indexing and scientific literature mining. I think that the applicability of this technology to text analysis in DH is undisputed. However, if we look at quantitative evaluations of keyphrase extraction systems, the figures are below 30% of F1 (micro-averaged) for the best systems. Kim et al., 2010 organized a keyphrase extraction task at SemEval 2010 (Task 5). The reference set contained keyphrases assigned to 284 scientific articles both by a group of human annotators and by the authors of the papers. Participant systems output a ranked list of keyphrases, and were evaluated on the subset of reference keyphrases found (as an exact match) among the top 5, 10, and 15 system outputs. Table 3.3 shows results for the best three systems. Besides these scores, note that, while agreement among annotators is not reported, an indication of the human agreement ratio for the task is the fact that 32.3% of the keyphrases assigned by human annotators to articles matched exactly keyphrases provided by the articles' authors, with many more partial matches (p. 23).

The keyphrase extraction task is of course different to Entity Linking, as there is no notion of a target knowledge base, and the evaluation methods are not the same, so the results cannot be compared directly. The reason

why I provided keyphrase extraction results was as an example that, a text mining technology that is generally accepted as useful by a community of users can obtain low scores when the technology is evaluated quantitatively in a formal task comparing to reference human annotations. Moreover, as in the EL case, agreement between human annotators can be limited. This must not be read as a criticism to NLP evaluation tasks and campaigns, which allow a basis for system comparison and systematic examination of NLP methods. Rather, as a suggestion that a technology's value for a DH scholar need not be obvious from the scores at such tasks.

In short, low figures for an automatic system at an NLP task need not mean that humans would obtain substantially better performance at the task, the way the task was defined in NLP terms. The implications for the DH scholar would be to be aware that the results may be limited in coverage or in level of detail. In this thesis, in Chapters 5 (5.2.5) and 6 (6.6), qualitative evaluations with domain-experts were carried out. Evaluations along those lines may be a good complement for a scholar to judge the value provided by a technology in his or her application scenario.

3.6 Summary and Outlook

A detailed chapter summary is provided below, outlining possible future work at the end.

An approach to combine the output of Entity Linking Systems was presented, to obtain combined results that outperform individual systems' outputs. The systems combined were publicly accessible via web services at the time of implementation. Outputs were combined with a weighted voting scheme (ROVER), based on the systems' precision and their rank by precision on two reference corpora (AIDA/CoNLL B and IITB). This improvement in combined results (as measured by Strong Annotation Match and Entity Match) generalized to two additional datasets (MSNBC and AQUAINT), besides the two datasets the systems were ranked on. The combination scheme relies on the reasonable assumption that, if KB-annotations for common-noun mentions are needed, results will improve if the better systems at corpora rich in such mentions are given more weight than the other ones in the vote (and conversely for annotators performing best at corpora where named-entity mentions predominate, if only named-entity mentions need to be disambiguated). However, as the voting scheme is not otherwise adapted to characteristics of the specific corpora to be annotated, it is open to question to what an extent the method can generalize.

The majority of the annotators combined were open-source; this was intended to improve the interpretability of results, since it is possible to look at

the implementation. All of the annotators had a publicly hosted deployment, accessible via web services. In most cases I used this public deployment, which proved a weakness of the approach. Using the public web services causes possible reproducibility issues if the team hosting the public instance change the system (e.g. configuration settings like confidence thresholds, the version of KB, or implementation aspects). In fact, one of the public instances I was accessing, and that gave competitive results on a variety of corpora (Wikipedia Miner) was taken down in early 2016. Hosting this system is not trivial and I cannot currently offer an instance for this system. This weakens the system combination described in this chapter, as one of the annotators it relies on is no longer available. A lesson to learn is that it is preferable to use tools that can be deployed locally with reasonable effort.

About the F1 scores attained by the individual and combined linkers, the range was approx. between 50 and 65 points, depending on the evaluation measure. These are indeed competitive figures and correspond to current implementations of the technology. A DH scholar may ask whether these figures are too low for DH research. I argued that low F1 figures at an evaluation task defined in NLP terms need not indicate lack of usefulness of the technology for DH scholars. I compared the figures with usual F1 scores for keyphrase extraction, which is clearly considered a useful technology for text analysis in DH, and for which F1 scores at NLP evaluation campaigns are below 30 points. I argued for a qualitative evaluation of results with a domain expert, as is done in Part III in this thesis, as a complement to intrinsic evaluation of a technology.

Interesting **future work** would be using result combination approaches from the machine learning literature. Stacked generalization would be a good approach to try with the systems combined here, in that it does not require retraining of the individual systems to be combined. It would involve creating a higher-level classifier that takes as inputs the outputs of each individual system, to provide better predictions.

Chapter 4

Extracting Relations between Actors and Statements

4.1 Introduction

Methods to detect important actors and concepts in a corpus were described in chapters 1 and 3. These methods can be the basis of useful co-occurrence based analyses to gain an overview of a corpus, such as the concept co-occurrence networks described in Venturini et al. (2012), and the clustering and topic modeling approaches surveyed in Grimmer et al. (2013) for political text analysis.

However, these techniques do not identify the nature of the relation between actors and concepts. For instance, if an actor like *France* and a phrase like *stricter regulations* are mentioned in the same sentence, is there a verb mediating between both? Is France in favour of, or against stricter regulations? To analyze a political negotiation corpus, identifying this type of information is useful, and the methods to extract relational information surveyed above in Chapter 2 can help in this respect.

Bearing in mind the usefulness of such relational information, this chapter describes a system for extracting propositions, i.e. triples for an actor, a message emitted by this actor, and the predicate relating both. This extraction is based on the output of a Natural Language Processing (NLP) pipeline, and on a domain-model for actors and predicates of interest. The system also determines which actors oppose each other, and regarding what issues.

The system was applied to analyze a corpus which reports on participants' statements at international climate negotiations (4.2.2). This corpus poses specific needs that the system was designed to address. The system was implemented primarily with this corpus in mind. Some elements in it may generalize to similar reporting corpora, but other aspects are corpus-specific. In other words, as will be elaborated on below, the system relies on a generic NLP pipeline which provides information about syntactic functions and semantic roles. Some of these functions (*subject*, *object*) and roles (*agent*, *theme*) are applicable across corpora, and this is a reason why it is useful to rely on a generic NLP pipeline. However, each corpus may use particular

constructions that benefit from a specialized treatment, and our system also involves means to deal with elements specific to our corpus of application.

The chapter structure is the following: The task fulfilled by the system, and the corpus of application, are introduced in 4.2. Previous work is discussed in 4.3. The system is described in 4.4, providing details about the NLP pipeline behind it, the domain model, the proposition extraction rules, and some discussion of the strengths and limitations of the approach. The evaluation method and results are presented in 4.5, with a discussion of the results. Finally, 4.6 provides a detailed summary and outlook.

4.2 Proposition Extraction Task

This section describes first how we defined a proposition (4.2.1). This was partly determined by the intended corpus of application, described in 4.2.2. Conventions adopted for representing propositions are also outlined, with some examples, in 4.2.3.

4.2.1 Proposition definition

The task consists in extracting propositions from a corpus that reports on political negotiations. A proposition is defined as an ⟨*actor*, *predicate*, *negotiation point*⟩ triple. An *actor* is a participant in the negotiation, generally a country or group of countries, but it can also correspond to other negotiation groups. The *negotiation point* is the message emitted by the actor. The *predicate* is the expression relating the actor and the negotiation point. It can be a verb or a noun, and it indicates the actor's position regarding issues expressed in the point: is there a relation of support, opposition, or a neutral attitude?[1]

As was seen on p. 43, time and location are usually considered as attributes for events (which are similar to the propositions defined here). We have not included time and location as part of the definition of a proposition. In the corpus, they do not need to be extracted automatically from the text, because they are available as metadata: A proposition's time and location are the date of the report, and the conference where it was uttered, respectively.

4.2.2 Corpus of application

The corpus analyzed, which has shaped some aspects of the task definition, comes from volume 12 of the *Earth Negotiations Bulletin* (*ENB*). This publication offers reports covering participants' statements in international climate policy conferences. The corpus covers a 20-year period, between 1995 and

[1] We use *predicate* in the sense of a relation expression on which arguments depend (see p. 32), not in the sense of *subject vs. predicate*, sometimes used in descriptive grammar to differentiate a subject vs. the rest of the sentence.

1 - Multiple verbal predicates

The EU, with NEW ZEALAND and opposed by CHINA, MALAYSIA and BHUTAN, supported including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."

Propositions				**Predicate Type**
	Actor	*Predicate*	*Negotiation Point*	
1	European_Union	supported	including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."	support
2	New_Zealand			support
3	China	~supported	including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."	opposition
4	Malaysia			opposition
5	Bhutan			opposition

2 - Nominal predicate

Much of the discussion was on a proposal by the G-77/China to include research and development in the transport and energy sectors in the priority areas to be financed by the SCCF.

Propositions				**Predicate Type**
	Actor	*Predicate*	*Negotiation Point*	
1	Group_of_77/China	proposal	to include research and development in the transport and energy sectors in the priority areas to be financed by the SCCF.	support

FIGURE 4.1 – Typical ENB corpus sentences. **Sentence 1** has predicates *supported* and *opposed*, with several actors each. **Sentence 2** shows a nominal predicate (*proposal*). For **Sentence 1**, five ⟨*actor, predicate, negotiation point*⟩ propositions are extracted by the system, and the opposing actors (*China*, *Malaysia*, *Bhutan*) are assigned a proposition which is a negated version (with *~supported* as the predicate) of the proposition for the main verb *supported*. Example from Ruiz Fabo et al. (2016b).

2015.[2] The corpus size is approx. 24,000 sentences (500,000 words). More details are provided in Chapter 6 (p. 164ff), the corpus description here is restricted to those aspects most relevant to understand the task.

The ENB corpus strives for an objective tone, neutral towards all participants. This has an impact on the predicates and style employed. A limited set of reporting predicates are used, that are considered non-interpretive of participants' intentions, e.g. *stated* or *emphasized* rather than *accused* or *attacked*. Additionally, the syntax tends to be regular, in order to avoid presenting certain actors more prominently than others.

In spite of regular syntax and a limited predicate range, the sentences contain a lot of information, regularly mention several actors, and can have several predicates of support and opposition. Typical sentences for the corpus are shown in Figure 4.1, which also shows the propositions identified by the system in them.

4.2.3 Proposition representation

In Figure 4.1, **sentence 1** has two predicates, *supported* and *opposed*. Several actors are related to each of the predicates. Five propositions, numbered *1* through *5* in the figure, can be extracted from the sentence. Propositions *1* and *2* have the main verb (*supported*) as their predicate. Propositions *3* through *5* correspond to the opposing actors (China, Malaysia, Bhutan). The opposing actors are assigned a proposition which is a negated version (with *~supported* as the predicate) of the propositions for the main verb *supported*. This is the convention adopted to indicate propositions where an actor opposes the proposition for the main verb. As for **sentence 2** in Fig. 4.1, its analysis shows an example of a proposition with a nominal predicate.

4.3 Related Work

Chapter 2 reviewed technologies used for identifying relations between lexical sequences in a sentence. The majority of those technologies extract relations between entities. For instance, they determine that a person and a location are linked via the relation *be born in*, or that a person and a company are related via the relation *is employee of*. Both rule-based and statistical approaches have been used to annotate such relations.

Relations between entities are relevant to analyze the climate negotiation corpus focused on here (4.2.2), in the sense that we need to determine countries that support or oppose another country. The example in Figure 4.1 is typical for how these relations are explicitly expressed in the corpus. In it, a predicate, *opposed by*, relates two groups of actors. The set of predicates explicitly expressing support and opposition between actors is limited, and we used a dictionary and rule-based approach to find these relations.

For our corpus, we are interested in which country supported or opposed which, but we also need to know about what statements this happened. Accordingly, we need to not only extract relations between entities, but also between an entity and a message emitted by that entity. In other words, we need to extract propositions as defined above (4.2.1), containing a corpus actor, its message, and the predicate mediating between both.

Related work for proposition extraction was also discussed in Chapter 2, such as Van Atteveldt et al. (2017), who describes the *rsyntax* clause extraction tool. The tool identifies events, as well as the source reporting them. Other relevant work is the political event extraction tool in Schrodt et al.

[2]The original HTML corpus is at `http://enb.iisd.org/enb/vol12/`. Its 12th volume covers the Conference of the Parties (COP) summits, which generally take place once a year. We created a raw text and XML version of the corpus based on the original HTML files.

(2014), called PETRARCH. Among other event-types, the tool extracts speech events.

Our corpus is restricted to speech events. In order to focus on these, and to systematically extract as much information as possible about them, as formulated in the corpus, a corpus-specific approach is justified. In the corpus sentences, besides the speech event whereby an actor emits a message, a set of actors supporting or opposing the statement are typically present. We need a procedure that "unpacks" this information, creating individual propositions for each of those actors. The generic systems above were not meant to output such information.

Another source of relation and event information reviewed in Chapter 2 is Open Relation extraction (ORE), where the predicates and relation types need not be known beforehand; ORE tools apply domain-independent predicate and argument detection procedures (p. 37). Three ORE tools were applied to the corpus, and the coverage of corpus statements was not systematic. The tools tested were Exemplar (Mesquita, 2015), OLLIE (Mausam et al., 2012), and Open Information Extraction 4.0.[3]

The corpus requires extracting propositions for nominal predicates. This is one of the shortcomings with the tools above. The ORE tools we tested extracted almost none of these propositions, and nominal reporting predicates are not considered in *rsyntax* or PETRARCH.[4] This is another reason motivating the implementation of a different workflow for our corpus.

A difference between the tools cited above and our system is that, in the tools above, the main source for the recognition of relation information is syntactic dependency parsing, whereas in our system it is semantic dependency parsing (SRL). This technology has sometimes been used successfully for open relation extraction (p. 41), and I find it interesting to test it for extracting speech-related propositions.

A final related study, applied to the same corpus we work with, is in Salway et al. (2014). They used grammar induction with the ADIOS algorithm by Solan et al. (2005),[5] to identify common patterns involving actors and representing issues in the ENB corpus. Their method is unsupervised and performs several iterations of pattern induction over the corpus, to infer

[3]The first two were introduced on p. 38ff. The third one is an evolution of OLLIE. It is available at https://knowitall.github.io/openie/ (v4.0 package) and I am not aware of a publication describing it.

[4]The claim about *rsyntax* is based on the list of speech predicates in the tool, at https://github.com/vanatteveldt/rsyntax/blob/master/R/clauses.R, which the tool uses to find statements' speakers. The claim about PETRARCH is based on its documentation, at https://github.com/openeventdata/petrarch2/blob/master/Petrarch2.pdf

[5]*ADIOS* stands for *Automatic Distillation of Structure*: http://adios.tau.ac.il/algorithm.html

increasingly abstract patterns. To judge from the example patterns in Salway et al. (2014, Table 1), the method is not intended to extract full propositions. It extracts patterns connecting an actor and the verb whereby it makes a statement. It also extracts subsentential sequences reflecting negotiation issues, without systematically linking them to the speaker who talks about them. The technology behind this approach is very different to ours, and it would be interesting to see if the rules induced by this approach complement the propositions annotated by our workflow.

4.4 System Description

Our system extracts propositions based on the information provided by a Natural Language Processing (NLP) pipeline, and on a domain-model of relevant actors and predicates, which are searched in the results of the NLP analysis. As mentioned above, propositions consist of an actor, a message emitted by the actor, and the verb or noun relating both; this verb or noun can express support or opposition, or be a neutral reporting verb.

The workflow starts with predicate identification. Then, actors and messages related to the predicates are identified.

The different elements of the workflow are described in following: The NLP modules used (4.4.1), the domain model (4.4.2), the rules that extract propositions based on both (4.4.3), and a procedure to assign confidence scores to the propositions extracted (4.4.4). Finally, following the system description, I provide some discussion about the choices made in our approach, addressing issues like the extent to which we need NLP technology to extract propositions from the intended corpus, instead of using simpler lexical patterns (4.4.5).

4.4.1 NLP pipeline

The NLP tools used in the system are from the IXA Pipeline toolkit (Agerri et al., 2014), as well as other modules compatible with that toolkit, since they use the same annotation format (details below). The results of this pipeline, like those for a number of other NLP libraries, are state-of-the-art or comparable. We chose the IXA toolkit for the following reasons: It uses a consistent XML annotation format, which can be processed with XML parsing libraries rather than ad-hoc string manipulation. Besides, its Semantic Role Labeling module (p. 79) provides information from several lexical databases at once. Finally, the toolkit has web-services, so it can be exploited easily from other programming languages than its original one—the toolkit is in Java but our system is in Python.

The NLP modules we used are the following:

Tokenization and **part-of-speech tagging** were performed with IXA Pipeline's default models for English. **Constituency parsing** (required by the coreference module) was also carried out with the toolkit's default model.

Coreference resolution was provided by the *CorefGraph* tool.[6] This is a Python implementation of Stanford's *dcoref* module (Lee et al., 2013), and its annotation format is compatible with the IXA pipeline. We did not exploit all coreference chains output by the tool, we only treated some cases of pronominal anaphora. Given non-standard pronoun use in the corpus, we created custom rules for pronominal anaphora resolution, based on CorefGraph's output and on dependency parsing, as explained on p. 83.

Dependency parsing and **semantic role labelling** (SRL) were carried out with *ixa-pipe-srl*.[7] This is a tool compatible with the IXA pipeline, and provides a wrapper around the dependencies and SRL modules in the Mate Tools library (Björkelund et al., 2010). The dependency and SRL format used are the CoNLL ones (Carreras et al., 2005). SRL is performed against the PropBank and NomBank databases (Palmer et al., 2005 and Meyers et al., 2004 respectively), which were introduced on p. 32. A useful feature of *ixa-pipe-srl* is that it includes a database of links across lexical resources called the *Predicate Matrix* (López de Lacalle et al., 2014, 2016). Thus, its SRL annotations, besides providing the relevant PropBank and NomBank predicate senses and roles, also include links to external resources like WordNet (Fellbaum, 1998) and FrameNet (Baker et al., 1998). In other words, this tool adds rich annotations to the corpus predicates and arguments, linking them to several lexical databases at once.

Annotation format: A convenient feature of the tools exploited is that they all use a common input/output format: the NLP Annotation Format, or NAF (Fokkens et al., 2014). This is an XML format composed of layers, each of which represents an analysis step (POS-tagging, parsing, SRL, etc.). The *KafNafParserPy* library[8] was used to manage NAF annotations.

Integration: As mentioned above, each module in the IXA Pipeline toolkit offers a web service, which I used to integrate the tools in my workflow. The toolkit is in Java and my system is implemented in Python.

4.4.2 Domain model

The domain model contains actors and predicates, a complete list of which is provided in Appendix B.

[6] https://bitbucket.org/Josu/corefgraph
[7] https://github.com/newsreader/ixa-pipe-srl
[8] https://github.com/cltl/KafNafParserPy

Actors represent participants in international climate negotiations, such as countries and country groups. They are formalized as a map between actor variants and their DBpedia URI. There are some exceptions to this: Some actors correspond to generic roles in the negotiations, like *the chair* or *delegates*, and no DBpedia term is assigned to those. The list of actors in the model is on p. 227ff.

The **predicate set** contains lemmas for both verbal and nominal predicates. Some verbal predicates are neutral reporting verbs (e.g. *announce*). Other verbs express notions like support or opposition and agreement or disagreement (e.g. *criticize*). The verbs are contained in PropBank. The nominal predicates (e.g. *announcement*, *objection*) express similar notions to the verbs, and were selected from NomBank. The predicate type (i.e. *support*, *oppose* or *report*) was determined manually and specified as an attribute in the model.

To obtain the predicates in the model, a list with all PropBank and NomBank predicates was obtained with NLTK's API for those databases.[9] Among these predicates, a total of 180 reporting verbs and 150 reporting nouns were manually selected. The list is on p. 231ff.

A list of reporting verbs and nouns actually attested in the corpus was also compiled, to verify their coverage in the domain model. Less than 20 infrequent reporting verbs attested in the corpus are not covered in the model, since they are not found in PropBank. The main reporting nouns in the corpus are also covered in the model.

Complex reporting predicates, where the lexical meaning of the verb is carried by a noun following the verb rather than by the verb itself are not currently part of the model. Examples of such complex predicates are *express concern* or *express sympathy*.[10]

In our system, these expressions are currently detected as neutral reporting predicates like *express*, whereas it would be more informative to tag them as predicates of support (e.g. *express sympathy*) or opposition (e.g. *express concern*). Practically speaking this is not a major shortcoming, because among the propositions extracted, only 1.79% contain this type of expression (see footnote 34 on p. 191). However, the treatment of these complex predicates could be improved as future work.

Regarding **implementation**, the model is represented as XML files containing the actors, predicates and their attributes.

[9] The APIs are documented at `http://www.nltk.org/api/nltk.corpus.reader.html`. NLTK (*Natural Language Toolkit*) is a suite of Python libraries for some NLP tasks.

[10] The term *complex predicate* is not used in a technical sense, but just descriptively to refer to these predicates where it is the noun that carries the semantic weight, not the verb. The linguistics literature shows a debate, beyond our scope here, about different manifestations of this phenomenon, and how to define and name each of them, e.g. *complex predicate* vs. *light verb* vs. *serial verb* (Butt, 2010), or *semi light-verbs* (Bonial et al. (2016); Bosque (2001)).

Discussion: is an actor list necessary? As a final issue to address regarding the way negotiation participants can be identified in a corpus, note that, whereas we used a model with actors (besides predicates), a predefined actor list is not a necessary condition for applying our proposition extraction workflow. For the ENB corpus we used a list of actors because a list of participants at United Nations Climate Conferences was readily available.[11] However, for corpora where actors are not known beforehand, Entity Linking (EL) or Named Entity Recognition (NERC) could be applied; these technologies were reviewed in Ch. 1 and EL was applied extensively in this thesis (Ch. 3; Ch. 5.2.3, 5.3.3). Countries or organizations found with EL or NERC, that occur in the subject function or agent role of a domain predicate, could be considered as participants that make a statement in the negotiation.

In fact, in our workflow we also extracted propositions with actors that are not in the model, but were found as the agent of a reporting verb (see 4.4.3 for a description of the *agent* role). This makes up for participants not covered in our model, which focuses on countries and country groups, but does not contain other types of participants like non-governmental organizations. Having access to information about participants that are not countries or groups was appreciated by the domain-experts who gave feedback on our work, as such participants are less commonly studied (pp. 191 and 192).

4.4.3 Proposition extraction rules

This subsection describes the rules implemented to extract propositions, as well as two procedures required by the rules, i.e. treating negation and resolving pronominal anaphora.

Several extraction rules were implemented, that identify propositions based on a predicate's semantic roles, as per the PropBank and NomBank repositories. Recall from p. 32 that these knowledge bases have two very general semantic roles, *Arg0* and *Arg1*, which correspond respectively to the *agent* and *patient or theme*.[12] The agent causes an event or a change of state in another participant, and the patient or theme is causally affected by another participant. In a speech event, the agent is the speaker, and the theme is the message emitted. These two basic roles were used to create a general extraction rule and a rule for cases where a set of disagreeing actors appears in the sentence, as explained below.

[11] The site of United Nations Framework Convention on Climate Change provides country lists at http://unfccc.int/parties_and_observers/items/2704.php, besides a detailed list of non-governmental and intergovernmental organizations, which we did not include in the model for time reasons. Country groups are listed at http://unfccc.int/essential_background/convention/items/6343.php

[12] In role names, *Arg0* and *A0* are notational variants. *ArgM* and *AM* are notational variants as prefixes to indicate adjunct roles, e.g. *ArgM-TMP* or *AM-TMP* for a temporal adjunct.

Rule: Generic proposition

1 **foreach** predicate *p* **do**
2 Resolve negation
3 **foreach** pronoun *he, she it* in *p's* A0 argument **do**
4 Resolve pronominal anaphora
5 **foreach** actor-mention *am* in *p's* A0 argument **do**
6 Create a proposition $\langle am, p, point \rangle$, where *point* is a concatenation of *p*'s A1 arguments

FIGURE 4.2 – Generic proposition rule

Rule: Proposition for an opposing actor

1 **foreach** *opposed by* sequence *ob* **do**
2 Find proposition *main* for the sentence's main verb
3 **foreach** actor-mention *oam* in *ob* **do**
4 Create a proposition $\langle oam, \sim p_{main}, point_{main} \rangle$, where $\sim p_{main}$ is a negated form of *main*'s predicate, and $point_{main}$ is *main*'s negotiation point

FIGURE 4.3 – Proposition rule for opposing actors

General rule: Most of our domain predicates involve an agent and a message (i.e. a negotiation point) expressed by that agent in a given manner. Some predicates indicate support, some indicate opposition, and some are neutral. Actor mentions in these predicates' *Arg0* argument correspond to an actor expressing a message, and the predicates' *Arg1* argument often corresponds to the negotiation point addressed by the actor. Based on this, the generic proposition extraction rule is given in Figure 4.2. More specific rules to complement the generic one were also created (p. 83), as well as procedures to treat negation and pronominal anaphora, referred to in Fig. 4.2 and also described below.

Rule for opposing actors: Some arguments contain actors that oppose other actors in the sentence. For instance, in constructions like *China, opposed by the EU, preferred...*, or in the first example in Fig. 4.1. The agent of *opposed by* is the agent of a proposition that contradicts the main verb's proposition. The rule to treat such constructions is in Figure 4.3. The "*opposed by* sequences" referred to in line 1 of Fig. 4.3 are searched with regular expressions in each argument.

Negation is treated by finding ArgM-NEG roles related to the predicates, as well as negative lexical items (e.g. *not*, *cannot*, *lack*) in a window of two tokens preceding the predicate. This simple way of addressing negation cannot cover all cases. Consider a (hypothetical) sentence like *There was no lack of disagreement about...* Such double negations are not addressed.

Pronominal anaphora resolution was performed with rules based on dependency parsing and on the coreference chains output by the *CorefGraph* module (p. 79). Custom rules were required given non-standard pronoun-use: In the corpus, a personal pronoun (*he*, *his*) can be the anaphor for a country, and the gender corresponds to the gender of the delegate who represents that country in the negotiations. However, the inanimate pronoun *it* can also refer back to a country, as in common usage. Two rules were created to deal with this non-standard pronoun use, and anaphora resolution was limited to cases covered by these rules:

- An actor (country) in the subject position of the main verb of a sentence is taken as the antecedent of a sentence initial *he/she* in the following sentence.
- Antecedents for a pronoun (from CorefGraph's coreference chains) are only accepted if they are in the same sentence as the pronoun, or in the sentence immediately preceding the pronoun.

Other extraction rules, more specific than the ones above, were created. For instance, in order to make up for uncommon SRL analyses, actors were sometimes searched in adjunct roles like AM-MNR (manner adjunct) or AM-ADV (a general untyped role for adjuncts).

Non-canonical propositions: Propositions where the A0 argument does not contain any of the model actors are also output by the system. This is meant to make up for potential missing actors in the model, like special negotiation participants that do not correspond to a country or group of countries. Propositions with these "non-canonical" actors, that are not part of the model, receive a lower confidence score. Incomplete propositions, containing an actor and a predicate, but where the negotiation point has not been identified, are also output, but with a very low confidence score. Confidence scoring is described in 4.4.4.

Some rules were implemented to correct likely errors in non-canonical actors. For instance, if an actor starts with certain prepositions, or a punctuation mark, those sequences, unlikely to be part of the actor, are stripped from the actor. When these filtering rules apply, the confidence score is also lowered.

Implementation: The proposition extraction rules and the rest of procedures described in this subsection, besides confidence scoring, are implemented as Python code.

Criterion	Bonus	Penalty
Actor is a country or group in model	2	
Actor is a generic actor in model	1	
Actor not in model	0.5	
Proposition point is not empty	3	
Point has only one token		1
Point has two or three tokens		0.5
Anaphora resolution applied		1
Actor has atypical punctuation		1
Actor starts with preposition		1

TABLE 4.1 – Proposition confidence scoring: main criteria

Proposition (actor, predicate, point)			Score
Nepal	advocated	using a range of technologies	5
Alliance of Small Island States	added	that for vulnerable countries a 2ºC increase is too high	5
Gender CC–Women for Climate	said	adaptation requires hundreds of billions of dollars per year	4
Indigenous Peoples' Organizations	stressed	the "dire and urgent" situation of indigenous peoples facing climate impacts.	4
A COP/MOP decision	supporting	the continuation of the Kyoto protocol	3
A ten-country study	comparing	national policies on energy and CO2 emissions	3
Co-Chair Anaedu	reported	agreement	2
Canada	announced	*[empty]*	1.5
They	accept	*[empty]*	0

TABLE 4.2 – Confidence score examples for different propositions, in decreasing confidence order

4.4.4 Proposition confidence scoring

A confidence score is output for each proposition. This gives an estimate of the proposition's quality, in terms of how complete its elements are, and how informative it is expected to be. The score range is between between 0 and 5, taking 0.5 point steps (i.e. $\{0, 0.5, \ldots, 4.5, 5\}$).[13]

Scoring takes place according to a set of ordered rules, which include "bonuses" and penalties. Bonuses reflect proposition characteristics that make it a likely complete and informative proposition. Penalties reflect features that negatively impact the informativeness of the proposition. These features, as well as the values for each bonus and penalty, were determined empirically by inspecting propositions extracted from the corpus. The most

[13]The scoring scheme does not guarantee that scores will fall within the range, but out-of-range scores are extremely rare. Scores above 5 were not attested. About 0.25% of the propositions got a score below 0. They are propositions with an empty message and where anaphora resolution has applied, and they were ignored for display on the UI.

important criteria for bonus and penalties are summarized in Table 4.1. Example propositions for several confidence scores are given in Table 4.2, and more examples can be seen on the user interface to navigate the corpus (6.5 in Chapter 6),[14] which exploits the proposition extraction workflow just described.

As future work, it could be interesting to learn a scoring function on the basis of propositions annotated with confidence scores by a domain-expert, rather than using manually determined weights for the scoring criteria.

4.4.5 Discussion about the approach

The regular style and syntax in the corpus we are applying the system to raises the following question: Do we need information about syntactic and semantic structure provided by an NLP pipeline? Or would rules based on lexical and part-of-speech patterns, without access to structure, be sufficient to extract propositions in our corpus? I would argue that rules that don't exploit structural information could work, but they would have some disadvantages over the rules based on semantic role labeling and dependency parsing used here. In general, rules based on semantic roles (or on syntactic dependencies) will abstract away from word order variation, whereas if the rules are formalized on the basis of token or POS sequences only, extra rules will be needed to account for varying word orders. And rules that have access to structure should generalize better to corpora where word-order variety is larger than in our corpus. By using syntactic dependency parsing or SRL, the system has a better potential to generalize to other corpora.

Having access to syntactic structure is also helpful for resolving pronominal anaphora, where it is useful to know which is the subject for the main verb and for verbs in subordinate clauses.

As regards using SRL, a useful feature in the specific NLP pipeline we integrated is that, as mentioned on p. 79, its SRL module links corpus predicates and arguments at once to several knowledge-bases (KBs), including WordNet and several semantic frame databases (PropBank, NomBank and FrameNet). Semantic type information in these KBs could help towards further analysis of the propositions extracted, e.g. how many of them are a proposal, how many express criticism etc., abstracting away from the predicate instantiating those semantic types.

A final observation about using syntactic and semantic analysis vs. shallower methods is the following. For very large corpora, it would be relevant to consider that SRL may require longer processing times than syntactic

[14] http://apps.lattice.cnrs.fr/ie/uidev/

Technology	Module name	Corpus	Measure	Result	Reference
POS tagging	ixa-pipe-pos	PennTreebk. (WSJ)	word accuracy	96.88	Agerri et al. (2014)
Coreference	CorefGraph	OntoNotes 4.0 dev-auto	MUC B^3 $CEAF_m$ BLANC CONLL-F1	55.1 68.5 45.6 71.5 56.4	
Dependency parsing	ixa-pipe-srl	CoNLL 2009	LAS	89.88	Björkelund et al. (2010)
SRL			Semantic Labeled F1	80.90	

TABLE 4.3 – Results reported in the literature for the different components, for English, in the NLP pipeline integrated in our proposition extraction workflow.

dependency parsing, and than shallow methods which use no syntactic structure. Mesquita (2015) performed a comparison of processing times for several Open Relation Extraction pipelines which use shallow methods only, dependency parsing or SRL. For related discussion, see chapters 6 and 7 in Mesquita, (2015, esp. pp. 101ff. and 120ff.). Processing time was not critical for the ENB corpus we analyzed, as its size does not pose high demands in this respect. But for much larger corpora, processing time would be a factor to consider.

4.5 Intrinsic Evaluation, Results and Discussion

As described above, the system consists of two components. First, a pre-existing NLP pipeline, developed by other researchers, that I integrated in the system. Second, the modules we developed: a domain-model and a set of proposition extraction rules which exploit the NLP output, including procedures to identify negation and resolve pronominal anaphora. This section evaluates the different components, providing a discussion thereafter.

4.5.1 NLP pipeline evaluation

The IXA pipeline, created by Agerri et al. (2014), has been evaluated in work already cited. The results reported in the literature are reproduced on Table 4.3 for ease of reference.

POS tagging and coreference were evaluated in Agerri et al.'s article. The test-corpus for POS tagging was the Wall Street Journal subcorpus of the Penn Treebank. Coreference was evaluated on the dev-auto subset of OntoNotes 4.0. References defining the coreference evaluation measures on Table 4.3 can be found in Pradhan et al. (2011). Depending on the evaluation measure, the CorefGraph tool is between approx. 1.5 to 5 points behind Stanford's

dcoref, the Lee et al. (2013) system. In spite of better results for the latter tool, we chose CorefGraph for the convenience of having all NLP analyses coming from tools that use the same input/output format (see p 79). Besides, the results for CorefGraph are comparable to participants' results at the CoNLL 2011 coreference resolution task Pradhan et al. (2011).

The dependency parsing and SRL models in the IXA Pipeline come from Mate Tools,[15] and they were evaluated in Björkelund et al. (2010) using the corpus for the CoNLL 2009 task. The metrics (LAS and Semantic F1) were introduced on p. 34. The semantic F1 reported is for an end-to-end evaluation; the predicates are identified by the system, not provided in the test-set. If predicates are provided in the test-set, F1 goes up to 85.58. The results are comparable to the top system at the CoNLL 2009 shared task (Hajič et al., 2009).

4.5.2 Proposition extraction evaluation

Proposition extraction was evaluated against a reference set, as described below. Recall that the workflow involved, besides proposition extraction itself, resolving specific types of pronominal anaphora (p. 83). Anaphora resolution was not evaluated formally. Some informal comments about its performance can be found on p. 89.

4.5.2.1 Evaluation method

The domain model and rules to create domain-relevant propositions were evaluated with a manually annotated test set. The test-set characteristics, including the choices made to annotate certain phenomena, and the criteria to define a correct result, are outlined below.

Test-set description

The test set comprises 100 sentences (313 propositions) from the COP climate summit issues in the ENB corpus, that the system was built to analyze. The test set was intended to primarily contain sentences representing the corpus challenges, with negation, multiple actors, multiple predicates, and covering both verbal and nominal predicates. To this end, sentences containing these features were grepped in the corpus, and a random subset of the sentences returned by each grep was kept for analysis. An entirely random selection, without previously filtering the corpus as just mentioned, may have returned a test corpus that is too easy or not balanced across the sentence types to cover.

[15] Based on Mate Tools, the IXA pipeline added predicate and argument resolution towards the Predicate Matrix, mentioned on p. 79, besides a wrapper to integrate the Mate package in the rest of the pipeline, producing NAF-format outputs (p. 79).

The test-set is publicly available online, along with system outputs and the evaluation script to reproduce the evaluation results.[16]

Annotation process and conventions

Propositions were tagged by one annotator only. This is a weakness, and, as future work, it would be relevant to add at least one more annotator. This would allow to calculate inter-annotator agreement, and detect potential inconsistencies in the annotations.

In order to understand the way certain propositions were annotated, it's useful to mention several conventions we adopted. These are described in following.

The system does not annotate the noun as the predicate in constructions such as *express concern* (p. 80). It currently considers the verb as the predicate. In the reference set, there were two sentences with such a construction, and they were annotated with the verb as the predicate. This is meant to not penalize the system based on a feature the implementation of which is almost trivial, but which was not carried out for time reasons. As for the potential impact of this feature on results in general, as stated on p. 80, these constructions amount to approx. 1.8% of propositions in the corpus.

In some sentences, part of the negotiation point appears *topicalized*, i.e. in an adjunct at the beginning of the sentence, as in example (1) below.

(1) *On coastal adaptation technologies, AOSIS noted that financial and human resource limitations have stifled progress in adaptation and urged the development of long-term approaches under the FCCC.*

In this sentence, the phrase *[o]n coastal adaptation technologies* is part of the issue AOSIS is making a statement about. Such topicalized phrases were annotated as part of the negotiation point in the reference set. The system needs to extract them for a correct result.

Definition of a true positive

As regards the criteria for a true positive, a system output was considered correct if all of the proposition components (actor, predicate, negotiation point) match the reference exactly.

4.5.2.2 Results

Based on the notion of a correct output just described (i.e. exact match of all proposition components), precision, recall and F1 values are shown on Table 4.4.[17]

[16] https://sites.google.com/site/thesisrf/proposition-extraction-test-set
[17] These metrics were defined in footnote 15 on p. 20.

Corpus	P	R	F1
ENB-COP	68.7	69.3	69.0

TABLE 4.4 – Exact-match proposition extraction. Precision, Recall, F1 on the ENB-COP corpus.

Error Type	Count	% of Errors
only predicate wrong	2	2.1
only point wrong	63	64.95
both predicate & point wrong	32	32.99
all three elements wrong	1	1.05

TABLE 4.5 – Counts and proportion of errors per error-type, for propositions of shape ⟨*actor*, *predicate*, *point*⟩ in the COP issues of the Earth NB-COP corpus.

We consider our evaluation conservative, since propositions partially matching the reference receive no credit. It could have been possible to achieve higher scores by computing F1 over individual proposition elements, or by using the slot error rate metric (Makhoul et al., 1999). Our conservative measure avoids overestimating the system's value for our users. The proposition elements for which the system made an error in the ENB-COP corpus are summarized in Table 4.5.

Most errors took place identifying the proposition's negotiation point. One reason for these errors is the difficulty posed by our evaluation, which requires exact matches. Another reason is that it can be challenging to delimit an actor's negotiation point based on semantic roles. Generally the A1 argument corresponds to the negotiation point. However, as stated on p. 83, sometimes the A1 argument does not cover the entire point, and other arguments need to be added to complete it; we created rules for this. The work we reviewed in 4.3 uses dependencies to extract clauses (Van Atteveldt et al., 2017) and events in general (Schrodt et al., 2014). It could be tested whether syntactic dependency information is a more robust source to extract propositions in this corpus than SRL.

About 33% of the errors involve a wrongly identified predicate. These errors occur with some types of multi-predicate sentences.

Regarding the custom rules for **pronominal anaphora**, a thorough evaluation against an annotated test-set has not been performed. What can be stated based on informal evaluation is that accuracy was fine for the application's needs, but given that the rules only consider sentence initial *he*/*she* pronouns, coverage may be lacking.

4.5.3 Discussion

A question to consider would be whether these results, at an F1 of 69, are useful for domain-experts wishing to do research on the corpus. Besides the intrinsic evaluation, a qualitative evaluation with three political scientists was carried out, reported in Chapter 6 (p. 184ff). The experts performed a total of 40 queries on a user interface (p. 174) that allows navigating the propositions output by the workflow, besides the corpus sentences and documents. The error rate observed in expert testing was well within the error ratio in the intrinsic evaluation just reported, and proposition extraction errors did not have a major impact for the experts.

Regarding specific error types that may have an impact on research, an expert pointed out that he would prefer to see complex predicates like *express concern* analyzed correctly. The user interface provides a workaround to identify these complex predicates, but it would be useful future work to treat them correctly. It would be easy to treat at least the most common cases using our dictionary and rule based approach.

It also needs to be considered to what an extent the system can generalize to other corpora. Some elements in the proposition extraction rules are very generic, such as basing speaker and message detection on the agent and theme roles of reporting verbs, and this should be applicable to other corpora. However, other components in the rules may not generalize well; the elements that were added to treat the detailed way in which the corpus lists supporting and opposing actors may pose problems in new corpora. The system may generalize to extracting propositions in other reporting corpora showing limited lexical and syntactic variety. A possible example of such corpora is parliamentary proceedings. Salway et al. (2014) make a similar claim about the potential for generalizability in their grammar induction system, which they applied to the same corpus as we analyzed here.

4.6 Summary and Outlook

A detailed chapter summary follows. Possible future work was already outlined in the course of the chapter, and is also recapped within the summary.

A system was presented to extract propositions, i.e. ⟨*actor*, *predicate*, *negotiation point*⟩ triples, and applied to volume 12 of the *Earth Negotiations Bulletin*.[18] This corpus covers international negotiations on climate change. The corpus provides detailed information about which actors support and

[18]The original HTML corpus is at `http://enb.iisd.org/enb/vol12/`, we scraped the corpus into clean text and XML, as will be described in Chapter 6 (p. 165).

oppose statements by other actors, and requires extracting propositions based on both verbal and nominal predicates.

Generic Open Relation Extraction and Event Extraction tools would not optimally extract information as formulated in the corpus, which justified implementing a corpus-specific workflow. It was possible to use an approach based on dictionaries and rules, because, even if a lot of information can be contained in typical corpus sentences (Fig. 4.1), the syntax is largely regular and the set of reporting predicates to consider is limited.

The **proposition extraction system** we implemented exploits the output of an NLP pipeline, which provides syntactic dependencies, coreference chains and semantic role labeling (SRL). The system also relies on a domain-model containing negotiation actors (countries and groups) and reporting predicates representative for the corpus. The main source of information to identify propositions is the SRL output. Actors, and the negotiation points they address, are searched among the arguments assigned by SRL to predicates from the domain model. The predicate relating the actor and the message indicates the actor's attitude towards the negotiation point (i.e. something they favour or they are against).

Proposition extraction also involves treating **negated predicates**; this relied on SRL output and on surface cues. The actor in some of the propositions was identified via resolving specific types of **pronominal anaphora**, based on coreference chains provided by the NLP pipeline. Custom rules were created for this, since personal pronouns for animate targets can refer to countries in this corpus.

Finally, propositions receive a **confidence score** indicating the extent to which the proposition is expected to be informative and well-formed. These scores are now determined with a set of rules, it would be interesting future work to train a scoring function based on propositions annotated for confidence.

The propositions extracted can also contain actors not present in the domain model. This allows the system to extract information about participants which are less commonly studied than countries and country groupings. This was appreciated by the domain-experts who evaluated the system, as will be described in Chapter 6 (p. 193ff).

Regarding **evaluation**, proposition extraction was assessed against a manually annotated reference set containing approx. 300 propositions, for 100 sentences from the ENB corpus. Exact match for all three proposition elements (actor, predicate, point) was required for a correct result. F1 was 69. In the course of domain-expert evaluation (Chapter 6, p. 184ff), the error rate was not seen as detrimental to experts' research on the corpus. Note that the

evaluation process could be improved, since the reference set was created by one annotator only. It would be better to have other annotators' input, to verify how consistent the tagging is across annotators.

An **improvement to make** in proposition extraction is treating complex expressions like *express concern* or *express sympathy* better. Currently, the verb *express* is identified as the predicate. However, it would be more informative to tag the noun as the predicate, because it is the noun that indicates the actor's attitude towards the issues in the negotiation point, rather than the verb. These constructions appear in a minority of corpus sentences, but as future work it would be useful to improve their treatment.

As regards **generalizability** to other corpora, some aspects of the proposition extraction rules are generic, and they can be expected to identify propositions in other collections of documents. However, some of the rules are specific to the corpus and it may be problematic to apply them elsewhere. The corpus may generalize to similar reporting corpora, with limited syntactic and lexical variety.

Another aspect discussed was the **extent to which we need NLP** for the task on this type of corpus. In other words, whether rules based on lexical or part-of-speech patterns, rather than rules exploiting a syntactic or semantic analysis, are necessary to extract propositions from the corpus, given its regular syntax. I argued that it may be possible to proceed by applying lexical and POS-based patterns, ignoring syntactic and semantic information. However, rules based on syntactic constituents or semantic roles abstract away from word order variability and could generalize better to other corpora. Syntactic information is also useful for anaphora resolution.

The proposition extraction system described here was the basis of the application to navigate the corpus using proposition elements enriched with several metadata, besides full-text search, presented in Chapter 6.

Part III

Application Cases

Application Cases: Introduction

Part I surveyed Entity Linking and Relation Extraction, especially as relevant to Digital Humanities applications, and Part II discussed developments I undertook around those Natural Language Processing (NLP) technologies, to be able to apply them to the corpora analyzed in the thesis' three case studies. The object of Part III is these application case studies: User interfaces that allow us to navigate corpora with the help of structured annotations, automatically obtained thanks to Entity Linking (in Chapter 5) and Relation Extraction (in Chapter 6).

Chapter 5 covers two interfaces: The first one (p. 100) gives access to the manuscripts of British philosopher Jeremy Bentham, which had been transcribed by the Transcribe Bentham project. Besides Entity Linking, Keyphrase Extraction was performed on the corpus, and concept networks were created, that are used as corpus maps. The interface permits users to navigate the corpus via the maps or via text-search. Domain-experts' feedback on each type of map and on the interface overall was also documented.

The second interface in Chapter 5 (p. 135) gives access to a subset of the PoliInformatics corpus, about the American financial crisis of 2007-8. The interface shows the combined results of several Entity Linking tools, giving information about the quality of each result, so that users can choose the concepts to model the corpus with. Besides, the corpus is represented as a concept network, as in the Bentham interface.

In the Bentham and PoliInformatics interfaces, information about a concept's role in the corpus is obtained from its position in a corpus map. The third interface (Chapter 6) gives access to negotiating parties' statements in the *Earth Negotiations Bulletin* (ENB), which documents international Climate Change conferences. Entity Linking was applied to identify speakers and negotiation issues. However, co-occurrence between actors and issues is missing one piece of information, i.e. where the actor stands with respect to the issue. Relation Extraction was applied, to identify whether the verbs and nouns mediating between actors and their statements indicate support or opposition. The interface can filter the corpus according to this information, e.g. displaying concepts found in statements where a given actor voices

opposition, or concepts found in statements over which a given pair of actors agree.

In summary, the Bentham and PoliInformatics case studies apply public domain Entity Linking systems (with a varying degree of customization) to annotate concepts in the corpus, representing its content as concept networks. A concept's role in the corpus is characterized by its location in the network. By contrast, for the ENB interface, relation extraction was applied to the documents, to make explicit where the actor stands with respect to an issue (in a relation of support or opposition). These NLP-based corpus annotations (concepts and relations) are exploited to navigate the corpus. Domain-expert input about the usefulness of these navigation workflows was documented and is discussed.

Chapter 5

Concept-based Corpus Navigation: Bentham's Manuscripts and PoliInformatics

5.1 Introduction

This chapter discusses two application cases that rely on concept annotation, and the user interfaces created to navigate the corpus using the concepts extracted. A concept is defined broadly as a relevant term in the corpus, that can contribute to gain an overview of the corpus content. A concept can be part of an ontology like DBpedia, in which case it will be expressed in the corpus by different, potentially diverse, mentions referring to the concept. As an example, DBpedia concept *English_Law*[1] may be referred to with phrases like *laws in England*, *the law of England*, or *English Law*. Core concepts in the corpora were identified with a technology called Entity Linking (EL), a review of which was provided in Chapter 1. EL finds in a corpus mentions to concepts in a knowledge base (KB), annotating the mention with the concept. This is used to relate passages mentioning the same concept to each other.

Section 5.2 discusses the first application: A corpus navigation interface for about 4.7 million words of the manuscripts of Jeremy Bentham (1748–1832), the British philosopher and social reformer, who wrote extensively on ethics, law, and politics. The knowledge base we performed Entity Linking with is DBpedia (Auer et al., 2007), which is an encyclopedic KB reflecting the content of Wikipedia. The question arises whether a general domain knowledge source developed in the early 2000s is relevant to annotate a specialized corpus from the 18th and 19th centuries; difficulties will be discussed below. Besides, as some terms specific to Bentham's thought are not part of DBpedia, in order to complement DBpedia annotations we performed keyphrase extraction on the corpus. This technology finds phrases (generally noun phrases) relevant to characterize a subset of the corpus documents, thus

[1] http://dbpedia.org/page/English_law

giving an overview of important notions in the corpus. Both DBpedia annotations, and the terms obtained via keyphrase extraction, were used to create different types of searchable and navigable concept networks of the corpus, used as corpus maps. Domain-expert feedback on each type of map was collected and is discussed below.

Section 5.3 presents the second application: A user interface to navigate the PoliInformatics corpus, about the American Financial Crisis of 2007-8. The subset of the corpus I worked with consists in a report by Congress on the causes of crisis, and interview transcripts by the Federal Crisis Inquiry Commission, which conducted hearings with individuals who played a major role in the crisis. The interface integrates the results of several Entity Linking (EL) tools, providing information to a user about the quality of each annotation, so that users can decide which annotations to keep for their work on the corpus. The interface also offers the option to select results automatically, based on a weighted voting procedure developed for this application, which was discussed in Chapter 3. Finally, EL results were used to represent the corpus as a concept map.

5.2 Bentham's Manuscripts

This section presents our user interface to navigate Jeremy Bentham's manuscripts, besides providing the relevant background about the corpus creation effort led by University College London, from which we obtained digital versions of the text of the manuscripts. In 5.2.1, the corpus is described, giving details about the document sample we selected for analysis, and preprocessing steps. This is followed in 5.2.2 by an overview of previous tools to navigate the corpus, and of previous work analyzing the content of the transcripts. In 5.2.3, our concept detection procedures based on Entity Linking and keyphrase extraction are introduced. Our user interface to navigate the corpus is presented in 5.2.4, and a domain-expert evaluation of the interface is provided in 5.2.5. Finally, a detailed summary of the work carried out can be found in 5.2.6.

5.2.1 Corpus Description

Jeremy Bentham (1748–1832) 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 et al., 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,[2] 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 et al., 2014b i.a.), also coordinated by UCL. We had access to a large number of these transcripts within a collaboration with UCL, and our user interface allows to navigate a subset of these transcripts.

5.2.1.1 Structure of the corpus and TEI encoding

Here we describe the structure of the complete corpus, from which we selected a sample (5.2.1.2) 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.[3] As of January 2017, 17,513 folios had been transcribed (45.17% of the total amount of then digitized material).[4] 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 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:[5]

- **Text Sheets** or 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.

[2] https://www.ucl.ac.uk/Bentham-Project
[3] For TEI, see http://www.tei-c.org/index.xml
[4] The Transcription Desk shows current progress: http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham
[5] http://www.benthampapers.ucl.ac.uk/help.aspx?subject=category

- **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 5.3.

Besides the document type, Bentham Project scholars recorded several **metadata** for each folio,[6] 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.

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.[7] 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 5.1 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. To our knowledge, 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.[8]

5.2.1.2 Corpus sample in our study and corpus preprocessing

As a collaboration between the LATTICE Lab and UCL's Centre for Digital Humanities, in 2015 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.

At first, we did not have access to the metadata described in 5.2.1.1 above, which include **document dates**. Since we wanted to analyze the temporal evolution of the corpus content, we needed to assign a date (a year) to each file. We used a simple heuristic to assign years: If the first sequence

[6] http://www.benthampapers.ucl.ac.uk/search.aspx?formtype=advanced
[7] http://www.transcribe-bentham.da.ulcc.ac.uk/td/Help:Transcription_Input_Form#Core_Guidelines
[8] http://www.transcribe-bentham.da.ulcc.ac.uk/td/Help:Transcription_Guidelines#Supplementary_Guidelines

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'.

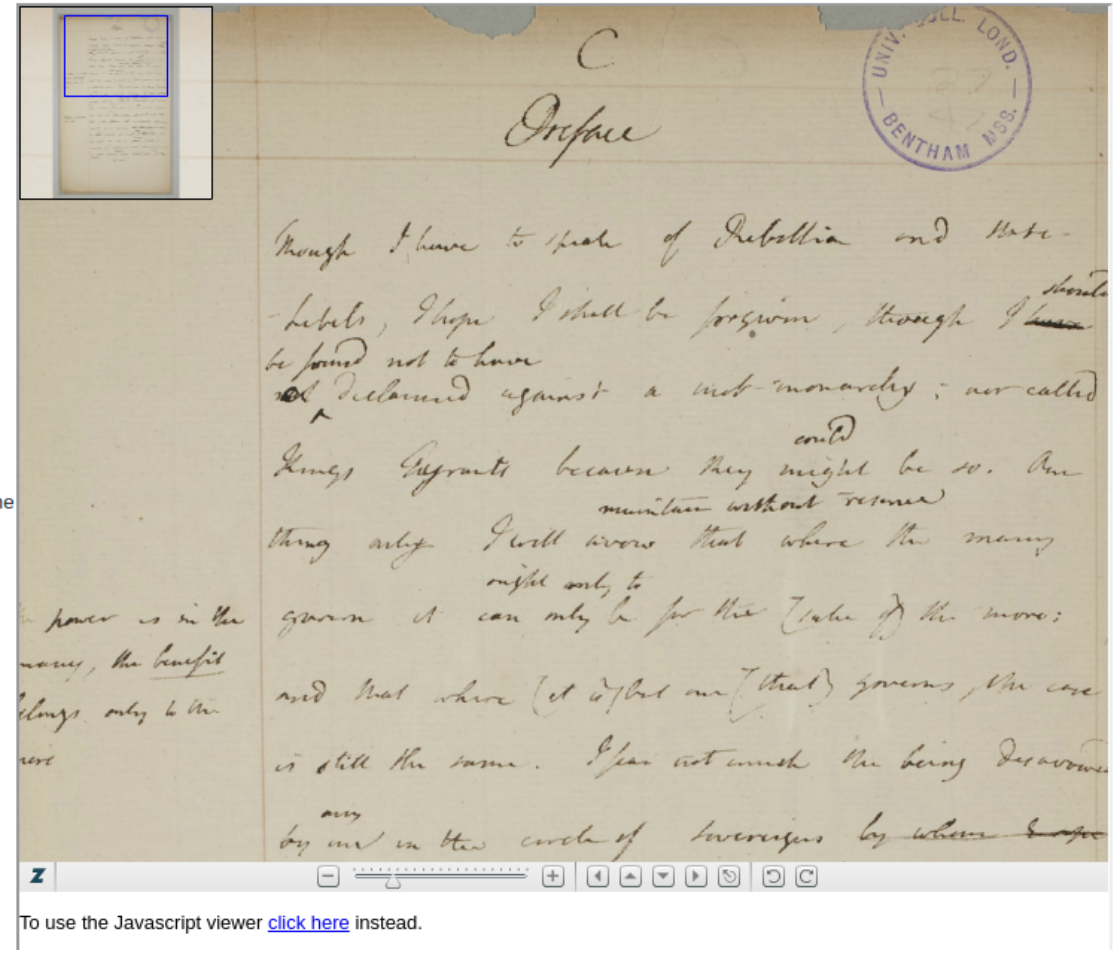


FIGURE 5.1 – UCL Transcribe Bentham interface. A manuscript digitized by UCL is on the right, which shows some of the characteristics encoded in TEI by volunteer transcribers, like added and deleted text, and marginal notes. The left pane shows an HTML rendering of the TEI, reflecting those annotations: Added superscribed text, crossed-out deleted text, and boxes for marginal notes. The screenshot was taken from http://www.transcribe-bentham.da.ulcc.ac.uk/td/JB/027/047/001

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 recently (Pearson's $r = 0.976$).[9] 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).[10] This classifies text against pre-trained models for many languages, using features 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.[11] In consequence, our content evolution analyses yield clearer results from 1800 onwards (p. 122). Now that all document dates established by Bentham Project scholars are available to us, as future work, it would be interesting to enrich our sample with more documents prior to 1800 and obtain new analyses. 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 Figure 5.2. The sample's distribution of pages across main document types (5.2.1.1) is in Figure 5.3.

Regarding **text preprocessing** performed before feeding text to the Natural Language Processing (NLP) tools, recall from p. 102 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 (see 5.2.1.1).

A TEI document's two outermost elements are a `teiHeader` tag, for meta-data, and a `text` tag, for the document content. Our preprocessing ignored the `teiHeader` tag, as it only encodes information relevant internally at

[9] As in https://docs.scipy.org/doc/numpy-1.10.1/reference/generated/numpy.corrcoef.html

[10] Using the default options: http://search.cpan.org/~ambs/Lingua-Identify-0.56/lib/Lingua/Identify.pm#langof

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

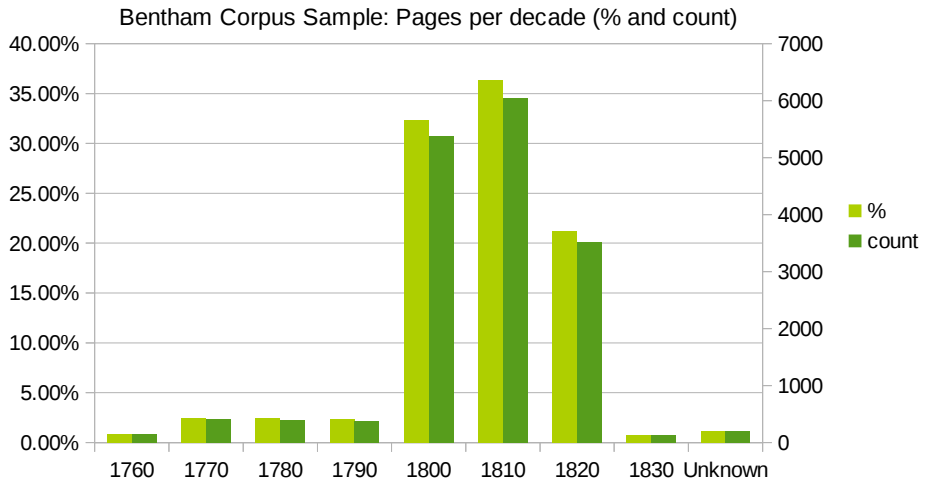


FIGURE 5.2 – Percentage and count of pages per decade in our sample of 16,618 pages from the Transcribe Bentham corpus. Decades correspond to dates assigned to the documents by the Bentham Project.

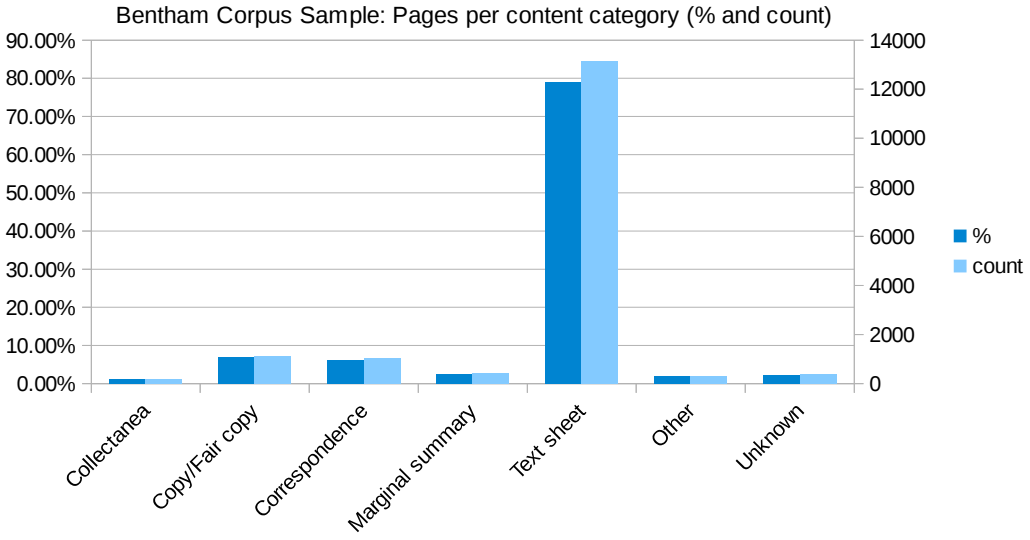


FIGURE 5.3 – Distribution of pages (count and percentage) in our 16,618 page sample across the main content categories described in 5.2.1.1, besides infrequent categories grouped as *Other*.

Transcribe Bentham—metadata mentioned above like dates or document types are available in a database and were not rendered in these TEI. To obtain unannotated text, our preprocessing consisted in the following steps, for the content under the `text` tag:

1. Removing text inside deletion tags (`del`), and the tag itself.
2. Protecting paragraph marks (`<p>`, `</p>`) so that paragraphs could be respected in further processing.
3. Replacing other tags (e.g. `add` tags, for additions) with spaces. While the tag itself was removed, its text content was kept.
4. Replacing repeated spaces with a single space.

In short: 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.

As regards the **file formats** we represented the corpus with, besides a plain text version of each file, sent to the Entity Linking and Keyphrase Extraction tools, two other versions were created:

1. Solr XML:[12] To index the content in the Solr search server (5.2.4.2).[13]
2. CorText CSV: To import the corpus into the CorText Manager text analysis and visualization platform (5.2.3.2).[14]

Besides a `text` field, which was populated according to the preprocessing described above, both formats require a `title` field. This was created from the first characters in the document body, as an overall title or heading for each page was generally not available in the TEI files.[15] A `date` field is required to perform chronological analyses: Date-based searches in Solr, and diachronic corpus maps in CorText.

5.2.2 Prior Analyses of 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 (5.2.1.1) and were transcribing the manuscripts prior to the crowdsourcing initiative. Several works analyze this corpus creation process (Causer et al., 2012; Causer et al., 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 et al., 2014b; Toselli et al., 2015).

[12] `https://wiki.apache.org/solr/UpdateXmlMessages`
[13] `https://lucene.apache.org/solr/`, version 4.9.0 was used.
[14] `https://docs.cortext.net/`, see *Data Processing/Upload Corpus*
[15] This is not surprising since some pages would not have an overall heading. Besides, such information is recoverable from a database with metadata for the whole corpus (see 5.2.1.1), which was not accessible to us when we indexed the documents.

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 some corpus navigation functions for those outputs. The Bentham Papers Database allows searching in metadata fields (5.2.1.1), returning matching records, and providing a link to the record's transcript when available (Figure 5.4).[16] 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 (Figure 5.5).[17]

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 et al. (2014a, Section 4) 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* (private interest, rather than common good, leading rulers' actions) and the timeline of his conversion to political radicalism.[18] The Bentham Project and Transcribe Bentham are authors of a large number of significant outputs, what I am mentioning in this paragraph is my understanding of 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.[19]

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

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

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

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

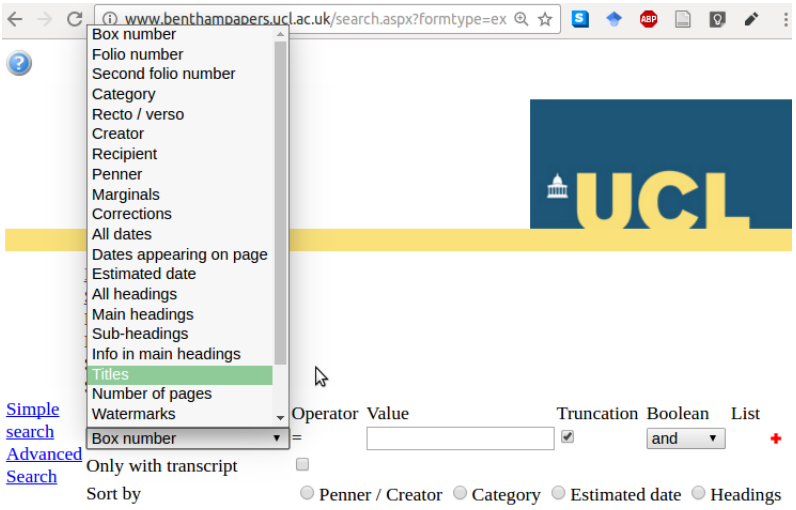


FIGURE 5.4 – UCL Bentham Papers Database, an early platform predating Transcribe Bentham, for metadata-based search on Bentham's manuscripts. The screenshot shows the search interface with a dropdown of some of the searchable metadata fields.
http://www.benthampapers.ucl.ac.uk/

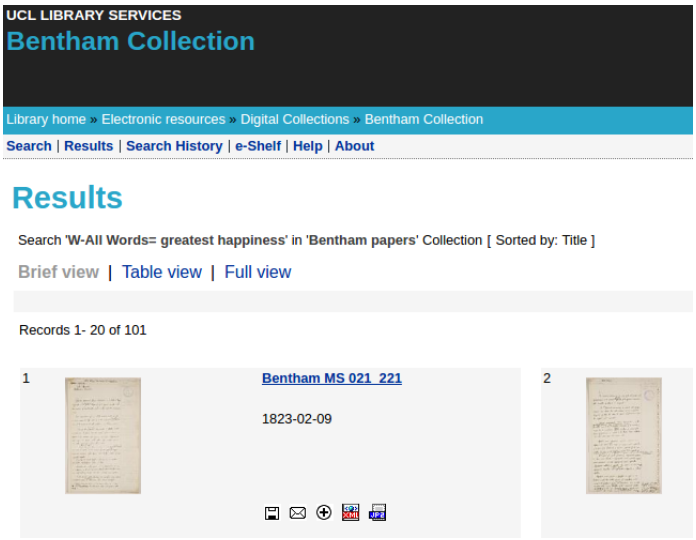


FIGURE 5.5 – UCL Libraries Digital Collections created an interface for metadata-based, or, where the transcript is available, full-text search. It integrates Transcribe Bentham outputs, returning a manuscript's image and its TEI transcript. The screenshot shows the first result for query *greatest happiness* (a Benthamic concept).
https://www.ucl.ac.uk/library/digital-collections/collections/bentham

5.2.3 Corpus Cartography based on Entity Linking and Keyphrase Extraction

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 in the following pages.

5.2.3.1 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 of a set of expressions to analyze the corpus, based on the output of each technology.

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 way of referring to that term in the corpus. For instance, textual mentions *amount* and *quantity* will be mapped to the *Quantity* concept in DBpedia.[20]

In our EL workflow, we are targeting KB terms that correspond to conceptual mentions as well as terms expressed by named entities. Conceptual mentions are usually noun phrases. Named entities were defined on p. 16; they are lexical sequences corresponding to a set of predefined types (like people,

[20] 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*

places and organizations) and are often proper nouns. When Entity Linking targets conceptual information, some scholars speak of Wikification instead of EL. In this thesis, both terms are used interchangeably, as argued on p. 16.

Tool: For EL, we used DBpedia Spotlight (Daiber et al., 2013; Mendes et al., 2011). 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 did pose 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 was described in Chapter 3 (p. 62). In summary, 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.

Annotation selection: After preprocessing the corpus sample as described in 5.2.1.2 above, it was sent to Spotlight's web service, using default **settings**. 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*).

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**, as 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. 1000 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 of EL (pp. 18, 110) was the main source of information to annotate the corpus, rather than the full results of linking to DBpedia. DBpedia concept structure 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 p. 110 that annotations whose confidence was below 0.1, and variants whose corpus frequency was below 100 had been removed automatically. A list of 285 $\langle variant\ set, label \rangle$ 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 (5.2.3.2), and corpus maps based on them (5.2.3.3). Appendix A shows the final list after manual filtering (p. 210), and the items filtered out (p. 212).

Keyphrase Extraction

Keyphrase extraction (Kim et al., 2010; Turney, 2000) 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 DH applications, it is sometimes used to give an overview of a corpus (e.g. G. Moretti et al., 2016; Rayson, 2008), which is the use intended here.

Tool: Keyphrase Extraction was performed with Yatea (Aubin et al., 2006), a rule-based keyphrase extractor.[21] It takes as its input part-of-speech tagged text in Treetagger output format.[22] Part-of-speech tagging (PoS-tagging) was done with Treetagger (Schmid, 1994).[22] 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.

Annotation Selection: 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. 1950 terms. From these, the most frequent 250 terms were used to create corpus maps (5.2.3.2 and 5.2.3.3). The list of terms is shown in the Appendix (p. 213).

The minimum frequency selected for keyphrases (10) is smaller than the one for Entity Linking mentions (set at 100, see p. 110). Keyphrases are generally multi-token and will consequently reach a smaller frequency than

[21] http://search.cpan.org/~thhamon/Lingua-YaTeA/lib/Lingua/YaTeA.pm
[22] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

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.

Other discussions: Note that, since Yatea only provides keyphrase frequency at document level, we computed corpus frequencies, besides each keyphrase's tf-idf weight (Salton et al., 1983).[23] This weight, commonly used in information retrieval (Manning et al., 2008, Chapter 6), favours phrases that are frequent in a subset of the corpus documents, but infrequent overall in the corpus. It is an indication whether the keyphrase likely refers to an important concept in the corpus, addressed in a subset of its documents, unlike phrases containing very general words; very general expressions (e.g. the phrase *another thing*) are likely to be used homogeneously in all documents, and unlikely to represent corpus concepts. The keyphrase ranking by corpus frequency was close to the ranking by tf-idf.

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., 2016b). A newer tool that may require less manual cleanup of results is Keyphrase Digger (KD), by G. Moretti et al. (2015).[24]. 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 Task 5 keyphrase extraction dataset (Kim et al., 2010), with competitive results (ranking 2nd to 4th depending on the evaluation mode). We are not aware of results for Yatea on a public benchmark, and we did not obtain such results ourselves in the interest of time, and since the plan was to manually verify its results on the Bentham corpus.

5.2.3.2 Lexical Clustering and Network Creation

The lists of terms described in 5.2.3, including their variants in the case of terms derived from Entity Linking, were clustered with the CorText

[23] The acronym means *term frequency – inverse document frequency*. The tf-idf for a term t in a document d was defined thus:
$tf\text{-}idf = tf \cdot idf$, where $idf = log \frac{|documents\ in\ corpus|}{|documents\ containing\ the\ term|}$, and two different variants for tf were used: (a) $tf = frequency\ of\ t\ in\ d$ (Salton et al., 1983, Section 3B), and (b) $tf = 1 + log(frequency\ of\ t\ in\ d)\ or\ 0\ if\ frequency\ of\ t\ in\ d\ is\ 0$ (Manning et al., 2008, p. 127)

[24] https://dh.fbk.eu/technologies/kd

Manager platform.[25] 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,[26] which can then be visualized with network analysis tools like Gephi.[27]

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 the fields `text`, `title`, `date` and `decade`. More details about corpus creation were given on p. 106.

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.[28] 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 (5.2.5) 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.

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

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

[27] 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.

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

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 et al., 2005), and, for reasons discussed there (p. 443ff.), it is asymmetric.[29] 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 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. Visualizations for each type of network are provided below (p. 120ff.).

Communities are computed on the network, i.e. groups of highly interconnected nodes. The algorithm is Louvain (Blondel et al., 2008). In visualization (5.2.3.3), 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 on p. 114, 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 5.7.

Networks and other visualizations that allow examining the **temporal evolution** of the corpus can also be created with the platform. Whereas the platform has more complex functions to analyze the evolution of lexical

[29] 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.

cluster structure,[30] 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 in CorText. 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 on p. 122, and examples are shown in Figure 5.9.

5.2.3.3 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 (Jacomy et al., 2014). 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 linked to from nodes in two 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.[26] Besides the positions of nodes and edges, other network attributes encoded in the GEXF, that are exploited for visualization, are the following:

1. **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.[31]
2. **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.

Besides the 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

[30] See (Rule et al., 2015) and https://docs.cortext.net/analysis-mapping-heterogeneous-networks/mapping-dynamical-analysis-options/

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

with other navigation tools like a search index. This UI is discussed in following.

5.2.4 User Interface: Corpus Navigation via Concept Networks

The user interface[32] (UI) gives access to our sample of the Transcribe Bentham corpus (p. 102), via a full text search index (5.2.4.2) and thanks to a navigable rendering (5.2.4.3) of the concept networks which were described in 5.2.3 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 (5.2.4.3). The goal of the UI is to provide an overview of the corpus. As a first requirement, the networks should reflect a 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 starting on p. 124 for discussion).

Our UI complements prior platforms to navigate the corpus, mentioned on p. 106: The Bentham Papers Database (Figure 5.4) and UCL Library's platform (Figure 5.5). 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 (see p. 117). 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.

5.2.4.1 User Interface Structure

The default view of the UI is the search index, displayed on Figure 5.6. 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 (see 5.2.3.1). 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.

5.2.4.2 Search Interface

The search backend is Solr (Lucene-based),[33] 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

[32] http://apps.lattice.cnrs.fr/bentham
[33] https://lucene.apache.org/solr/

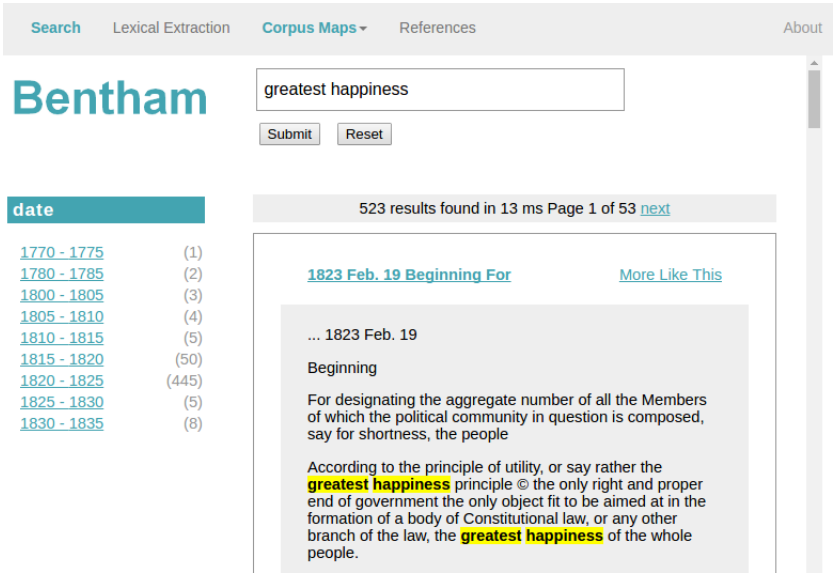


FIGURE 5.6 – Structure of our User Interface to navigate transcripts of Bentham's manuscripts using concept networks and full text search. The screenshot displays the search index, showing results for query *greatest happiness*. Date facets are available on the left, to filter results per 5-year period (the number in parenthesis indicates records returned per period). Concept maps are accessible from the `Corpus Maps` menu.

or document body only. Logical operators, proximity search and fuzzy matching are also possible.[34] For relevance scoring, the main factors used by Solr are tf-idf weighting with raw term-frequency counts,[23] 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).[35] 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.[36] All these features are visible on Figure 5.6.

5.2.4.3 Navigable Corpus Maps

Recall that, following a lexical extraction on the corpus (5.2.3.1), based on concept mentions obtained via Entity Linking, and based on keyphrase extraction, concept networks were created with the CorText platform (5.2.3.2). As will be described in the paragraphs below, the networks were then exported in GEXF format, and were rendered navigable with two libraries,

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

[35] 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`

[36] How these years were established was discussed on p. 102.

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, for reasons discussed on p. 114. Two navigable versions of each network were created, using two tools. The first tool is the Gephi Sigma JS exporter plugin[37] and the second one is the TinawebJS project explorer.[38] Both tools rely on the Sigma JS graph drawing library[39]

The navigable networks obtained with both tools allow us to search for a node in the network. Another point in common is that, 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.

There are some differences in the way graphs are navigated with each tool, that in my opinion make the tools complementary. On the TinawebJS rendering, **searching** for a term returns the list of matching nodes and highlights those nodes on the network directly. With the SigmaJS exporter, a search does return the list of matching nodes, but they are not highlighted on the network. So for getting a **global overview** of a term's position in the corpus, I find the Tinaweb representation more useful.

Another difference that makes TinawebJS helpful for a network overview is that it provides a **legend** matching community colours with community labels. The legend provided by the Sigma JS exporter plugin is not so user-friendly, as it lacks spelled-out labels, using an integer label instead. Figure 5.7 shows how the TinawebJS map provides an overview of the way the notion of power is addressed in the corpus, displaying nodes mentioning *power* distributed over several communities.

A third difference is that, with the SigmaJS exporter, clicking on a network node (or on a node list returned by a node search) isolates the node and its immediate neighbours; the rest of the network is hidden, and can be displayed again by deselecting the node. On the TinawebJS rendering, clicking on a node highlights the node by greying out or decreasing saturation for all other nodes. I find that, depending on the network colours, this highlighting is ineffective. For this reason, to get a quick impression of the **local context** of a node, and to follow the **path between any two nodes**, I find the SigmaJS exporter plugin easier to work with.

[37] `https://marketplace.gephi.org/plugin/sigmajs-exporter/` The tool was created at the Oxford Internet Institute.
[38] `https://github.com/moma/ProjectExplorer` The tool was created at two CNRS labs, the ISC-PIF and the CAMS.
[39] `http://sigmajs.org/`

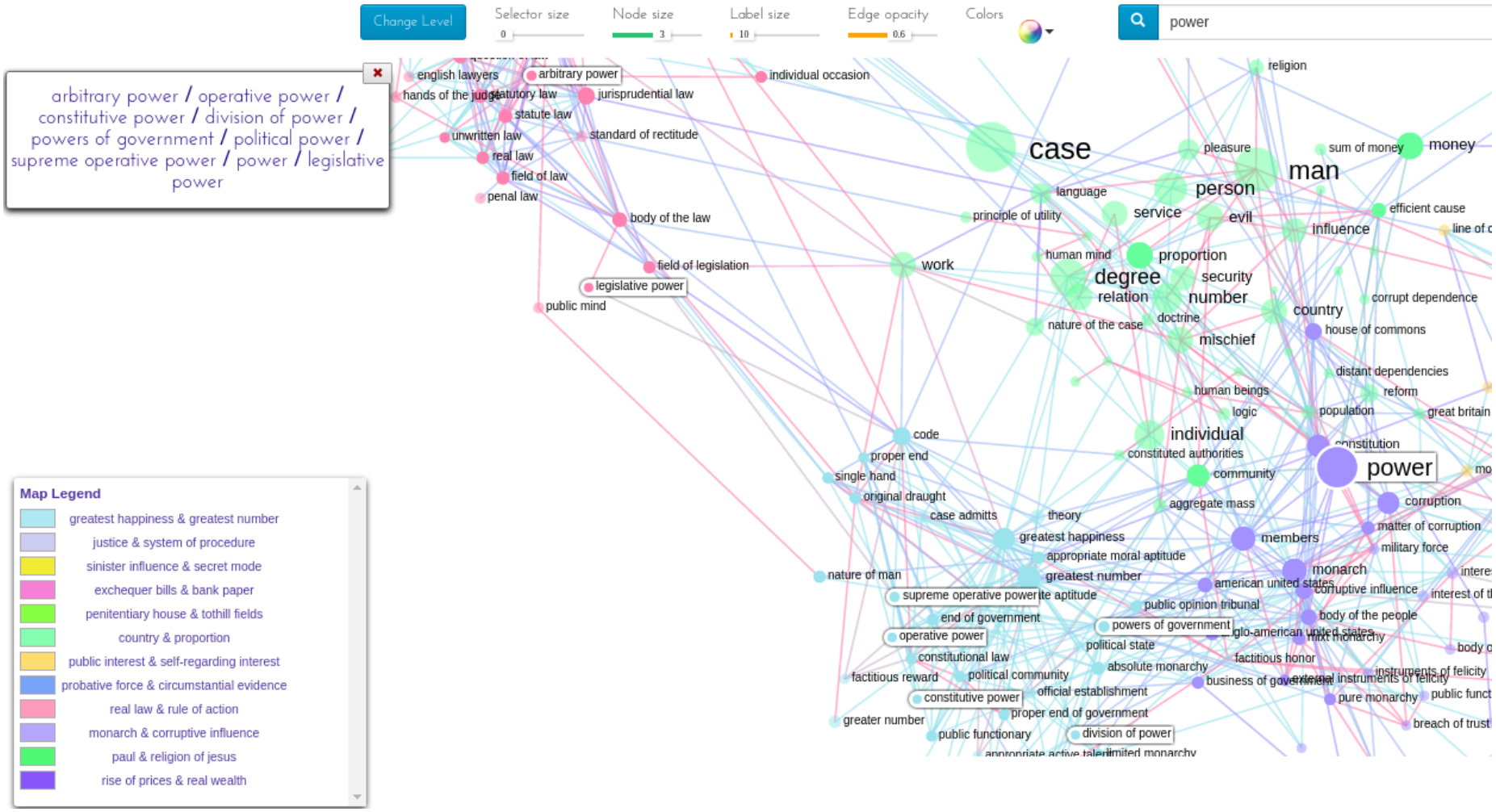


FIGURE 5.7 – TinawebJS rendering of the 250 keyphrase network. Nodes matching *power* have been searched for, and the tool highlights them in the network. Terms belonging to different communities are highlighted, providing an overview of how power is discussed in the corpus. The communities are identified 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.

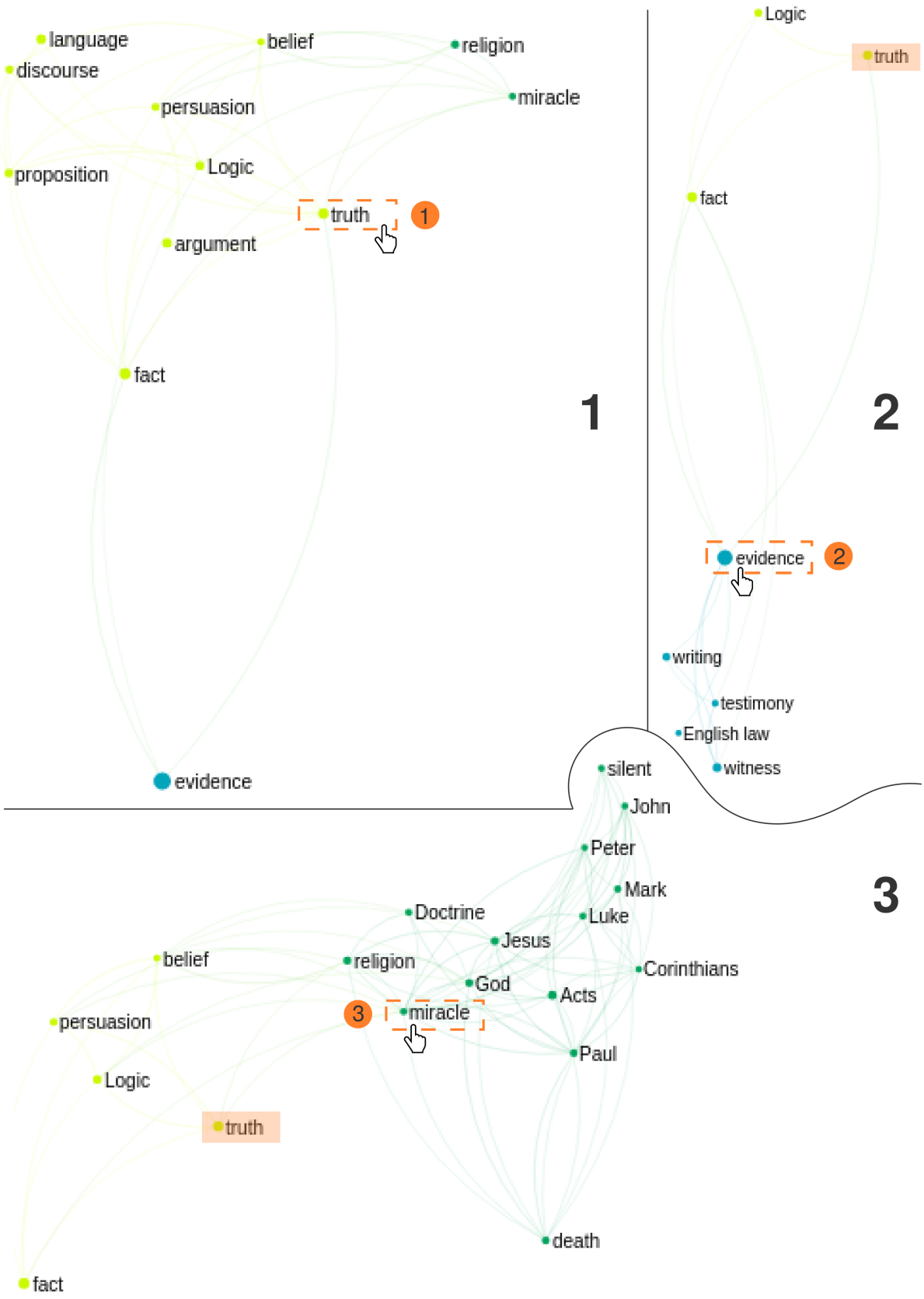


FIGURE 5.8 – Corpus navigation and local context exploration by successively selecting neighbours, as can be tested on http://apps.lattice.cnrs.fr/bentham/bentham-js-more.html and the other maps rendered with the Gephi Sigma JS exporter. **(1)** Node *truth* is selected. It has neighbours in the bright green *discourse & proposition* cluster, and is linked via *evidence* to the blue *court & procedure* cluster. It also has neighbours in the dark-green religion-related cluster. **(2)** Selecting *evidence* we see the closest nodes to *truth* in the judicature-related blue cluster. **(3)** Selecting *miracle* we see the closest nodes to *truth* in the religion cluster. In summary: Starting from a given node like *truth*, we can navigate the network by sequentially clicking nodes from its neighbour-set and the node-set linked to each neighbour. This might suggest connections, not previously known to a researcher, between corpus terms.

A way for an interested reader to verify this would be by testing one of the UI's Gephi exports.[40] For instance, by looking at the neighbours of the concept *truth* in the 250-term Entity-Linking based map[41] (see also Figure 5.8). This would show different ways in which truth is discussed in the corpus: many of its neighbours are related to human reasoning, but it also links via *evidence* to a judicature-related cluster, and finally some other of its neighbours come from a religion cluster. Figure 5.8 shows the nodes just mentioned and their neighbours, although using the UI interactively is a clearer way to access this information.[41]

In fact, with the Gephi Sigma JS export, a way to navigate the network is by sequentially clicking on nodes from the node-sets linking to each neighbour of a given concept. This might reveal connections a user had not thought of in a serendipitous manner.

Note that in these navigable networks the corpus context for a node cannot be accessed directly from the network. This would have required additional development on the Gephi exporter plugin and on TinawebJS, which was not carried out for time reasons. Currently, users can search manually for the context of a node or a node combination, by going to the `Search` tab on the UI. Linking the corpus maps with the search index would be useful future work.

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. 115. In the UI screenshots in Figure 5.9, the areas shaded in red show how in the 1810s the manuscripts focused on human reasoning and religion, i.e. the communities labeled as *discourse & proposition* and *God & Jesus*, whereas in the 1820s, the manuscripts turn their attention to the *Constitution & government* cluster. A Bentham expert's feedback on the heatmaps can be found on p. 128 in the UI evaluation section.

In the case of heatmaps, on the UI we rendered them as image files based on the PDF files generated by CorText. Navigability is restricted to choosing the decade by clicking on the top bar, and there is no search function. This leaves room for improvement, but was implemented this way since, by using images, it took much less time to make the site compatible with all major browsers than by using searchable formats.[42]

[40] Any of the maps mentioning *Gephi*, in the `Corpus Maps` menu.

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

[42] As a final remark on the heatmaps, one of our users asked if it is meaningful that nodes are represented by triangles in them, rather than as circles like in the navigable maps. This does not have a special meaning; it is just the way CorText renders nodes on PDF files.

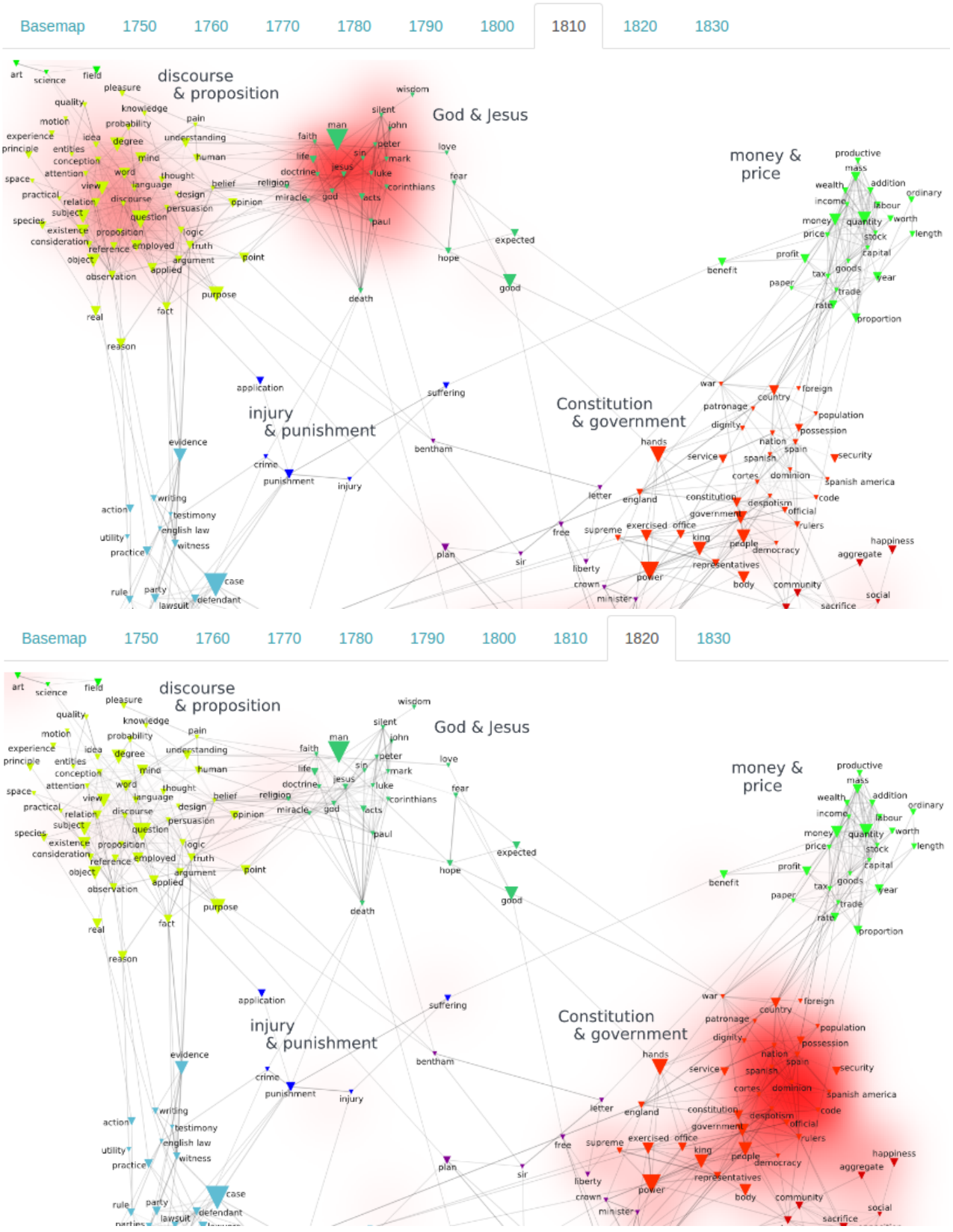


FIGURE 5.9 – Heatmaps per decade, based on concept mentions obtained with Entity Linking: Red shading becomes darker as saliency of the corpus area in the decade increases. **Top:** In the **1810s**, two areas the corpus focuses on are the *discourse & proposition* and *God & Jesus* clusters. **Bottom:** 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. 128): 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.

In **summary**, the UI shows navigable maps rendered with the Gephi Sigma JS exporter and TinawebJS, and heatmaps per decade rendered as images. A search interface with boolean and proximity search and year facets is also available. The Gephi exports are useful to examine a node's local context and navigate the path between nodes. The TinawebJS rendering is more useful for a global overview, given the better search functions and legend. It would be practical future work to combine the best aspects of both of these tools. Finally, the corpus context for a node is not directly accessible from the corpus maps. It would be useful future work to link the maps with the search index; currently the user accesses the corpus map and search index separately. The maps were intended to give an overview of the corpus, and ideally to suggest new research ideas to domain experts. The next section reports on a UI evaluation with domain experts, to assess to what an extent these goals were fulfilled, and how the UI is perceived by experts.

5.2.5 User Interface Evaluation with Experts

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

5.2.5.1 Introduction and basic evaluation data

Feedback was gathered from one Bentham scholar and one DH researcher. 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 from University College London, where she is the Director of the UCL Centre for Digital Humanities and Professor of Digital Humanities at the Department of Information Studies. She has also published research within the Transcribe Bentham project.

5.2.5.2 Expected outcomes

The evaluation was informal, with only one domain-expert and one DH researcher. The task 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.[43] Second, keyphrases salient in the corpus, regardless whether they are covered by DBpedia or not. My intuitive expectation was that DBpedia concept mentions would be perceived by the domain-expert as clearer for a non-expert public. Conversely, I expected the terms obtained via keyphrase extraction to be more informative for the domain-expert than the DBpedia mentions, since 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.

5.2.5.3 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 (see 5.2.3 above). 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 later transcribed (non-verbatim) by myself.

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

- **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,[44] *degree* and *aptitude* connect the discourse-related purple cluster and the

[43] The terms *Entity Linking* and *Wikification* are used interchangeably in this thesis, as justified on p. 16.

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

government-related green cluster (Figure 5.10). Is this relevant for a scholar?[45]

- **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.[46] E.g. in the 250 node wikification-based map,[47] *vote* and *bribery* are connected; we can verify the contexts connecting both terms with the search index as described (Figure 5.11).
- **Using the maps' search functions**: The *Navigable* maps can be searched. E.g. searching for *power* we see[48] 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 5.12).

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.

5.2.5.4 Results, discussion, and possible UI improvements

General comments by users

The DH researcher suggested to give more details to the users about the methods to create the corpus maps. I followed this recommendation.[49]

[45] I chose an example where the connection seemed irrelevant to me, thinking that this may help not bias the expert towards thinking that these connections *will* be relevant. See p. 131 for discussions of possible biases in experts' feedback in a visualization evaluation task.
[46] http://apps.lattice.cnrs.fr/bentham
[47] http://apps.lattice.cnrs.fr/bentham/bentham-js-more.html
[48] The example comes from the 250-term wikification-based navigable map, http://apps.lattice.cnrs.fr/bentham/bentham-js-more.html
[49] E.g. in http://apps.lattice.cnrs.fr/bentham/maps-intro.html or http://apps.lattice.cnrs.fr/bentham/lexical-extraction.html

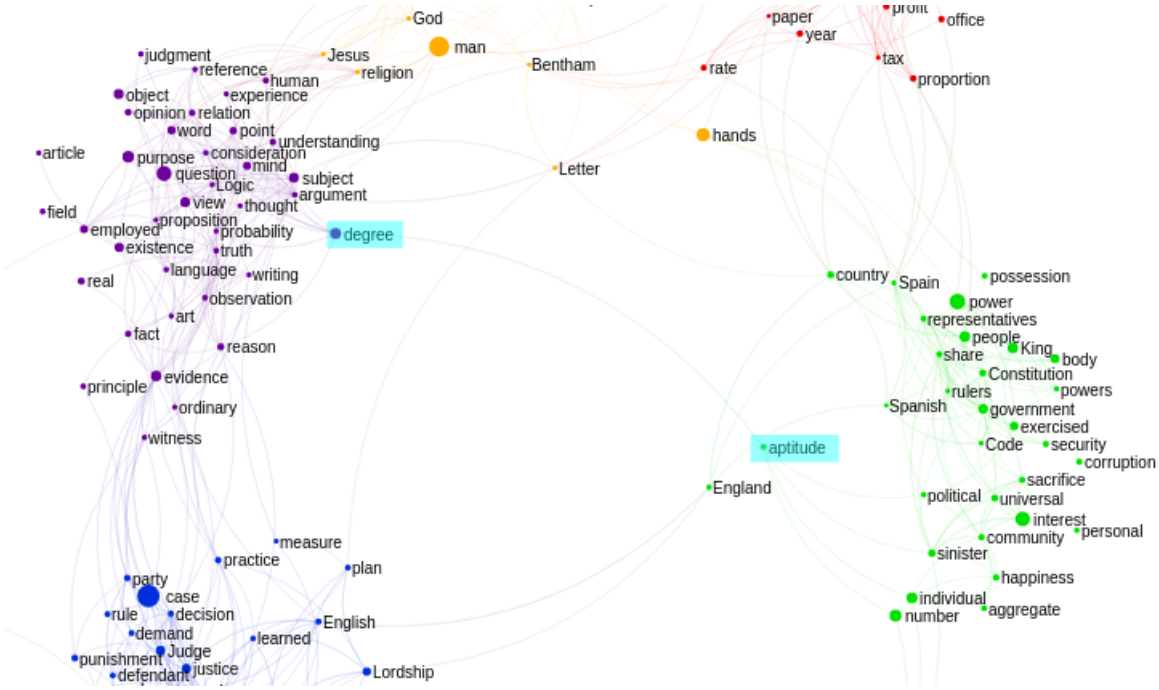


FIGURE 5.10 – Nodes *degree* and *aptitude* connect two clusters in the 150-concept-mention map (Gephi Sigma JS export)

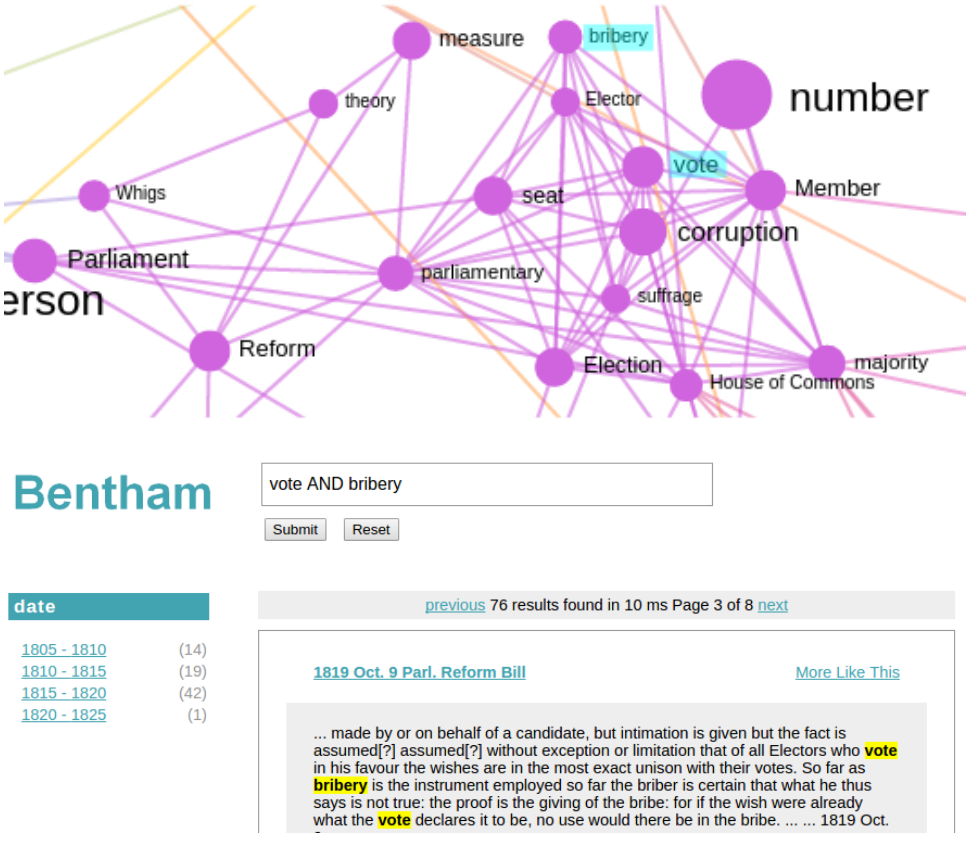


FIGURE 5.11 – The concept-network (top) shows *vote* and *bribery* connected. The search index (bottom) allows us to find those connection's contexts

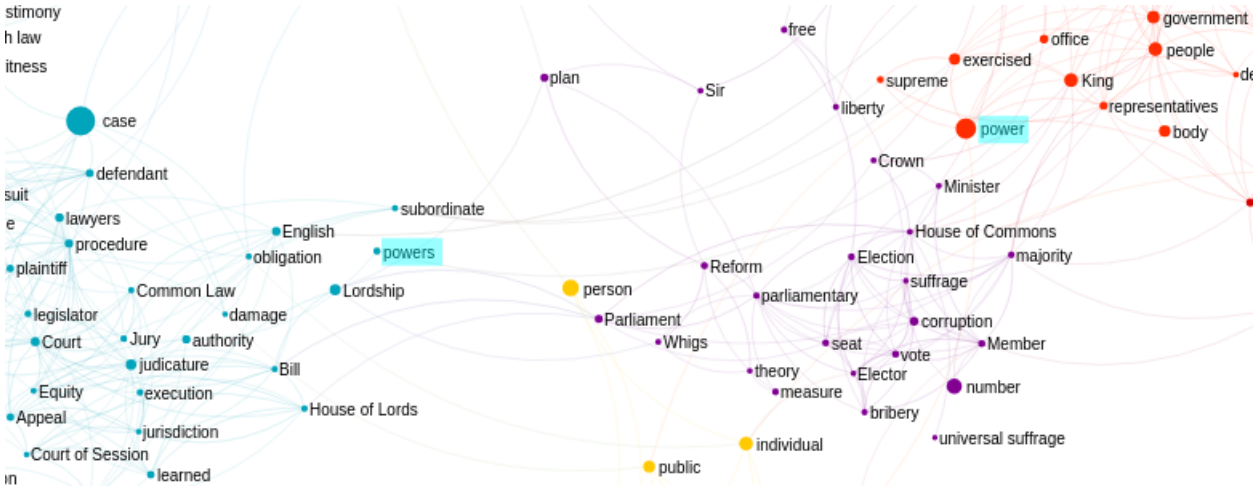


FIGURE 5.12 – Nodes matching query *power* in the 250 concept-mention map (Gephi Sigma JS export)

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,[50] but I agree that this would be useful future work.

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.

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 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 5.13), e.g. *private interest* and *self-regarding interest*

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

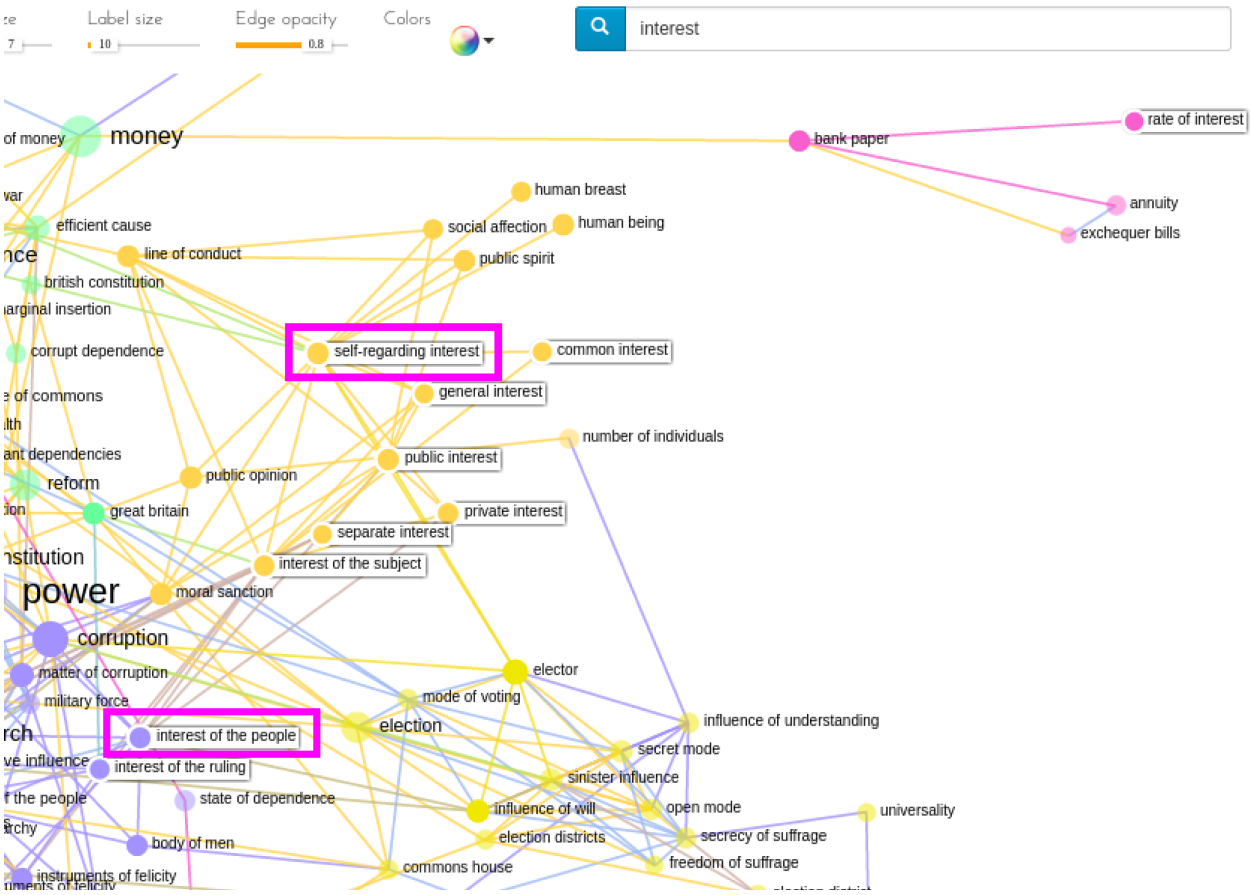


FIGURE 5.13 – Results for query *interest* in the 250-keyphrase network (Tina export). Given a core corpus term like *sinister interest*, the domain-expert identified in these results near-synonyms (e.g. *self-regarding interest*) and near-antonyms (e.g. *interest of the people*) for that concept (pink 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

(near-synonyms of *sinister interest*), or *general interest* and *interest of the people* (near-antonyms to that term).

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. I'd like to point out 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 like TXM (Heiden, 2010, Heiden et al., 2010) or Le Trameur (Fleury et al., 2014), which compute statistically salient terms in the context of a pivot-term would help.[51] Still for the same purpose, but using a very different paradigm to TXM's, 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

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

be tested,[52] or other models that have been shown to perform similarly (Levy et al., 2015).

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.

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[47] shown in Figure 5.14, 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*).

In summary, in terms of potential for new insight for a Bentham specialist, the domain-expert perceived the keyphrase-based networks as better. He considered the networks based on concept-mentions useful for non-specialist users. My interpretation based on other comments by the expert, is that the concept-mention networks express with non-technical terms basic elements

[52] E.g. using the Gensim distribution:
https://radimrehurek.com/gensim/models/word2vec.html

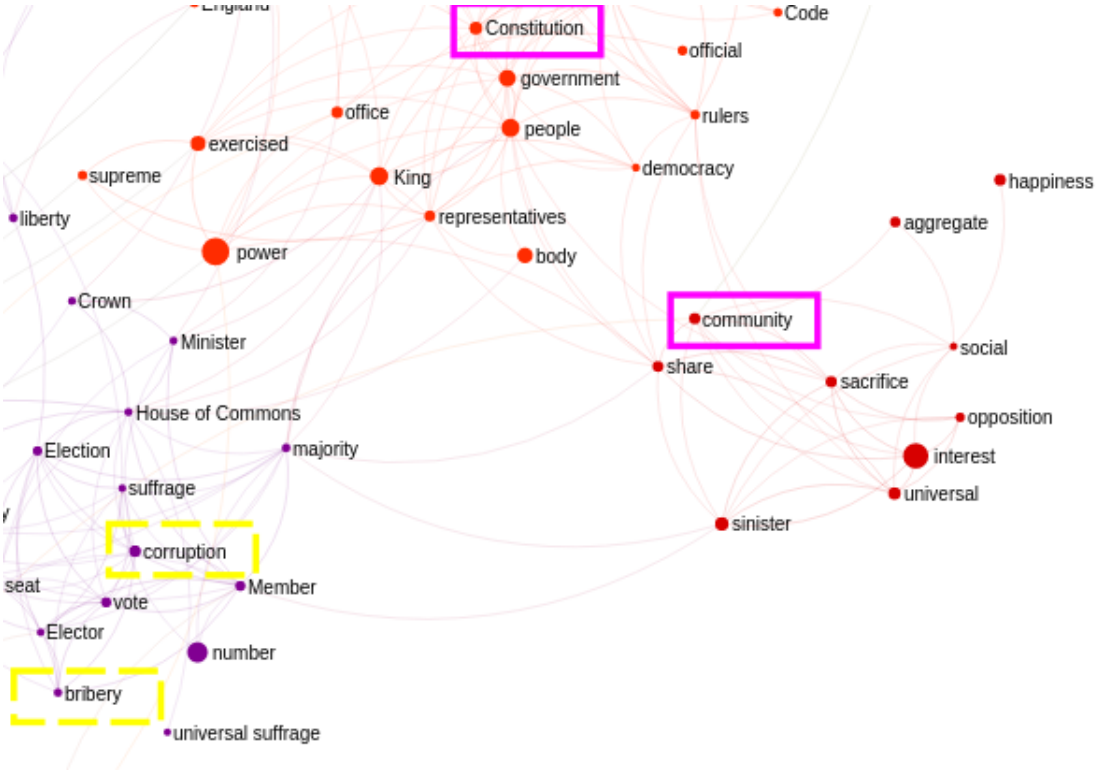


FIGURE 5.14 – Area focused on by domain-expert as representing general Bentham concepts and the relation between them: He describes that the terms surrounded in full pink squares (e.g. *community*) are part of the remedy Bentham proposes for problems like those in the terms with dashed yellow squares (e.g. *corruption*). The map corresponds to the 250 concept-mention network, rendered with the Gephi Sigma JS exporter.

of meaning present in some of Bentham's basic notions, like *interest* or *community* (see Figure 5.14).

Interpretability of the task's results

I consider this task as preliminary validation of the potential of the methods and products developed, rather than exhaustive evidence, for the reasons below.

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 et al. (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 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.

5.2.5.5 Summary of the UI evaluation

This subsection reported on user validation by one Bentham scholar and one DH researcher. In terms of potential gain of insight through the corpus maps and the search interface, the Bentham expert considered that clustering semantically related terms can help find alternative formulations for core Bentham concepts. The representations were found to reflect the expert's knowledge of the corpus, without obvious artifacts that would make them questionable. The domain-expert found that networks whose nodes are concept mentions capture general elements of Bentham's thought, and could be useful for non-specialist students, whereas keyphrase-based networks are more useful for Bentham scholars. The DH researcher suggested adding detail to the user-facing descriptions of some technical details of the lexical extraction and visualization; this recommendation was already adopted. Both experts pointed out that it would be valuable to have access to the corpus contexts for a network node directly by clicking on it. Although the contexts can be obtained from the search index, this would be useful future work.

5.2.6 Summary and Outlook

A detailed summary of the work we carried out with the Bentham corpus follows. Future work possibilities are also recapped for those points in the summary where they are relevant.

We reported on our work to create an application to navigate a subset of ca. 4.7 million words of Jeremy Bentham's manuscripts, previously transcribed and made available by University College London. The user interface offers navigation via a search index and via concept networks of the corpus, where the concepts have been obtained with Natural Language Processing methods. No previous examples of applying automatic text analysis to this corpus existed at the time of writing, which increases the interest of the work reported here.

To obtain corpus maps, three steps were performed: Lexical extraction, network creation via lexical clustering and network visualization. For **lexical extraction**, two technologies were used: Entity Linking to DBpedia with the Spotlight tool, and keyphrase extraction with the Yatea tool. Entity Linking results had some problems, like anachronisms in concept reference resolution. For that reason, concept-mentions themselves rather than the DBpedia labels proposed by our linking tool were used to model the corpus. Keyphrase extraction results were complementary to concept-mention identification: keyphrases can represent specialized notions in Bentham's thought not covered by a general encyclopedic knowledge-base like DBpedia.

The lists of terms and term-variants thus obtained were used to create **concept networks** with the CorText platform. This platform clusters terms based on their common contexts of occurrence, as an indication of semantic similarity between them. It creates a network where lexical items used in similar contexts in the corpus are drawn together. The network represents thematic areas in the corpus and how those areas interact with each other. The network is spatialized as a corpus map which reflects the clusters, using a force-based graph drawing algorithm. Besides, **heatmaps** were created, which depict the corpus areas focused on per decade.

Navigable versions of the **corpus maps** were created with TinawebJS and the Gephi Sigma JS exporter plugin. Each tool's results complement each other. TinawebJS maps provide clearer overview functions, and the Sigma JS exporter maps make it easier to examine the local context of a node and to navigate the network sequentially by selecting related nodes. As future work, it would be practical to do further development to unite the strengths of both tools.

Besides the navigable maps, the transcripts were indexed in a Solr **search server**. This complements the maps, allowing us to retrieve passages and documents mentioning the maps' terms. At present, the search index is not connected to the maps, but users can search for map terms, or combinations of them, by entering these terms on the search interface. It would be useful future work to connect the search index with the corpus maps, so that users

can select the map terms to query the index with, and perform the query, by clicking on the map directly. Other useful future work regarding search would be to take advantage of Transcribe Bentham's TEI annotations about the textual composition process (additions and deletions), and index deleted and added text as such, providing an option on the interface to restrict searches to added or deleted text.

A **user interface** was created to display the corpus maps and the search index.[53] This UI complements the two previously existing platforms to access Bentham's manuscripts, i.e. the Bentham Papers Database and UCL Libraries Digital Collections. Those platforms do not offer concept maps, and their search functions do not identify the context of occurrence of search terms inside each document, unlike our UI. Our work complements but does not replace those tools: The Bentham Papers Database has the most detailed metadata search, and UCL Libraries' platform returns manuscript image files (unlike our UI), besides the text of transcripts matching a query. It would be useful future work to tie together the strengths of those two prior tools and our interface.

Besides complementing prior work by adding new navigation possibilities, the goal of our UI was to provide an overview of the corpus, and, ideally to suggest new research ideas to Bentham experts.

To assess the extent to which the UI's goals were fulfilled, a Bentham scholar and a DH researcher provided feedback on the UI and on the corpus maps, as a preliminary **evaluation of the UI**. The Bentham scholar found that the corpus representations provide an overview that reflects his knowledge of the corpus, without obvious artifacts. As regards the differences between maps based on Entity Linking vs. based on keyphrase extraction, the Bentham expert pointed out that the maps based on keyphrases were more informative for a Bentham scholar, since they contain specialized terms absent from the maps based on DBpedia concept mentions. The concept-mention maps may be useful for non-specialist users. The domain expert commented on potential gain of insight on the corpus by using the UI: Since the networks cluster similar terms together, he finds that the networks may suggest to an expert alternative ways to refer to the same notion (e.g. *sinister interest*). Similarly, that the networks may suggest search terms to find contexts where the same notion is discussed, but perhaps with different words. Since the Bentham expert points out that finding alternative expressions for the same notion is useful for their editorial process, relevant future work would be to create distributional similarity models on the corpus (e.g. with word2vec), that would allow a user to retrieve the most similar expressions to a given term.

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

5.3 PoliInformatics

The second application we built is used to navigate a subset of the PoliInformatics corpus (Smith et al., 2014). This corpus contains heterogeneous materials about the American Financial Crisis of 2007-8. The subset of the corpus analyzed here consists in a Congress report about the sources of the crisis, and in transcripts of hearings by the Federal Crisis Inquiry Commission, which interviewed economics experts and individuals who played a major role in the crisis. The PoliInformatics corpus was the object of an international challenge at the 2014 Conference of the Association for Computational Linguistics (ACL). The aim of the challenge was for participants to produce information based on the corpus that would be relevant for social and political science research, through the application of Natural Language Processing (NLP) methods. The challenge had an open prompt, asking participants to address the following two questions: *Who was the financial crisis?* and *What was the financial crisis?*

The approach to these questions presented here consists in applying several Entity Linking (EL) tools to annotate relevant actors and concepts in the corpus. A user interface (UI) was then created to navigate the corpus using these annotations as facets. The UI also provides measures indicating the quality of each annotation, so that a researcher can decide which to keep. Alternatively, an automatic annotation selection can be performed, using a voting procedure we developed for this application, described in Chapter 3. Finally, some networks were created based on the annotations.

The description of our work is structured as follows: The corpus is presented in 5.3.1, and an overview of prior analyses of the corpus is provided in 5.3.2. The Entity Linking workflow and the annotation attributes it outputs, like entity types or measures to assess annotation quality, are described in 5.3.3. The user interface is introduced in 5.3.4, showing how it exploits the Entity Linking results and annotation attributes. In 5.3.5, application examples of the UI are discussed, assessing strengths and weaknesses. A detailed summary of the work carried out can be found in 5.3.6.

5.3.1 Corpus Description

The PoliInformatics corpus (Smith et al., 2014) consists in diverse texts about the American Financial Crisis of 2007-2008. This corpus was the object of an NLP challenge at the 2014 ACL Workshop on Language Technologies and Computational Social Science.[54] The complete corpus contains legislation, Congressional reports on the crisis, transcripts for monetary policy meetings

[54]The corpus is available from the task's site: https://sites.google.com/site/unsharedtask2014/home

at the Federal Reserve, and transcripts for Congressional hearings on legislation and policy related to the crisis. The corpus also contains the transcript for the first public hearing of the Federal Crisis Inquiry Commission (FCIC), which was appointed by Congress to investigate the causes of the crisis.

5.3.1.1 Corpus sample in our study and preprocessing

The subset of the corpus analyzed here consists in the following documents:

A. Congressional report *Wall Street and the Financial Crisis: Anatomy of a Financial Collapse* (AoC):. This is an official report of Congress' conclusions about the causes of the financial crisis, on the basis of a two-year long investigation involving court hearings and expert interviews. The report contains ca. **318,000 words** in 643 files.

B. Transcripts of the first public hearing of the Federal Crisis Inquiry Commission (FCIC): The first hearing of this Congress-appointed commission interviewed officers from major banks who played a role in the crisis or were affected by it, besides representatives from financial firms and experts. The hearings comprise approx. **82,000 words** in 859 speaker-turns.

As regards **preprocessing**, the task's materials included plain-text versions of the corpus. However, the text required some cleanup before it could be treated with NLP tools. Accordingly, we carried out some preprocessing, the steps of which were the following, in the order listed:

1. Markup left over in the files was normalized, e.g. paragraph marks like `<p>` or `</p>` were replaced by a newline.
2. Hard line-breaks were eliminated: Newlines that did not indicate a paragraph break were replaced by a space, to prevent newlines from cutting up phrases. This was possible because paragraph breaks had been indicated by several blank lines in the original corpus.
3. Sequences of successive spaces or tabs were normalized to one space.
4. Lines consisting in a number only (page numbers) were removed.

In the case of the FCIC hearings, speaker names and the text for each turn were also identified in preprocessing, using regular expressions.

In terms of the formats we represented the corpus with, we created two versions of the corpus:

1. Our preprocessed plain-text version, just described, which was annotated with the Entity Linking tools.
2. An XML version in Solr-format,[55] to index the corpus in the Solr search server.[56] Besides the fields required by default in our Solr setup

[55] https://wiki.apache.org/solr/UpdateXmlMessages
[56] https://lucene.apache.org/solr/, version 4.9.0 was used.

(`title`, `date`), for the FCIC hearings we added a `speaker` field to each turn.

In the AoC corpus, the titles were created from the first 50 characters in the text. For the FCIC hearings, the title consists of the speaker name, besides a sequential turn-ID we assigned to each turn. Note that the date field is redundant in this application, since all documents in each subcorpus have the same date.

5.3.2 Related Work

We review two types of relevant prior work. First, papers that have analyzed the PoliInformatics corpus. Second, tools that have applied similar approaches to ours, using automatic annotation of actors and concepts as a means to gain an overview of a corpus.

5.3.2.1 Prior work on the corpus

The most important source of work on the corpus is the 2014 ACL challenge described in 5.3.1, an overview of which can be found in (Smith et al., 2014). As an answer to the task's open prompt about *who* and *what* the financial crisis was, participants covered a range of issues and technologies.

Regarding the the *who was the financial crisis* question, two participating teams studied central bankers' behaviours: Baerg et al. (2014), which was elected as the task's best paper, proposed a procedure to position central bankers on a scale of aversion vs. tolerance towards inflation. They developed an inference method based on topic models of bankers' speech in transcripts of monetary policy meetings at the Federal Open Market Committee (FOMC),[57] which is part of the US central bank (the Federal Reserve or *Fed*). Zirn et al. (2014) used the same transcripts to examine which Fed members' positions are closer to each other, using cosine similarity between word vectors for their turns at the meetings. Clark et al. (2014) turned their attention to Congressional hearings, analyzing participants' sentiment towards issues and people, as well as opinion shifts. Bordea et al. (2014) applied expertise mining to the corpus to identify which participants are expert in what topics. The goal in Morales et al. (2014) was creating a social network depicting connections between banks and other economic and political actors.

As regards the *what was the financial crisis* question, Miller et al. (2014) assessed the evolution of concerns at the FOMC meeting transcripts[57] based on period-specific topic models covering several years before and during the crisis. Li et al. (2014b) studied text-reuse in crisis-related legislation, to

[57] https://www.federalreserve.gov/monetarypolicy/fomc.htm

determine the extent to which it contained novel policies or adopted pre-existing ones. Wang et al. (2014) created a system to provide summaries of meeting transcripts and Congressional hearings relevant for a user query, where estimated speakers' expertise was factored into relevance scoring for text retrieval. Kleinnijenhuis et al. (2014) examined the reciprocal influence of the media and Congress based on US and UK news and Congressional hearings.

Finally, our lab's participation at the task (Bourreau et al., 2014; Poibeau et al., 2015) performed named-entity recognition and keyphrase extraction to find majors actors and topics in the crisis. Networks and diachronic maps for the content of the corpus were created with network analysis and visualization tools: Gephi[58] (Bastian et al., 2009) and the CorText platform[59] (Chavalarias et al., 2013; Rule et al., 2015). By contrast, in this thesis we have used several Entity Linking tools for annotating actors and concepts, and have created a custom interface to navigate the corpus.

5.3.2.2 Prior tools related to our user interface

The tool that we find closest to the goals of our user interface is ANTA. This tool is not currently maintained,[60] but our approach is inspired by its objectives. The tool applied keyphrase extraction and Entity Linking with Alchemy API's web services (part of IBM Watson since 2015.)[61] The tool focused on helping users choose keyphrases and entities, based on their corpus frequency and document frequency; it also provided facilities for manual merging and classification of terms (Venturini et al., 2012, p. 10ff.).

The tool was created by social scientists and developers from the Sciences Po médialab in Paris. Discussing the tool, its authors state what they see as desirable features, and problems to avoid, in an automatic entity annotation tool (Venturini et al., 2012, pp. 7, 15, emphasis added):

> *[W]e **don't like** that the [Alchemy] service is offered as a **"black box"** and that the exact algorithm is secret. Something that is perfectly reasonable from a commercial viewpoint may be a problem for research.*
>
> *Expression extractions should be improved and implemented on **open source** software. The careful use of natural language processing algorithms could provide **better filtering metrics** and support in expression merging.*

[58] https://gephi.org/
[59] https://docs.cortext.net/
[60] https://github.com/medialab/ANTA
[61] The relevant current Watson services would be https://www.ibm.com/watson/developercloud/alchemy-language/api/v1/#entities and https://www.ibm.com/watson/developercloud/alchemy-language/api/v1/#concepts

To address those needs expressed by social scientists, our interface is based on open source Entity Linking tools. Additionally, we provided filtering measures reflecting annotation confidence (how likely the annotation is correct) and a notion of coherence between an annotation and other annotations in the corpus (to what an extent it is thematically consistent with other annotations). More details are provided in the following pages.

5.3.3 Entity Linking Backend

This subsection describes the Entity Linking (EL) workflow applied to the corpus, the results of which are exploited on the UI. In 5.3.3.1, the following elements are discussed: The EL tools applied and their settings, a procedure to combine multiple tools' outputs selecting the best ones, and our entity classification method. In 5.3.3.2, I describe our implementation of two measures that can help assess annotation quality: confidence and corpus-level coherence.

5.3.3.1 DBpedia annotations: acquisition, combination and classification

We used Entity Linking tools to annotate any type of resource present in Wikipedia or DBpedia,[62] whether the resource is expressed in the corpus by a named entity or by a common-noun phrase. Named entities are lexical sequences from specific types, like persons, organizations, locations, products, laws and others; they are generally expressed by proper nouns (see p. 16). However, many resources in our target Knowledge Bases (KBs) represent conceptual information, expressed with common nouns, like DBpedia concepts *Loan* or *Risk*. Both types of resources are useful to get an overview of the PoliInformatics corpus; named entities will more likely correspond to actors, and conceptual resources will refer to topics in the corpus. When annotating conceptual resources besides named-entities, the literature sometimes uses the term *Wikification* instead of *Entity Linking*. In this thesis, both terms are used interchangeably, as justified on p. 16. For related reasons (p. 16), I speak indistinctly of a Knowledge Base's *concepts*, *entities* or *terms*, irrespectively of the types of expressions used to refer to them in the corpus.

The annotations shown on the UI come from three different Entity Linking tools: TagMe2 (Ferragina et al., 2010; Cornolti et al., 2013),[63] Wikipedia Miner (Milne et al., 2008a)[64] and DBpedia Spotlight (Mendes et al., 2011; Daiber

[62] Strictly speaking, among the three tools we applied, only DBpedia Spotlight uses DBpedia as the target Knowledge-Base, whereas TagMe2 and Wikipedia Miner use a database version of Wikipedia. The distinction is not essential for the application described here. For conciseness, I often speak of linking to *DBpedia* in the pages below, instead of using the more precise formulation *DBpedia/Wikipedia*.

[63] https://tagme.d4science.org/tagme/ The system is a reworked version of TagMe (Ferragina et al., 2010)

[64] https://github.com/dnmilne/wikipediaminer

et al., 2013).[65] These tools were chosen since they are open source, which can help interpret the results, as the algorithm is open to inspection.[66] Also, since an evaluation of their results on different test corpora suggests that their results complement each other, in the sense that the tools' performance is affected differently by characteristics of the corpora like the types of named-entities found in them, as was discussed on p. 25.

The tools were accessed via their public web services, with default options. Outputs were filtered according to optimal **confidence thresholds** for each tool, assessed on the IITB reference corpus (Kulkarni et al., 2009), using the evaluation framework by Cornolti et al. (2013).[67] Annotations whose confidence score did not reach the threshold were filtered out. These confidence scores are an estimation of how likely an annotation is correct; they are discussed further when presenting the information provided on the UI for users to assess annotation quality (5.3.3.2).

The reason to assess optimal confidence thresholds based on the IITB corpus was the following: This is a web corpus covering various domains (news, science and culture), and has been annotated not only for Knowledge-Base resources expressed by proper nouns, but also for many common-noun concepts. Since in the PoliInformatics corpus we want to annotate both types of KB resources, using optimal thresholds for IITB seemed a reasonable choice.

For obtaining the optimal thresholds, results were evaluated with the Weak Annotation Match (WAM) measure (p. 21), which requires Knowledge-Base (KB) concepts to match the reference exactly, whereas their corpus mentions only need to overlap with the reference. We used WAM since it is not essential for concept mentions in the corpus to be perfectly delimited in the results in order to navigate the corpus on our UI.

The thresholds used were 0.094 for TagMe2, 0.016 for Spotlight, and 0.219 for Wikipedia Miner.

After obtaining Entity Linking annotations from the different services just mentioned, our **workflow combines the results**, automatically selecting annotations more likely to be correct. To this end, we implemented a weighted voting procedure inspired by the ROVER method (Fiscus, 1997; De La Clergerie et al., 2008).[68] Our procedure was described and evaluated in Chapter 3 (p. 60ff.). The basic idea in this procedure is that annotations that have been

[65] https://github.com/dbpedia-spotlight/

[66] Note however that Wikipedia Miner no longer has a publicly accessible instance, unlike when we used it for this application. Accordingly, even if the code is still public, reproducibility of results for this tool is difficult, as local deployment of the tool is not trivial.

[67] The *BAT Framework*: https://github.com/marcocor/bat-framework

[68] *ROVER* stands for *Recognizer Output Voting Error Reduction*.

proposed by a smaller number of systems are less likely to be selected. Each system's performance on a series of reference corpora is also taken into account, weighting a system's annotations accordingly.

A further step in the workflow is **classifying the annotations**, assigning them a category among *Concept*, *Person*, *Organization* and *Location*. Category *Concept* consists in terms usually expressed by common nouns, like *Loan* or *Investment*. The content of the other categories is respectively person names, organization names and geographical locations. Classification is rule-based. It involves searching for category indicators in the category or type labels output by the EL services in their response. Some rules involve an exact match against the categories or types in the response, e.g. "Assign type *Location* if the annotation has type *DBpedia:Place*". Other rules involve a partial match, e.g. "Assign type *Person* if one of the Wikipedia category labels for the annotation contains the word *births*". The classification results were assessed by informal inspection and their quality seems sufficient for navigating the corpus according to resource type. A formal evaluation was not performed, this would be relevant future work.

5.3.3.2 Annotation quality assessment: confidence and coherence

A requirement expressed by social scientists we have discussed Entity Linking with is filtering metrics to guide a researcher's manual filtering of annotations (Venturini et al., 2012). Our user interface (UI) provides two measures as an estimation of annotation quality: a confidence score and a corpus-level coherence score. Our implementation of these measures is described in following. Both measures can help for manual filtering, besides other information presented in 5.3.4.1.

Confidence scores provide an indication of how likely the annotation is correct. Ways for Entity Linking tools to arrive at a confidence score were described on p. 19. In essence, annotations for expressions that are highly ambiguous in a context will receive less confidence than expressions whose reference in the Knowledge Base is clear in the context.

Since the minimum confidence threshold for each Entity Linking tool was different, the original confidence-score range in each tool's results was different. Each tool's confidence scores were normalized to a range between 0 and 1, using *min-max* scaling, the definition for which is in Equation 5.1.

Min-max scaling: Given a value v from the original range between min_o and max_o, the value v_{mm}, which represents v min-max scaled to a new range between min_n and max_n, is obtained thus:

$$v_{mm} = \frac{(v - min_o) \cdot (max_n - min_n)}{max_o - min_o} + min_n \tag{5.1}$$

The extent to which this type of scaling was useful to compare results across tools is discussed on p. 151.

Besides confidence scores, **coherence scores** are another type of information relevant for manual filtering. Coherence scores indicate how related an annotation is to other annotations for expressions in its context, according to a given definition of relatedness. Most EL tools factor in coherence scores in their disambiguation process (see p. 19 for an overview).

A widespread measure of coherence is based on a semantic distance measure defined by Milne and Witten in 2008. This computes a semantic distance between Wikipedia pages based on the number of common inlinks between them, with a larger amount of common inlinks indicating less distance. The definition, provided in (Milne et al., 2008a, p. 27) and (Milne et al., 2008b, Section 3.1), represents a *distance*, because its output increases as the number of common inlinks decreases. Based on that distance, in order to obtain a *relatedness* measure whose increasing values reflect an increasing number of common inlinks, an approach is to deduce the distance from 1. The resulting relatedness measure (Equation 5.2) has been called by several authors (Hoffart et al., 2011; Vieira, 2015) the **Milne-Witten coherence**,[69] and it is used in Wikipedia Miner (Milne et al., 2008b), TagMe (Ferragina et al., 2010) and AIDA (Hoffart et al., 2011). Its definition follows.

Milne-Witten coherence: Given set N, containing all pages in Wikipedia, the set of pages IN_{c_1} linking to the page for concept c_1 and the set of pages IN_{c_2} linking to the page for concept c_2, the Milne-Witten coherence $Coh_{MW}(c_1, c_2)$ between c_1 and c_2 is defined thus, following Hoffart et al. (2011):

$$Coh_{MW}(c_1, c_2) = 1 - \frac{\log\left(\max\left(|IN_{c_1}|, |IN_{c_2}|\right)\right) - \log\left(|IN_{c_1}| \cap |IN_{c_2}|\right)}{\log\left(|N|\right) - \log\left(\min\left(|IN_{c_1}|, |IN_{c_2}|\right)\right)} \tag{5.2}$$

As seen in Equation 5.2, the Coh_{MW} computes relatedness between two Wikipedia concepts c_1 and c_2 as a function of the number of common inlinks to both of their pages, i.e. the number of third pages that simultaneously contain a link to the pages for both c_1 and c_2. Highly coherent concepts according to this measure have a large number of third pages linking to both of their pages.

[69] Some clarification on the literature may be relevant. The formula named *relatedness* in (Milne et al., 2008b, Section 3.1) and *sr* in (Milne et al., 2008a, p. 27) encodes a *distance*, as argued above, in spite of the formula's name. In fact, the same authors' implementation of relatedness in the Wikipedia Miner software deduces the formula from 1, see `https://github.com/dnmilne/wikipediaminer/blob/master/wikipedia-miner-core/src/main/java/org/wikipedia/miner/comparison/ArticleComparison.java#L117` as well as `ArticleComparer.java#L457` in the same package. Later authors who adopt Milne-Witten coherence like Hoffart et al. (2011, p. 787) or Vieira (2015, p. 20) explicitly provide a definition for relatedness deducing the distance from 1, as in Equation 5.2.

We obtained the Milne-Witten coherence values using the `compare` endpoint of Wikipedia Miner's public web service. The service is no longer publicly available since early 2016, and local setup is not trivial. To obtain similar results, a possibility (other than reimplementation applied to a Wikipedia dump) would be to use TagMe2's relatedness web-service.[70]

Corpus-level coherence: Based on the Milne-Witten coherence (Eq. 5.2), which applies to two concepts, we implemented a more general measure Coh_corpus, to determine how coherent an annotation is with other annotations in the corpus overall. In essence, our measure computes the averaged Milne-Witten coherence between an annotation and a subset of annotations that are considered representative for the corpus, because they are overrepresented in it and have above average confidence scores. Implementation details follow.

1. Selection of a comparison concept-set: To compute a concept's coherence Coh_corpus with concepts in the corpus overall, the first step is to select a subset of corpus concepts $\mathcal{C}_r$ considered representative for the corpus. The composition of $\mathcal{C}_r$ relies on annotations' corpus frequency and average confidence in the corpus. The set of concepts $\mathcal{C}_{rcan}$ to consider as candidates for inclusion in $\mathcal{C}_r$ consists in concepts whose corpus frequency is at least 3 times the average corpus frequency, and whose normalized annotation confidence is higher or equal than the average normalized annotation confidence for all concepts in the corpus, using *min-max* normalization (Equation 5.1). Concepts in $\mathcal{C}_{rcan}$ are ranked by decreasing normalized confidence, and the top 5% concepts in $\mathcal{C}_{rcan}$ are included in $\mathcal{C}_r$. The thresholds for corpus frequency, confidence and percentage of $\mathcal{C}_{rcan}$ to keep were established empirically.

2. Partial scores combined: The corpus-level coherence $Coh_corpus(c, \mathcal{C}_r)$ between a concept c and the set $\mathcal{C}_r$ of concepts representative of the corpus is articulated into two scores:

1. $Coh_{type}(c, \mathcal{C}_{rtype})$, between concept c and the subset $\mathcal{C}_{rtype}$ of concepts in $\mathcal{C}_r$ which are unequal to c and whose type matches the type of c.
2. $Coh_{all}(c, \mathcal{C}_{rall})$, between concept c and all concepts in $\mathcal{C}_r$ unequal to c.

[70] `https://tagme.d4science.org/tagme/rel`, documented at `https://services.d4science.org/web/tagme/documentation`. A quick test shows similar results to Wikipedia Miner's. TagMe2' source code (available from its authors) implements concept relatedness using the Milne-Witten coherence from Eq. 5.2 above, in class `preprocessing.graphs.OnTheFlyMeasure`, method `rel`. However, at this writing I have not re-contacted the authors to confirm if their current public web-service is still using the implementation in the code they had made available earlier.

$Coh_{type}(c, \mathcal{C}_{rtype})$ is defined as follows, with Coh_{MW} as in Eq. 5.2:

$$Coh_{type}(c, \mathcal{C}_{rtpye}) = \frac{\sum_{c' \in \mathcal{C}_{rtype} \setminus \{c\}} \left(Coh_{MW}(c, c') \cdot \frac{\sum_{c'' \in \mathcal{C}_{rtype} \setminus \{c,c'\}} Coh_{MW}(c', c'')}{|\mathcal{C}_{rtype} \setminus \{c,c'\}|} \right)}{|\mathcal{C}_{rtype} \setminus \{c\}|} \quad (5.3)$$

$Coh_{all}(c, \mathcal{C}_{rall})$ is defined the same way as $Coh_{type}(c, \mathcal{C}_{rtype})$, substituting $\mathcal{C}_{rall}$ for $\mathcal{C}_{rtype}$.

In both $Coh_{type}(c, \mathcal{C}_{rtype})$ and $Coh_{all}(c, \mathcal{C}_{rall})$, the Milne-Witten coherence Coh_{MW} between c and each concept c' in $\mathcal{C}_{rtype}$ or $\mathcal{C}_{rall}$ is itself weighted by the Coh_{MW} between c' and each concept c'' unequal to c or to c' in $\mathcal{C}_{rtype}$ or $\mathcal{C}_{rall}$. This is meant to decrease the impact of possible concepts in $\mathcal{C}_r$ which may be weakly related to the core themes of the corpus, by assigning a lower weight to the relatedness between c and those concepts.

3. Final score: Finally, $Coh_corpus(c, \mathcal{C}_r)$, the corpus-level coherence between a concept c and the corpus overall (as represented by the set of entities $\mathcal{C}_r$) is obtained by adding $Coh_{type}(c, \mathcal{C}_{rtype})$ and $Coh_{all}(c, \mathcal{C}_{rall})$, weighted by parameters α and β respectively, which indicate the relative importance of relatedness to concepts of the same type vs. relatedness to concepts of any type.

$$Coh_corpus(c, \mathcal{C}_r) = \alpha \cdot Coh_{type}(c, \mathcal{C}_{rtype}) + \beta \cdot Coh_{all}(c, \mathcal{C}_{rall}) \quad (5.4)$$

Parameters α and β were set at 0.5. This setting was established empirically.

The Coh_corpus scores displayed on the UI are normalized to a range between 0 and 1, using *min-max* scaling (Equation 5.1).

The extent to which our Coh_corpus score helps identify incorrect annotations is discussed on p. 151. In essence, it worked fine for resource types *Organization* and *Concept*, but not so for types *Person* and *Location*.

As a final comment on corpus-level coherence, in (Ruiz Fabo et al., 2015c, p. 47), we mentioned tests with other coherence measures, based on distance between nodes in the Wikipedia category graph. The results with those methods did not improve over the corpus-level measure presented here. For simplicity, the scores displayed on the user interface were computed with the measure presented here only.

5.3.4 User Interface: Corpus Navigation with DBpedia Facets

The goal of the User Interface (UI)[71] is to help users choose a representative set of terms to model a corpus based on the output of Entity Linking tools, with the help of term frequency and other measures like annotation confidence, or the corpus-level coherence scores described above (p. 143). On the UI, users can assess the validity of a term by simultaneously looking at the measures, and at its context of occurrence in the documents where that term was annotated.

This subsection describes the interface's functions, discussing how it makes use of the results of the Entity Linking workflow from 5.3.3 above, and of the annotation attributes it provides. In 5.3.4.1, the colour coding used to represent concepts' confidence and coherence scores is introduced. The UI's search and filtering functions are described in 5.3.4.2. Besides such information, relevant for manual annotation filtering, the UI also shows the results of an automatic selection of annotations (5.3.4.3). Finally, result ranking is described in 5.3.4.4.

Two limitations in the current implementation of the UI should be noted at this point. First, whereas concepts can be filtered in several ways, there is not a function to allow users to export the filtered set of concepts. This would be necessary future work to exploit users' annotation selection for purposes like creating networks for the corpus. Second, the UI currently permits navigating the PoliInformatics corpus only, i.e. it is not possible to upload a new corpus to it. Enabling the UI to do so would be useful future work. In spite of these limitations, the current implementation permits assessing the potential of the approach (see 5.3.5).

The UI was implemented in PHP and Python and it looks best on Chrome-based browsers.[72]

5.3.4.1 Visual representation of annotation quality indicators

The user interface (UI) attempts to provide a convenient way for users to get an idea of the quality of a result by representing confidence and coherence scores with a colour scale.

An annotation's confidence score provides an indication of how likely it is correct (p. 141). The corpus-level coherence score represents to what an extent the concept is thematically related to other concepts in the corpus overall (p. 143). Both scores are output in a range between 0 and 1; the score can be seen on the interface by hovering over each of the coloured cells on the left pane.

[71] http://apps.lattice.cnrs.fr/nav/gui/
[72] The figures in the pages below are screen captures on Chromium.

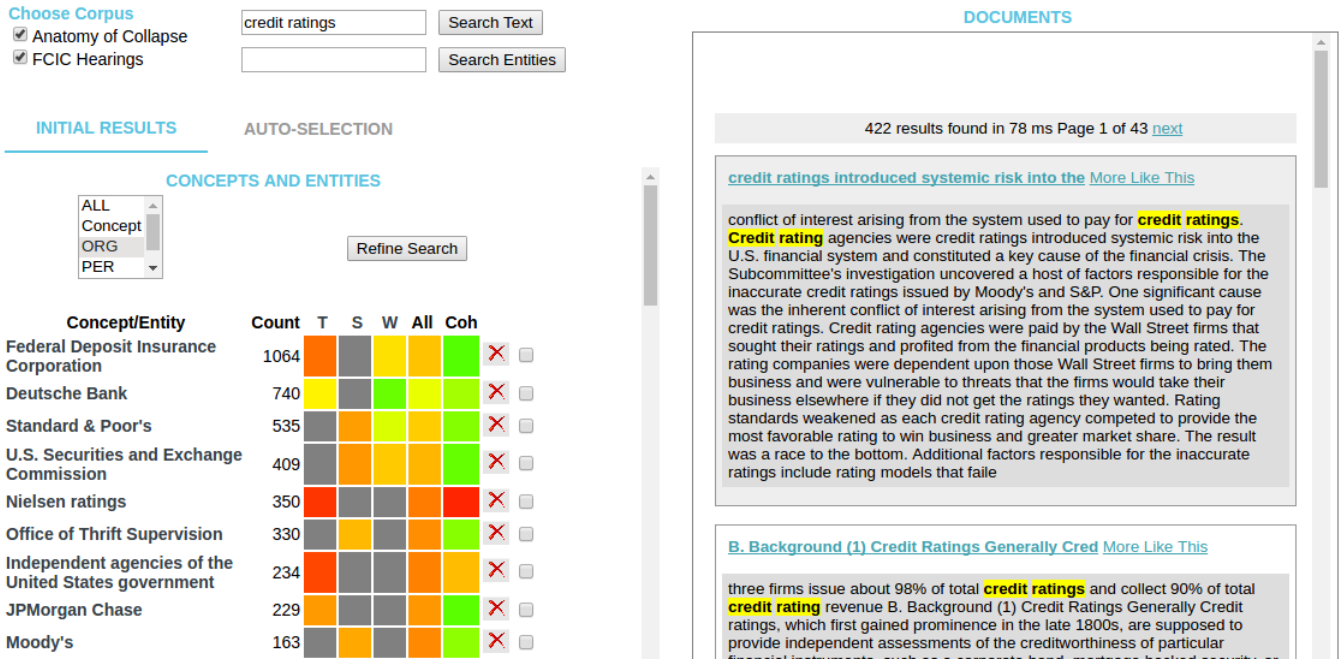


FIGURE 5.15 – Results for text query ***credit ratings***, restricted to annotation type **ORG**anization, chosen in the select box. The panel on the right shows documents matching the query. The left panel shows DBpedia concepts annotated in those documents by TagMe2 (column **T**), Spotlight **(S)**, and Wikipedia Miner **(W)**, with their frequency in the result-set and their Coherence score **(Coh)**. Colours indicate confidence scores.

The colour columns on the concept pane (Figure 5.16, Items 2 and 3) have the following meaning: Columns ***T***, ***S*** and ***W*** show a concept's average confidence score in the corpus for each Entity Linking tool: ***T*** for TagMe2, ***S*** for DBpedia Spotlight and ***W*** for Wikipedia Miner. Column ***All*** is the average of the previous three columns. Finally, column ***Coh*** shows the coherence score at corpus level.

Scores are represented visually as a *red–yellow–green* colour scale,[73] where red stands for a confidence score of 0 and pure green is a confidence of 1. Shades of orange and yellow-green represent intermediate values.

Dark grey cells correspond to cases where one or two of the three services (TagMe2, Spotlight, Wikipedia Miner) have not extracted the concept. The number of dark grey cells for a concept is itself an indication of its reliability: Concepts extracted by only one of the services are less likely to be correct than concepts output by several services.

Light grey cells occur in isolated cases in the *Coh* column; they represent cases where the corpus-level coherence algorithm (p. 143) did not reach a valid result. This can happen given missing values or error responses from the `compare` function in Wikipedia Miner's web service (p. 142), which we

[73] The colour scale was implemented following `http://stackoverflow.com/a/26204509`. This uses a colour model known as HSL. This model (as does a similar one called HSV) results in visually pleasing colour scales, as discussed at the same source.

used to obtain the Milne-Witten coherence scores (Coh_{MW}), which are the basis of our corpus-level coherence measure (Coh_corpus).

Figure 5.17 (left) on p. 154 shows examples how the colour coding can help identify wrong annotations. Several concepts in the image for *Initial Results* have only been extracted by TagMe2, as seen from the dark grey colour in the cells for Spotlight and Wikipedia Miner (e.g. *Gemstone_Publishing*, *Italian_Socialist_Party* and *Portfolio.com*). This is an indication that those results are probably wrong. Moreover, the coherence scores are moderate (yellow) for *Portfolio.com* and very low (red) for the other two concepts. This also helps flag the concepts as likely errors.

5.3.4.2 Search and filtering functions

The interface allows navigating the corpus through text-search and through facets for Wikipedia/DBpedia terms.[74] The functions are described below, and summarized in Figure 5.16.

Backend

Regarding the backend, the corpus was indexed in the Solr search server,[75] which was described on p. 117. The Entity Linking annotations were stored in a MySQL database. Using the same document IDs in the search index and in the database allows to combine the results from both. A `Search Text` query runs first on the Solr index and then obtains from the database the annotations for the documents in Solr's response. A `Search Entities` query runs first on the database, returning a list of annotations. Then it retrieves from the Solr index the documents where mentions for those annotations occur.

User-facing functions

A **Search Text** query displays, on the right panel, the documents matching the query, with the query term highlighted. The Wikipedia/DBpedia terms annotated in those documents are shown on the left panel. Figure 5.15 shows the results for query *credit ratings*, with annotations restricted to organizations (thanks to the `Refine Search` function described on p. 148 below).

A **Search Entities** query displays concepts whose DBpedia *label* matches the query on the left panel, and, on the right, the documents where those concepts were annotated. Note that this is an imperfect implementation of

[74] DBpedia is a semantic web repository reflecting Wikipedia information. Strictly speaking, only Spotlight links to DBpedia among the tools we applied; TagMe2 and Wikipedia Miner disambiguate against a Wikipedia database. For conciseness, I refer to the interface terms as *DBpedia* terms rather than *DBpedia/Wikipedia terms*. The distinction is not essential for the application described here.

[75] http://lucene.apache.org/solr/

entity search. A more complete way to implement this would be to look for the query term not only in the labels, but also in the corpus mentions (i.e. corpus variants) for each label, which are available in the database. A similar limitation exists for highlighting: The query string itself is highlighted in the documents, but other variants contained in the database as mentions for the same concept are not currently highlighted in the documents. The implementation was simplified in the interest of time, and since the UI was meant to show the potential of the approach, rather than to be used as a research tool for the corpus. It would be useful to improve this in the future.

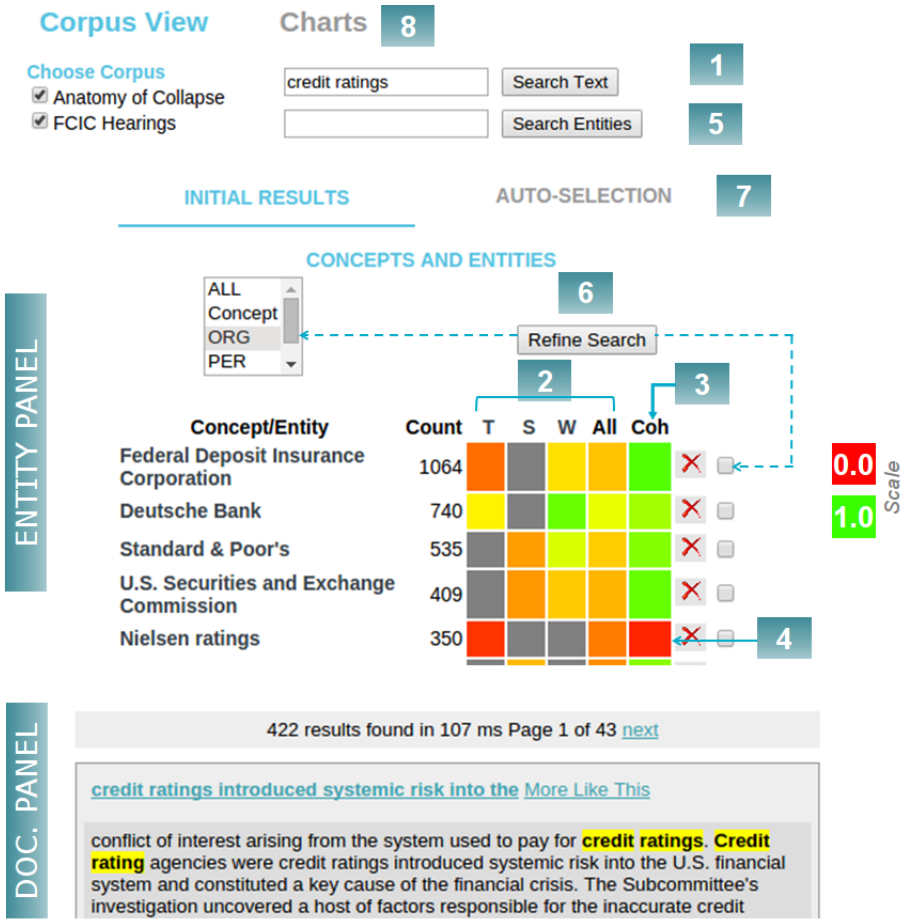
Refine Search filters results according to a user selection, which can be performed one of two ways:

- By choosing one or more annotation types (*Concept*, *ORG*anization, *PER*son, *LOC*ation) on the select box above the DBpedia concept list. For instance, in Figure 5.15, results have been restricted to organizations. The select box can also be used without a query on the text boxes, to return all terms of a given type.
- By selecting individual annotations with the checkboxes to the right of each term. Highlighting in the document panel is limited to the label for the selected term; other variants stored in the database for the same term are not highlighted. This could be improved in the future.

5.3.4.3 Automatic annotation selection

Social scientists we have been in contact with prefer manual filtering based on estimates of result quality and on an examination of terms in their context, rather than an automatic filtering of results (p. 138). This provides them with a more transparent and interpretable term selection method than applying an automatic algorithm whose settings they do not necessarily control. For that reason, we showed annotation quality measures on the interface and give access to concepts' context of occurrence thanks to the full-text search panel. However, for some purposes, like getting a quick idea of the most important annotations for a corpus, it may be helpful to perform an automatic selection of annotations. The interface can also show the results of such a procedure.

The results of automatic annotation selection are accessed on the *Auto-Selection* tab of the UI (see Figure 5.16, Item 7). This restricts the concepts displayed, and available for searching, to those automatically selected by the weighted voting procedure from Chapter 3 (p. 60ff.). This procedure combines all annotators' results, selecting those for which there is sufficient agreement among annotators, and filtering out the rest. On the *Auto-Selection* tab, the concepts' average confidence scores for each annotator (columns



(1) Query via *Search Text* displays:

- **Document Panel: Documents matching the query**
- **Entity Panel: Entities extracted in the documents matching the query displayed on doc. panel, plus:**
  - **(2) Confidence Scores for each annotator, normalized to a 0-1 range. (T=Tagme, S=Spotlight, W=Wikipedia Miner).**
  - **(3) Coherence score between the entity and a representative subset of the corpus entities.**
  - **(4) Entities not coherent with the corpus are flagged in red.**

(5) Query via *Search Entities* displays:

- **Entity Panel: Entities matching the query.**
- **Document Panel: Documents containing one of the entities displayed on the entity panel.**

(6) *Refine Search*: Entities can be selected with a list of types (like *ORG*) or selected individually with checkboxes.

(7) The *Auto-Selection* tab shows the output of an automatic filtering via weighted voting of annotations.

(8) *Charts*: examples of co-occurrence networks, created offline exploiting workflow information (sentence number, confidence, ...)

FIGURE 5.16 – Summary of User Interface functions: Numbered items on the bottom text describe the functions indicated by numbers on the top image. From our NAACL demo poster (Ruiz Fabo et al., 2015c).

T, *S*, *W*) are based on the scores for each annotator's output in the selected annotations only.

5.3.4.4 Result sorting

Concept sorting on the left panel, as can be seen in the *Count* column in Figures 5.15 or 5.16, was by decreasing frequency of occurrence in the corpus. Making the concept table sortable by the other columns would be a useful enhancement.

As regards document sorting on the right panel, for `Search Text` queries, Solr relevance scoring applies. Solr result ranking was described on p. 117. In short, Solr ranks result-documents according to how many query terms were found in them and how frequent those terms are in the document, using tf-idf weights with raw term-frequency counts; see variant *(a)* in footnote 23.

For `Search Entities` queries, Solr relevancy does not apply to sort the documents returned, since they are retrieved by document-ID based on the results of a MySQL query. The results of this SQL query were sorted by the ID of the corpus mentions for each of the concepts returned. This sorting criterion does not reflect document relevance, it roughly corresponds to the sequence of appearance of the mentions in the corpus. In future work, it would be better to sort the documents in a way that reflects the ranking of entities returned for the entity query, displayed on the UI's left pane. E.g. ranking documents according to how the entities mentioned in them are ranked in the results for the entity query.

5.3.5 User Interface Example Uses and Evaluation

The interface features were described in the previous subsection. This subsection presents example uses of each of those features, discussing strengths and weaknesses. In that sense, the discussions can be seen as an evaluation of the interface.

An evaluation with domain experts was not performed for this interface, unlike for the other two corpus navigation applications in the thesis (see 5.2.5 for the evaluation of the Bentham interface and 6.6 for the Earth Negotiations Bulletin application). Instead, I present and comment on several examples of my own experience using the interface, which illustrate the potential of the approach as well as shortcomings in the current implementation.

The use of confidence scores and coherence scores is discussed in 5.3.5.1 and 5.3.5.2 respectively. Example results of the automatic selection of annotations are shown in 5.3.5.3. An intended application of the UI is helping select concepts for purposes like creating corpus networks. The networks cannot be created online on the UI, but the Entity Linking workflow provides the

information required for network creation. Example networks are presented in 5.3.5.4, focusing on how the annotation attributes in the UI can help validate the network. Finally, a limitation of the UI is that corpus actors not present in DBpedia/Wikipedia are not annotated; this is discussed in 5.3.5.5.

5.3.5.1 Using confidence scores

As explained above (pp. 19, 141), these scores give an indication of the extent to which we can trust an annotation (i.e. how likely it is to be correct). The possible score range is between 0 and 1. However, within that range, given different minimum confidence thresholds and scoring methods for each service, the actual score range attested in the outputs differs across services.[76] To help comparability, the scores were scaled to a range between 0 and 1 using *min-max* scaling (p. 141).

A first thing to notice is that in the *min-max* scaled scores each annotator is covering a distinct range, reflecting its original range. To improve on the comparability of the scaled confidence scores, it would be relevant future work to test a scaling method that takes into account the distribution of each service's original scores, besides each service's original score range.

5.3.5.2 Using coherence scores

These scores correspond to our Coh_corpus measure, whose purpose is to quantify the extent to which a concept annotated in the corpus is related to the main corpus topics (pp. 19, 142). A low corpus-level coherence indicates that the annotation is likely an error. The score is meant to complement confidence scores as a means to assess annotation quality. As discussed in following, our measure worked fine to identify incorrect annotations of type *Organization* and *Concept*, but not so for types *Person* and *Location*.

In the case of terms of type *Organization* and *Concept*, our corpus-level coherence measure Coh_corpus teases apart satisfactorily results not thematically coherent with the corpus vs. results that are coherent with it. Consider the mention *Gemstone*. In the corpus, *Gemstone* is a brand for one of the high-risk financial products considered to have played a role in triggering the crisis. This was sometimes disambiguated by Entity Linking as a publishing company (*Gemstone_Publishing*), and in some cases as *Gemstone* in the sense of a precious stone. Both disambiguations are wrong, and the terms proposed by Entity Linking are unrelated to the corpus themes.[77] Accordingly, the corpus-level coherence measure has given both annotations a very low score (below 0.1).

[76] The ranges attested were: For TagMe2, between 0.10 and 0.96; for Spotlight, between 0.10 and 0.60 and for Wikipedia Miner, between 0.40 and 0.97

[77] *Gemstone* in the sense of a very specific financial product is absent from the target Knowledge Base (Wikipedia/DBpedia); it is not surprising that the disambiguations are wrong.

A similar case would be *Nielsen_ratings*, which can be seen in Figure 5.16 (Item 4). This is an incorrect disambiguation for mention *ratings*. In the corpus, this has the sense of *credit ratings*, but it has been disambiguated by EL as the name of an audience ratings agency. The corpus-level coherence score suggests the error, as it is very low.

In the case of person names and locations, our corpus-level coherence measure does not give uniformly correct results, as it provides too low scores even for entity annotations coherent with the corpus' themes. For instance, the score for annotation *David_Viniar* is very low, even if this entity does fit thematically in the corpus: He was an officer at Goldman Sachs, one if the banks who played a major role in the crisis. An interested reader could verify these and similar cases on the interface.[78]

For person names, the source of these errors is the following: The target knowledge-base (KB) for Entity Linking was DBpedia/Wikipedia. Several people that are important in the corpus (frequently mentioned in it) do not have a Wikipedia entry, and mentions to some of those frequent person names are wrongly disambiguated as an existing Wikipedia term. E.g. the mention *Tom Casey*, for a former officer at the failed Washington Mutual bank, and who is not mentioned in Wikipedia, is erroneously tagged as a diplomat with the same name, who does have a Wikipedia page: *Tom_Casey_(diplomat)*. This concept (the diplomat) is not thematically related to other important persons in the corpus, who come from the finance and business domain. Corpus-level coherence scores for a concept (p. 143) rely on assessing the relatedness between that concept and a set of frequent concepts in the corpus. However, as in the example just given, some of the frequent person names in the corpus are incorrectly disambiguated as Wikipedia terms unrelated to the corpus domain. As a consequence, the corpus-level coherence scores for person annotations are unreliable.

It would be useful for users if important corpus actors that do not have a Wikipedia entry were also identified by our application. We discuss how this could be implemented, as future work, below (p. 158).

As regards locations, it would require further analysis to determine why the coherence scores are generally low. A possible reason is the following: Our notion of coherence relies on common inlinks in the Wikipedia link graph (see p. 142). The Wikipedia pages for organizations and technical terms mentioned in the corpus, which are from the finance and government/regulation domains, are likely to receive common incoming links from other Wikipedia pages about those domains. By contrast, corpus locations do not have much

[78] Filtering results by choosing entity types *Concept*, *ORG*anization, *PER*son and *LOC*ation on the select box and pressing the `Refine Search` button displays results illustrating the strengths and weaknesses just mentioned.

in common other than being majoritarily US locations and common inlinks to corpus locations may be scarce. Accordingly, our coherence measure may not be appropriate to capture whether a given location is a plausible disambiguation for this corpus.

In summary, the examples mentioned for term types *Concept* and *Organization* suggest that it is useful to have a measure of the general relatedness of an annotation with representative annotations in the corpus overall. However, for term types *Location* and *Person*, our implementation does not perform well, and it would be relevant to improve this in future work.

5.3.5.3 Examples of automatic annotation selection

The automatic selection (p. 148) is based on the weighted voting procedure from Chapter 3 (p. 60ff.) This combines the results of a set of Entity Linking tools, selecting those annotations for which there is sufficient agreement among the tools.

A formal evaluation of automatic annotation selection would require creating reference Entity Linking annotations for the PoliInformatics corpus, and comparing results obtained on the reference set by each individual tool vs. the automatic selection based on the tools' combined outputs. The combination and selection procedure was assessed in Chapter 3 (p. 66ff.) with four pre-existing, publicly available test corpora, noting improvements in the combined selection vs. each individual tool's output. However, we cannot take for granted that those improvements will generalize to other corpora. Besides, the setup used for the interface is different to the one previously evaluated. For the interface, three annotators were used, whereas the procedure in Chapter 3 combined five annotators.[79]

I did not create a manually annotated reference corpus to evaluate Entity Linking on the PoliInformatics corpus, in the interest of time. This would be relevant future work. I have nevertheless collected informal evidence about the performance of the automatic selection by inspecting its results on the interface. Figure 5.17 shows the initial results vs. the automatically selected ones, for organizations, between frequencies 100 to 50 approximately. The initial results contain several errors like *Gemstone_Publishing*, *Italian_Socialist_Party*, *United_States_federal_courts*[80] or *Portfolio.com*. These

[79] The reason for this difference is that we implemented the interface, without the *Auto-Selection* tab, and with results for three annotators, before the combination procedure. After I had implemented the combination, for time reasons, I did not rework the interface and the database feeding it to integrate the additional annotators used in the combination workflow, and performed the result combination for the *Auto-Selection* tab based on the results that were already available in the database.

[80] Most annotations for this concept are wrong, as can be seen by looking in the database; in most cases, this concept occurs as an annotation for the mention *federal*.

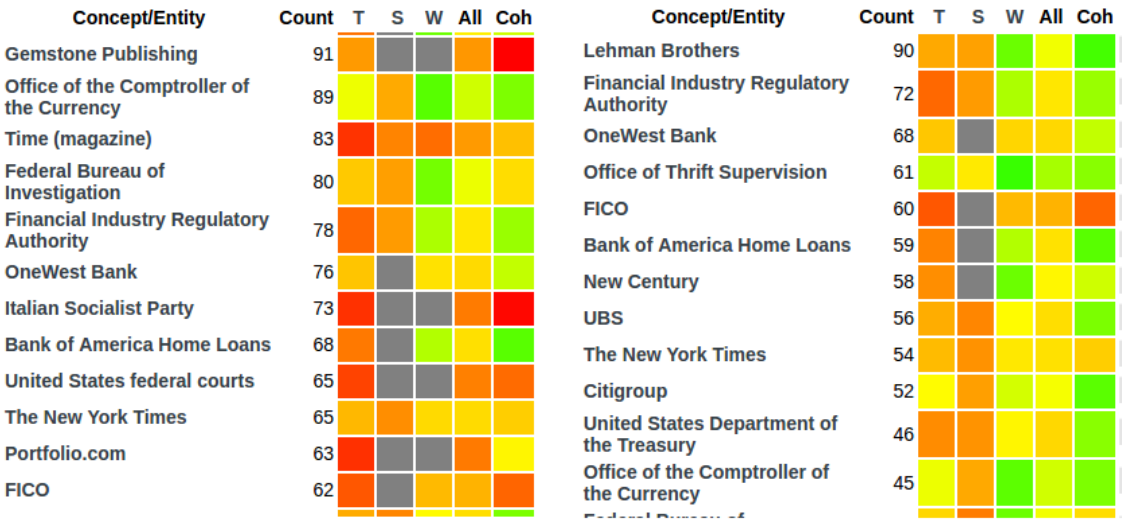


FIGURE 5.17 – Original results (left) vs. automatically selected results (right) for type *ORGanization*. The automatic selection has excluded wrong annotations, like *Gemstone Publishing*, *Italian Socialist Party* or *Portfolio.com*. The correct DBpedia concept *Federal Bureau of Investigation* was part of the selected results, but at a lower frequency, therefore not seen on the screen capture for *Auto-Selection*. Such results are preliminary evidence for the usefulness of the automatic selection.

erroneous annotations have been correctly filtered out by the automatic selection. Note that the correct DBpedia concept *Federal_Bureau_of_Investigation*, seen in the initial results, was also part of the selected results, but at a lower frequency (41), therefore not seen on the *Auto-Selection* screen capture on the right, where the lowest frequency is 45. Finally, *Time_(magazine)*, which was an incorrect annotation in the initial results, decreases from frequency 83 in the original results to 3 in the automatic selection.

Such results are informal evidence suggesting the usefulness of the automatic selection. However, this is only preliminary evidence, based on a small number of examples for term type *Organization*. To draw solid conclusions about the performance of the procedure on this corpus, we would need to assess results quantitatively, against a sufficient number of reference annotations of several types for this corpus.

5.3.5.4 Validating a corpus network

An organization network for the corpus was created, performing the following exercise: It was attempted to follow the steps that a user wishing to get an overview of organizations in the corpus might take, in order to interpret the network, and to evaluate it using the information on the interface about annotation quality (described in 5.3.4.1). This may serve as an indication of the potential usefulness of the UI in this respect.

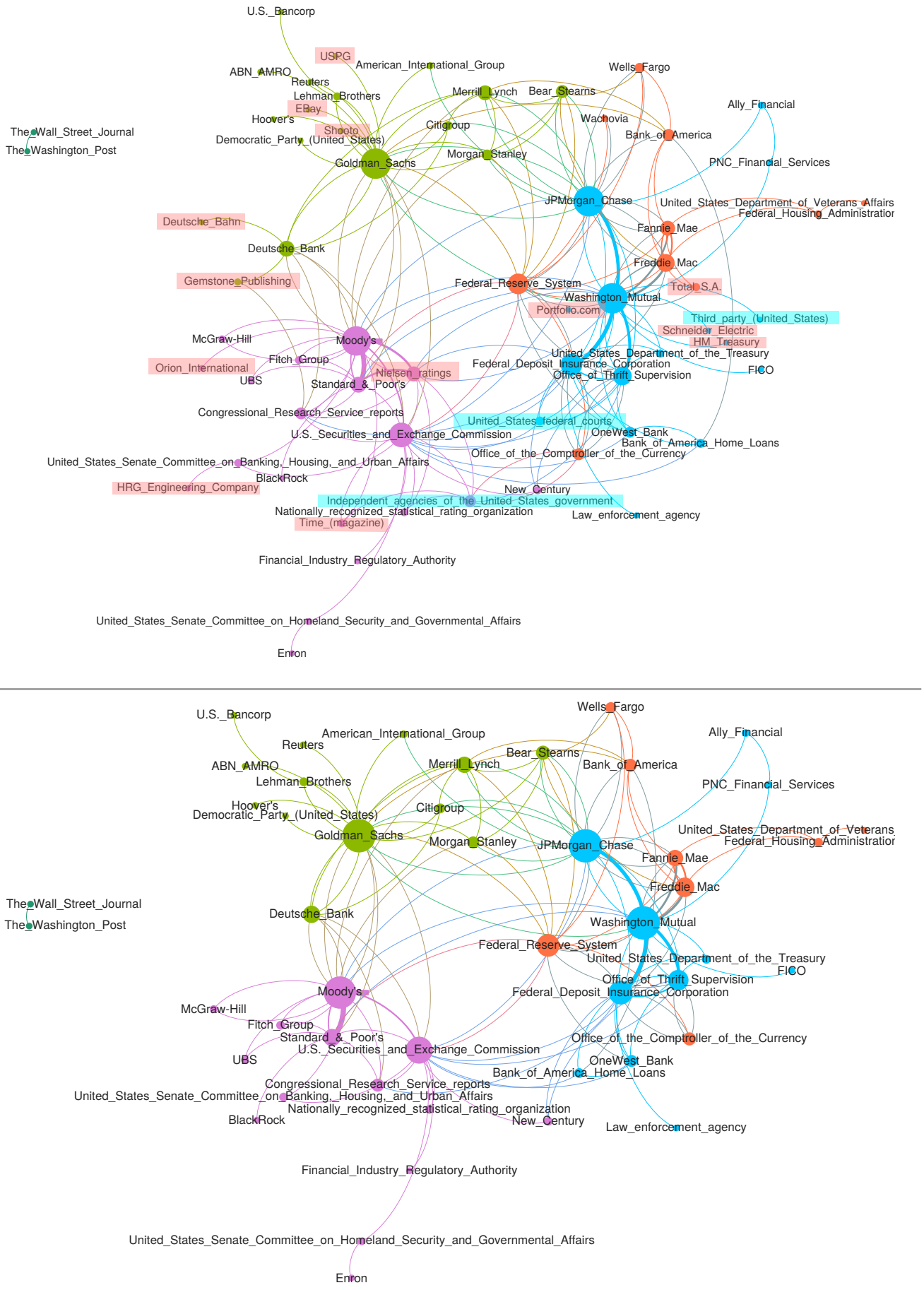


FIGURE 5.18 – PoliInformatics Corpus Network, representing the cooccurrence of organizations in the same sentence. In the top network, incorrect entities, found with the help of the UI (p. 157), are highlighted in red and blue. In the bottom version incorrect entities have been removed.

I first describe the characteristics of the network and the network creation process. Then, I provide my interpretation of the network clusters, and comment on how information on the UI helps assess whether given network nodes are valid or not.

In the network, nodes represent organizations, and edges encode their co-occurrence within the same sentence. Edge-weight corresponds to co-occurrence frequency; the minimum frequency considered was 4. The nodes and edges are listed in an appendix (p. 216). The network was visualized with Gephi, spatialized with the Force Atlas layout (Bastian et al., 2009; Jacomy et al., 2014). The layout settings are also in the appendix (p. 224). Community detection (i.e. identifying sets of highly interconnected nodes) was performed with Gephi's modularity tool,[81] which applies the Louvain method (Blondel et al., 2008).

Direct co-occurrence is a very simple measure of proximity between corpus terms. The assumption is that terms that are mentioned together within a given stretch of text behave in a related manner in the corpus. The proximity measure applied in the networks for the Bentham corpus (p. 114) is more advanced; it is an indirect measure, based on words in common in the context of the network's terms as an indication of term proximity. This indirect measure can provide more nuanced results than direct co-occurrence. For instance, two companies that are mentioned in the corpus as offering the same types of products would be drawn close to each other in the network, even if they are never mentioned together. However, for the PoliInformatics networks, I chose direct co-occurrence for the following reason: The point of the exercise was to see how the information on the UI can help validate the network, not to obtain the most nuanced representation of the corpus possible. And obtaining data for direct co-occurrence was faster than using more advanced methods.

Node labels in the network are DBpedia term labels. Co-occurrence is based on mentions to those concepts in the corpus, i.e. variants that have been disambiguated by Entity Linking as referring to those DBpedia terms. The mentions need not match the label exactly. For instance, *Washington_Mutual* is often referred to as *WaMu* in the corpus.

As regards how the network was created, the interface does not currently allow an export of the entities a user wishes to select, and it does not show co-occurrence information, which would be required to create the type of network just described. However, the information is available in the database feeding the interface, which I used to create the network. As future

[81] https://github.com/gephi/gephi/blob/master/modules/StatisticsPlugin/src/main/java/org/gephi/statistics/plugin/Modularity.java

work, it would be useful to make such information accessible for users on the interface.

My interpretation of the network clusters is the following. The clusters represent main actors in several aspects of the financial crisis. In general terms, the cluster composition is as follows:

1. *Investment banks (green)*: Either active ones like Goldman Sachs or failed ones like Lehman Brothers.
   Risky investment products sold by some of these banks are considered to have contributed to the crisis.
2. *Mortgage-related (orange)*: Consumer banks, who were issuing mortgages (Wells Fargo, Bank of America), and the two Government Sponsored Enterprises whose role is to ease consumers' access to mortgages (Fannie Mae, Freddie Mac).
   The basic relevance of this cluster is that loan defaults were a factor in the crisis.
   The Federal Reserve (the Fed) is part of this cluster, as well as a Treasury agency called the Office of the Comptroller of the Currency (OCC). This may be due to their role in bank supervision.
3. *Failure of Washington Mutual (blue)*: Contains this failed bank, plus JPMorgan Chase, who bought it, and the governmental organizations who oversaw the process, like the Department of the Treasury and the Office of Thrift Supervision, who put Washington Mutual into the receivership of the Federal Deposit Insurance Corporation.
4. *Ratings and Regulation (purple)*: The cluster contains rating agencies like Moody's or Standard & Poors. Also, bodies related to regulation, like the Securities and Exchange Commission (SEC), who regulates securities trading in the US, or the Senate Committee on Banking, Housing and Urban Affairs. Other bodies involved in regulation (the Fed and the OCC), in the orange cluster, are linked to nodes in this cluster.

Whereas the general area of each cluster seems clear, it may not be clear whether some of the entities contained in them are correct or not. The quality measures on the interface, besides the full-text search index to find corpus contexts containing the annotations, can help decide about those unclear cases. If we look at the concept-pane entries for the entities I have highlighted in red on the network, we'll see that the UI suggests that they

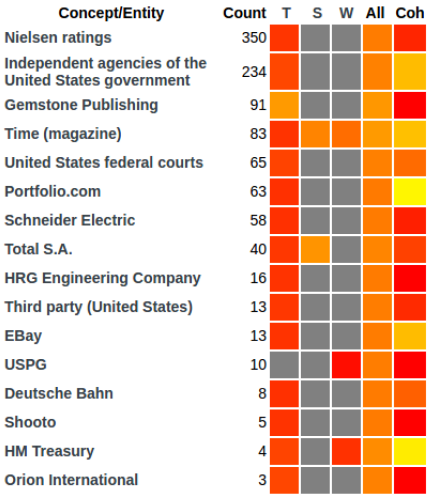


FIGURE 5.19 – Low-annotation quality indicators on the UI for these concepts from the network in Fig. 5.18 suggest they are errors.

are likely incorrect (Figure 5.19), because the number of annotators having produced them is generally only one (out of three), their average confidence is low and their coherence scores are also generally low. For a description how to read the information on Fig. 5.19, see p. 145.

A function that is not currently available on the UI, but that would help users validate the corpus annotations in more detail, would be to give access to not just the concept and its confidence score averaged over all mentions, but also to the mentions themselves, including their sentence of occurrence. This information, available in the database, would be helpful in order to more easily validate cases like those highlighted in blue in the network (p. 155), e.g. *United_States_federal_courts* or *Independent_agencies_of_the_United_States_government*. Their overall scores are low, however, they seem plausible concepts for the corpus—more plausible for me than other incorrect entities like a publishing house (*Gemstone_Publishing*) or an electric supplies company (*Schneider_Electric*). Looking at the database, we can see that the most frequent mentions for the two plausible-looking examples above are *federal* and *agencies* respectively. These mentions are ambiguous, and, whereas in some cases the corpus shows that the annotations are correct, in other cases they refer to other uses of the word *federal* or to other types of agencies. Enabling access to mentions and their contexts on the UI would be useful future work in this respect.

5.3.5.5 A limitation: Actors unavailable in the knowledge base

Our application annotates and displays corpus actors found in Wikipedia (or DBpedia). It would be informative for users to have access to other relevant corpus actors, even if they cannot be linked to a Wikipedia page. As discussed on p. 152, several people important in the corpus are not covered by Wikipedia; coverage for organizations was better. To annotate person-names unavailable in the knowledge base, the method in Coll Ardanuy et al., (2016a; 2016b) could be applied, as it was created specifically for that purpose, i.e. being able to find coreferential person-names across documents, independently of their presence in a target knowledge-base. Their method integrates several sources of information in order to decide whether two person mentions should be considered as coreferential: The inherent ambiguity of the person name (based on lists of first, middle and last names), the probability that up to three person names within a given inherent ambiguity range are coreferential across two documents (based on a development set), and similarity of context vectors for the person-name mentions. The advantage of this method over others reviewed at the same work is that it has been tested multilingually; otherwise an approach reaching similar efficacy for English would be Rao et al. (2010).

Another way to potentially resolve person-name references for people not covered by Wikipedia would be by applying a domain-specific knowledge-base. Frontini et al. (2015) created a tool (REDEN) to simultaneously disambiguate against generic KBs like DBpedia and domain-specific ones. However, I am not aware of a readily available domain-specific KB listing the types of people relevant for the PoliInformatics corpus (e.g. American bank officers). The US Government's open data portal (`data.gov`) provides information about banks, but not their officers. Such a knowledge base would need to be created from corporations' financial statements or professional directories.

5.3.6 Summary and Outlook

A detailed summary of our PoliInformatics case-study follows. Future work possibilities, which were already mentioned in the chapter, are also outlined for several points in the summary as relevant.

Work to create a user interface (UI) for the PoliInformatics corpus was presented. The interface[82] shows DBpedia terms annotated in the corpus by three Entity Linking tools. These terms can be used as facets to navigate the corpus. Full-text search is also available.

A goal of the UI is to help users find relevant actors and concepts in the corpus. To this end, several **annotation quality attributes** are displayed for each DBpedia term: Its corpus frequency, a confidence score for each of the tools that have output the term, and a corpus-level coherence score. The confidence score represents how likely the term is correct. The corpus-level coherence score indicates the extent to which the term is thematically related to a set of terms considered representative for the corpus overall. Low scores for these attributes suggest that the annotation is likely incorrect. As a second source of information to validate the terms, the documents in which each term was annotated can be accessed on the UI by choosing the term as a facet.

Providing annotation quality indicators to guide manual term selection, based on the output of Natural Language Processing (NLP) tools, responds to a need identified by social scientists we have collaborated with. The measures we implemented were shown to be useful in some cases, but they also show some shortcomings.

As regards **corpus-level coherence scores**, examples showing how they are useful were provided: The scores managed to tease apart a set of incorrect organization names proposed by Entity Linking from correct ones. Some examples of incorrect annotations flagged by low corpus-level coherence

[82] `http://apps.lattice.cnrs.fr/nav/gui/`

scores were also provided in the case of terms of type *concept*, expressed by common nouns like *warehouse* or *gemstone* (pp. 151, 157). However, whereas for terms of type *organization* and *concept* the coherence scores were satisfactory, for term-types *person* and *location* a non negligible number of scores were low even for terms thematically coherent with the corpus (p. 151). As future work, it would be relevant to find a better measure of coherence that would handle all annotation types adequately.

As for the **confidence scores** we provided, they are based on scores output by each of the Entity Linking tools we integrated. Each tool outputs scores in a different range. With a view to facilitating comparison across tools, we scaled the scores to a range between 0 and 1, using a linear transformation called *min-max* scaling. The comparability of scores obtained with this method was limited, since distribution of scores in the scaled range is still affected by the original score distribution. It would be relevant future work to test other scaling methods.

Besides search and annotation attributes that can help for manual term selection, the UI also shows the results of an **automatic annotation selection**, which rejects terms for which there was not sufficient agreement among tools. Some examples were provided showing how this annotation selection manages to reject some incorrect organizations, while keeping correct ones (p. 153). The automatic selection procedure had been tested in Chapter 3 on other corpora, with positive results. However, to ascertain the performance of this method on the PoliInformatics corpus, it would be necessary to manually create reference annotations for the corpus. This was not done for time reasons, but it would be interesting future work.

A limitation of the application is that it does not treat **NIL mentions**: In other words, corpus actors which are not present in the knowledge base used for entity linking (Wikipedia/DBpedia) are not annotated. Improving on this would be useful future work. Possible methods to annotate such actors were outlined (p. 158), e.g. clustering coreferential actor mentions, or trying to create and apply a domain-specific knowledge-base.

In terms of **implementation** aspects of the UI, the current version allows us to see the main features of the approach, but several improvements would be relevant. First, entity search is currently based on entity labels, not on both the labels and the variants for each entity as attested in the corpus, and available in the database. Entity highlighting in the texts also has similar limitations. The implementation was simplified in the interest of time, but it would be useful to improve it in the future. Displaying the mentions for each term on UI (not just the term labels) would also be useful, especially since it would help assess the validity of each term.

As a second implementation aspect to improve on, it is not currently possible for users to store terms they wish to select, and to export the selected set of terms. This would be necessary future work for users to exploit the results of their work with the interface for purposes like creating networks of the corpus. Currently, storing or exporting a subset of terms can performed through queries on the backend database only.

Finally, our **evaluation of the interface** could be improved: Obtaining domain-experts' feedback on the interface, as we did for the other applications in the thesis, would be a relevant validation exercise to perform in the future. Strengths and weaknesses were discussed, but they were based on our own examples of use of the interface rather than on domain-expert feedback.

Chapter 6

Relation-based Corpus Navigation: The *Earth Negotiations Bulletin*

6.1 Introduction

This chapter presents an application to navigate the climate conference negotiation reports in the *Earth Negotiations Bulletin* (volume 12), which covers yearly climate summits since 1995.[1] As this is a negotiation corpus, it is important to know not only what points are being addressed in the negotiation, but also who is addressing them, and whether actors are voicing support or opposition regarding an issue. To this end, actors' statements were analyzed with the system described in Chapter 4 (pp. 78ff), which extracts propositions, i.e. triples formed by actors, their messages and the predicate relating both, relying on semantic role labeling and on a domain model with actors and predicates. Proposition messages were enriched with automatically annotated metadata: keyphrases, DBpedia terms and concepts from a thesaurus about climate issues. An interface makes all the information navigable.[2] A user can search for propositions (as well as sentences and documents) for a given actor, or for propositions containing a given predicate or predicate-type (support, opposition or neutral). The metadata extracted from the messages can be displayed on the interface to get an overview of issues addressed by actors. Issues over which actors agree or disagree are also shown on the interface. The application is an example of how Natural Language Processing (NLP) technologies can be exploited to obtain analyses that go beyond establishing the cooccurrence between actors and concepts, providing information about the nature of the relation between them. A domain-expert evaluation with three climate policy experts suggested the relevance of the approach for their research needs.

[1] http://enb.iisd.org/enb/vol12/
[2] http://apps.lattice.cnrs.fr/ie/uidev/

The chapter is structured as follows: The corpus and preprocessing steps are described in 6.2. Related work is discussed in 6.3. The NLP components used for analyzing the corpus are introduced in 6.4. This includes a brief summary of the proposition extraction system which had been presented in Chapter 4 and an account of the keyphrase extraction and entity linking tools we employed for annotating metadata in the propositions' messages. The user interface is described in 6.5, and examples of its use are provided. The domain-expert evaluation is in 6.6. Finally, 6.7 summarizes the work discussed in the chapter and possible future work.

6.2 Corpus Description

This section describes the original corpus, our preprocessing to make it amenable to treatment with NLP tools, and the corpus sample we analyzed. An alternative version of the corpus created by other researchers, that could complement ours, is also presented.

6.2.1 The *Earth Negotiations Bulletin*

The *Earth Negotiations Bulletin* is a publication covering international climate policy negotiations. Its 12th volume covers the yearly Climate Change Conference of the Parties (COP) summits, besides other meetings. The volume had about 620 issues at the time we compiled the corpus (August 2015). For our analyses we focused on the issues covering COP summits, which at the time were 258 issues. COP meetings consist in international climate policy negotiations between countries, country groups, and other interest groups. International climate policy treaties like the Kyoto Protocol from 1997 or the Paris Agreement from 2015 get negotiated at COP meetings.

The ENB corpus provides reports on participants' statements at the negotiations. It strives for an objective tone and enforces the use of a specific set of reporting predicates, which can be considered non-interpretive of participants' intentions. For instance, *objected* or *stated* rather than *attacked* or *accused*. The corpus tends to use a limited variety of syntactic structures, to avoid featuring some participants more prominently than others.

The corpus is published by the International Institute for Sustainable Development (IISD).[3] Corpus editors are experts in climate policy, with related post-graduate degrees.[4]

[3] http://www.iisd.org/
[4] http://enb.iisd.org/about/team/

6.2.2 Corpus sample in our study and corpus preprocessing

Each of the ENB issues covering COPs is either a daily report for the negotiations, or a summary report for the complete COP. The summaries for the complete COP tend to reproduce large parts of the content of the daily issues, but aggregated as a single document. For that reason, we based our analyses on the daily issues, and excluded the summaries. If we had included summaries, this could misrepresent the information given in the corpus, since many propositions, keyphrases and concepts would be duplicated, annotated once in the daily issue where they occur and a second time in the related summary. After excluding summaries, we had 235 issues to analyze instead of the original 258 COP issues.

Our sample covers 23 climate summits, starting with a meeting in New York in 1995 called the INC or Intergovernmental Committee for a Framework Convention on Climate Change, which served to prepare the first COP on the same year. The last COP covered is the Geneva one in February 2015. The corpus size is approx. 23,700 sentences (505,000 words).[5]

The original corpus format at the time we crawled it was plain text or PDF for some issues, and HTML for most issues.[6] We crawled the text and HTML versions. The original text versions were normalized to remove hard line-breaks (i.e. newline characters that do not correspond to the end of a paragraph, but that had been inserted for lines to respect a fixed length in characters). In HTML versions, the markup was removed to yield clean text. For each file, we created a plain-text version and an XML version. The plain-text versions were used to run NLP tools on them. The XML ones were used for indexing in the Solr search server.[7] Several metadata for each issue were extracted from the table of contents for the corpus:[6] The date for each issue, its conference number and the conference location. These metadata were added to each file in our XML version of the corpus, and kept in a standoff metadata file for our plain-text version. Our clean corpus is publicly available online.[8]

Note that, besides our version of the corpus, another public version of the corpus[9] was created by Venturini et al. (2015), who have performed research on it and who created an interface to navigate the corpus (see 6.3.3).

[5] This is the figure after excluding the summaries for the reasons mentioned above. With the summaries, the size of the corpus goes up to ca. 41,400 sentences and 950,000 words.

[6] http://enb.iisd.org/enb/vol12/ The site was updated after we had crawled the material. Currently HTML and PDF versions are available for all issues.

[7] https://lucene.apache.org/solr/. The Solr server was already described on p. 117 and its XML input format is documented at https://wiki.apache.org/solr/UpdateXmlMessages.

[8] https://bitbucket.org/pruizf/enb/src/master/out/

[9] http://www.climatenegotiations.org/assets/data/ClimateNegotationsBrowser-ENB-verbatims.csv.zip

Their corpus version was not publicly available when we created ours. As explained in their documentation,[10] they split corpus documents into several sections, based on HTML markup in the original corpus. Those sections were tagged with several metadata. For instance, the topic of each section, which they had previously identified manually in earlier research (Venturini et al., 2014). The sections were also tagged for the actors mentioned in them (but without specifying their role in the section, i.e. whether they favoured or opposed a statement is not annotated). The corpus covers the remaining issues in volume 12 of the ENB, besides the COP issues which we covered. The corpus contains metadata created or verified manually by a team of several researchers. These metadata could be an interesting complement to the information we display on our corpus navigation application (see section 6.5).

6.3 Prior Approaches to the Corpus

Prior work on this corpus includes the application of language technologies and network visualization for its analysis (Venturini et al., 2014; Salway et al., 2014; Baya-Laffite et al., 2016) as well as the creation of a user interface to navigate it (Venturini et al., 2015). This section surveys such work.

6.3.1 Corpus cartography

Venturini et al. (2014) applied a corpus cartography process to the corpus, similar to our approach to the Bentham corpus, discussed in Chapter 5. Like we did for the Bentham corpus, they used CorText Manager, a platform that performs lexical extraction and network visualization.[11] With CorText Manager, they first performed a keyphrase extraction. Based on it, experts selected relevant terms to represent the corpus content. Concept co-occurrence networks for those terms were created with CorText. Besides, experts established several thematic areas based on the term-clusters generated by CorText. Each paragraph was tagged for those thematic areas based on terms contained in the paragraph, and graphs showing the temporal evolution of those thematic areas in the corpus were created. The approach yielded useful results. For instance, the climate policy literature has spoken of a turn in later COP meetings from discussing measures to mitigate climate change to proposing measures for adapting to its consequences. However, the analysis of the evolution of lexical clusters in Venturini et al. highlighted that this description is inaccurate, since discussions around how to *finance* adaptation to climate change were already present in the early COPs. A similar approach to the corpus was applied by Baya-Laffite et al. (2016).

[10] `http://www.climatenegotiations.org/about.html`
[11] We described this platform in 5.2.3.2 (p. 113).

Whereas useful results were arrived at, the approach does not systematically provide information about which actors in the negotiations are addressing which topics and in which manner, as no syntactic or pattern analysis that would allow acquiring such knowledge was performed. Offering such information is one of the goals of our corpus navigation application (6.5).

6.3.2 Grammar induction

A study that automatically extracts some information on how actors stand with respect to an issue was carried out by Salway et al. (2014); this work was already introduced in Chapter 4. They applied a technology called *grammar induction*, which permits identifying patterns containing actors and issues. Patterns are inferred in an unsupervised manner from the corpus. Several iterations are performed, to induce increasingly abstract patterns. Some of the example patterns reported in Salway et al. (2014, Table 1) connect an actor and the verb introducing their statement. More abstract patterns cover some cases of sequences relating actors and issues. In this sense, the system provides information about how actors speak about the issues addressed and how actors stand with respect to each other. The study analyzes sequences of type *Country A, supported by Country B* (or *opposed by*), which correspond to one of the patterns induced by the system. Also, the study compares actors' negotiation positions. The comparison is based on the text following the verb in sentences which match a pattern relating a country and a reporting verb. To judge from the examples provided in the study, the approach focuses on identifying patterns that indicate the presence of propositions, rather than propositions themselves understood as ⟨*actor, predicate, message*⟩ triples. However, some of the more complex patterns could be the basis for extracting propositions and it would be interesting to see how their approach complements our outputs.

6.3.3 Corpus navigation

An interface to navigate Vol. 12 of the ENB was created by the Paris Sciences Po médialab among other partners.[12] The interface gives access to the version of the corpus described on p. 165, created by the same team. Unlike our corpus version, which includes Conference of the Parties (COP) reports in Vol. 12 of the ENB, the médialab's corpus includes COPs and the remaining meetings covered in the volume. Corpus documents are divided into sections according to headings in the original HTML files (see p. 165). The sections were tagged with metadata describing the section content: First, the countries or country groups mentioned in the section, besides the conference and city. Second, topics prominent in the section, based on term lists

[12] http://www.climatenegotiations.org/explore/#/?

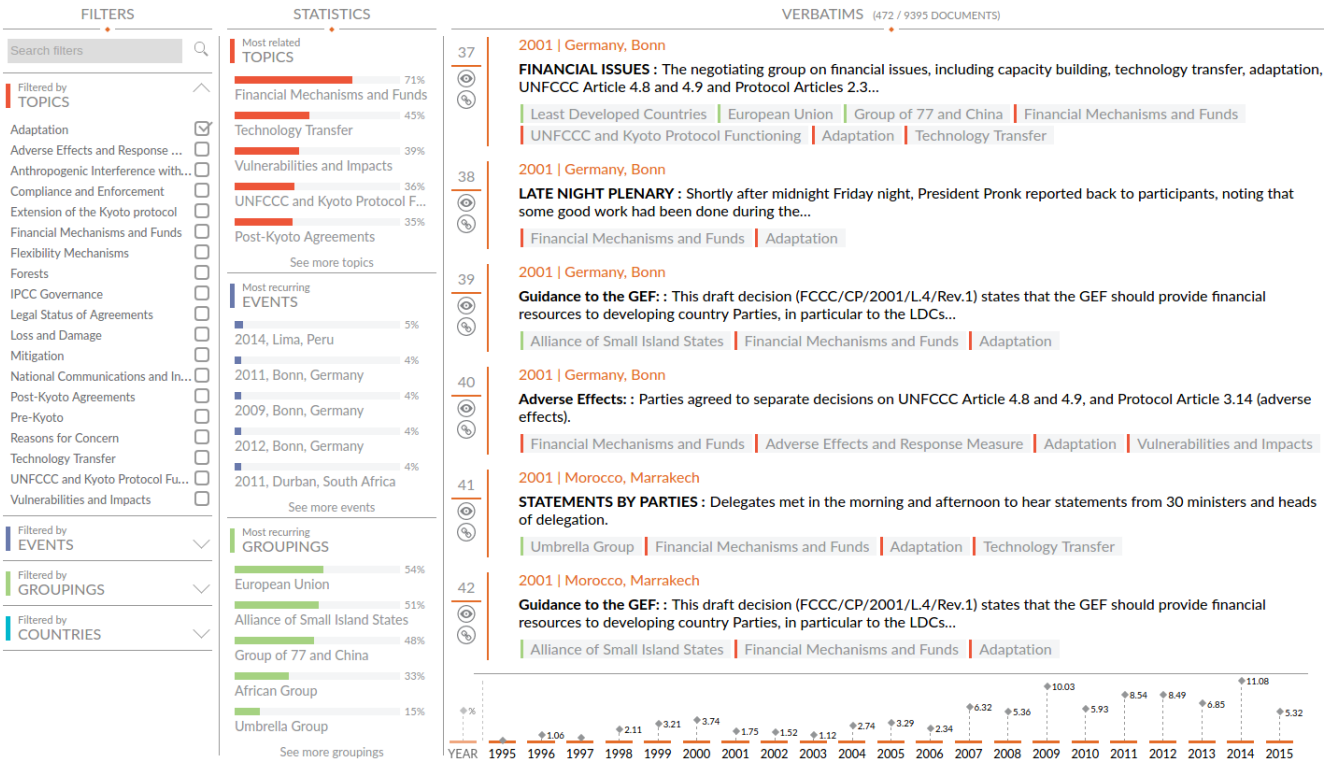


FIGURE 6.1 – The Sciences Po médialab (and partners) created this interface to navigate the *Earth Negotiations Bulletin*, vol. 12. Results for facet query *Adaptation* are shown. The FILTERS pane contains metadata facets. The STATISTICS pane shows the most frequent topics, conferences and country groups in results matching the query. The VERBATIMS pane shows these results. The screenshot shows the initial view for a query's results, with snippets for each result item, but the section and document for each item can also be displayed. Below each result snippet, the interface shows actors and topics previously tagged by experts in document sections matching the query. The bottom pane shows the distribution of results per year. The screenshot corresponds to `http://www.climatenegotiations.org/explore/#/?topics=Adaptation`

representing each topic, that had been created by domain-experts (cf. their earlier research like Venturini et al., 2014) and that were matched against the section. The metadata are used as facets on the interface; the result set can be restricted to records matching a set of facets. The metadata are searchable, but not the full text of the documents: For each search query, the interface returns document sections whose metadata match the query.

The médialab's interface is shown on Figure 6.1. The leftmost panel (*Filters*) displays the metadata facets and the rightmost one (*Verbatims*) shows the corpus records matching a query. The interface has several result aggregation and overview functions. For each query, sections in the mid panel (*Statistics*) show the most frequent country groupings, conferences and topics in the result set; the topics are based on a list created by domain-experts, as explained above. A panel below the results shows the temporal evolution per year of records matching the query.[13]

[13]The interface also has another view, under the *Discover* heading (`http://www.climatenegotiations.org/`). The *Discover* tab is however not a corpus navigation tool. It is a set of visualizations of the results of experts' analyses of prominent actors and their evolution in the corpus, along with experts' comments on the results.

Some differences between this interface and the one we created, presented in 6.5, are the following. The former offers metadata-based search and returns document sections whose metadata match the query, also giving access to the complete document. Some of the metadata had been created manually by domain-experts. By contrast, our interface offers full-text search against proposition elements (actors, predicates, messages) and corpus sentences. The queries return propositions, sentences or documents. Metadata for the messages emitted by actors are also annotated in our interface, but they have been extracted with automatic means: keyphrase extraction, entity linking to DBpedia, and domain-specific tagging with a thesaurus. As regards overview functions, the médialab's interface offers a panel depicting the temporal evolution of records in the result-set, and such a function is not available in our interface, although it would be a valuable addition. Integrating the expert-created metadata available in the médialab's corpus could also be useful, to complement our automatic metadata.

6.4 NLP Backend: Proposition Extraction and Enrichment

The goal of the application presented in this chapter is to help analyze actors' statements in climate negotiations, thanks to a user interface that allows navigating the corpus, integrating the results of several NLP analyses. A system architecture diagram is in Figure 6.2.

The first component of the workflow which generates the analyses exploited in the interface consists in **proposition extraction**. Actors' statements are formalized as propositions, i.e triples consisting of the actor, its message, and the predicate relating both. Our system to extract propositions was already described in Chapter 4, and a summary is provided below in 6.4.1.

The second component consists in **enriching the propositions** with metadata that describe the content of their messages. Once propositions have been extracted, automatically generated metadata are added to their messages: Keyphrases, DBpedia concepts and terms from a climate thesaurus. These metadata, and the tools providing them, are described in 6.4.2 below.

The **reason for enriching** proposition messages with those metadata is the following: The interface intends to provide an overview of issues that actors address, showing whether actors speak favourably of an issue, or express opposition towards it. Keyphrases and entities extracted from the propositions' messages are considered to reflect issues addressed by an actor. In response to a user query on the interface, the metadata can then be aggregated over a result-set, in order to provide an overview of messages in the result set. For instance, if we select propositions where a given actor is using predicates of opposition, keyphrases and other metadata from the

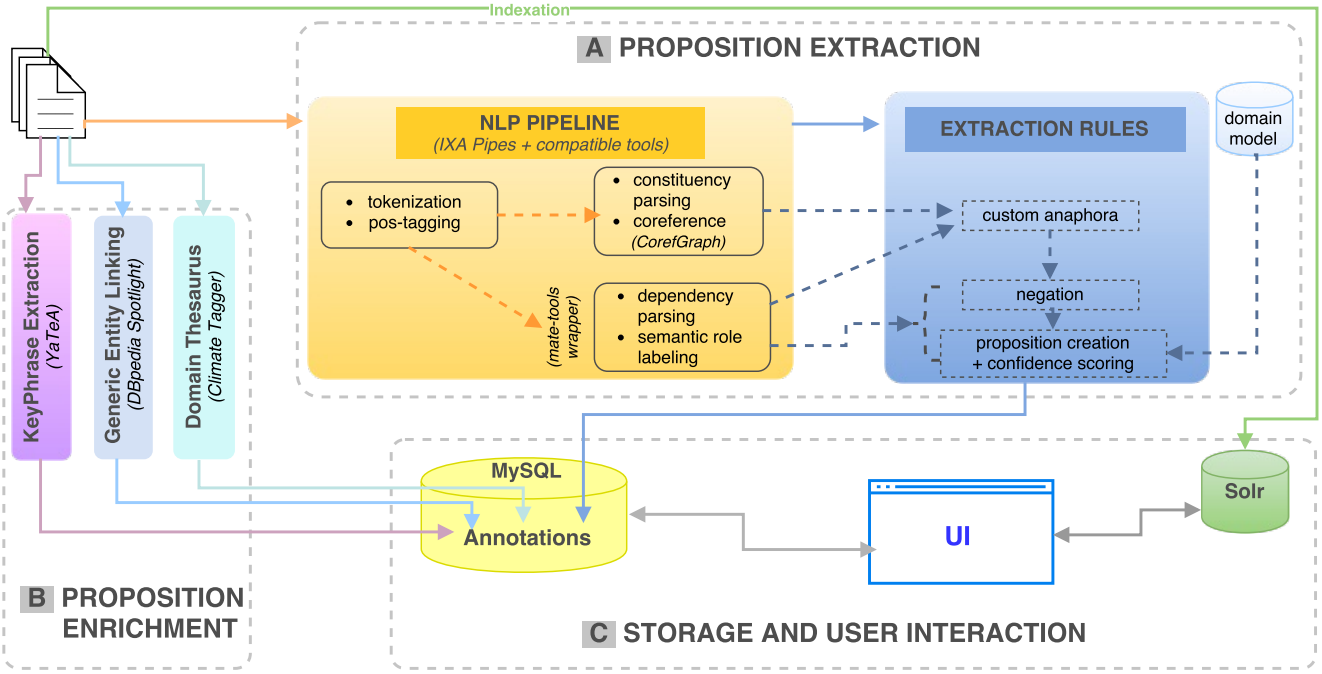


FIGURE 6.2 – System Architecture for Corpus Navigation via Enriched Propositions. **A. Extraction:** As seen in Chapter 4, an NLP pipeline provides Semantic Role Labeling, dependency parsing and coreference chains. Propositions (i.e. ⟨*actor*, *predicate*, *message*⟩ triples) are extracted with rules based on the NLP output and on a domain model. **B. Enrichment:** The proposition messages are annotated with keyphrases, DBpedia concepts and Climate Thesaurus concepts. These metadata will allow users to compare actors' statements. **C. Storage and User Interaction:** The corpus is indexed in Solr, and the different annotations generated for it (propositions, keyphrases, etc.) are stored in a MySQL DB. Users access the information on a Django interface at `http://apps.lattice.cnrs.fr/ie/uidev/`

propositions' messages give an overview of issues towards which the actor shows a negative attitude. Similarly, these overviews can be provided for propositions in sentences showing agreement or disagreement among actors, as an indication of issues over which actors have conflicting views. This will be discussed further in 6.5.1 and 6.5.2.

6.4.1 Proposition extraction

The proposition extraction workflow was described and evaluated in Chapter 4 (p. 78ff). A system architecture diagram for the complete corpus navigation application which exploits its results, and which is the object of the current chapter, is provided in Figure 6.2. This shows how proposition extraction (component **A** in the figure) integrates in the complete application.

A proposition was defined as a triple of shape ⟨*actor*, *predicate*, *negotiation point*⟩, where *negotiation point* is the message emitted by a participant at the negotiation, i.e. the *actor*. The *predicate* is a reporting verb or noun whereby the message is emitted.

The proposition extraction workflow is summarized below—more details are available in Chapter 4 (p. 78ff), and pointers to that chapter are provided in the summary here. An NLP pipeline performs Semantic Role Labeling

(SRL) on the corpus (4.4.1). A domain-model containing predicates and actors is used to find relevant predicates in the SRL output (4.4.2). The actor emitting a message, and the message itself, are identified among the roles assigned by SRL to arguments for predicates from the domain-model (4.4.3). Negation is also identified, using SRL roles and using surface cues (p. 82). As some actors are expressed by pronouns, some cases of pronominal anaphora are resolved based on coreference chains output by the NLP pipeline, but applying custom rules given non-standard pronoun use in the corpus; these rules also exploit dependency parsing (p. 83). Finally, propositions receive a confidence score which reflects their expected informativeness. Propositions whose speaker is not an actor in the model are also output, with a smaller confidence. Incomplete propositions, lacking a message, or with uninformative messages, are equally output, but with a very low confidence score (see p. 84). The location and date for each proposition correspond to the conference and day where they were uttered, which were extracted as metadata when the corpus was crawled and cleaned up.

The user interface (UI) in 6.5 allows navigating the corpus making queries for each proposition element, besides filtering by date and confidence score.

As regards evaluation, proposition extraction was evaluated intrinsically against a manually annotated reference set. The F1 score for exact match of all proposition components was 0.69. The qualitative evaluation in 6.6 for our *Earth Negotiations Bulletin* navigation interface, which exploits these propositions, showed that the extraction quality was sufficient for domain-experts using the interface to explore the corpus.

6.4.2 Enriching proposition messages with metadata

As stated on p. 169, the reason to annotate the proposition messages with keyphrases and concepts is that these metadata will allow us to get an overview of the content of propositions matching a user query on the interface. Three types of metadata were annotated, as described below.

6.4.2.1 Keyphrase extraction

Keyphrases are noun phrases that represent important notions in the corpus. They are defined according to part-of-speech patterns and frequency criteria. They correspond to a corpus-driven extraction, without reference to external knowledge. In this sense, they can provide a more detailed view of the corpus than knowledge-based metadata like those provided by entity linking (6.4.2.2) or by tagging the corpus against a thesaurus (6.4.2.3), since they are not limited by how well external knowledge resources cover the corpus content.

Keyphrase extraction was performed with the Yatea tool (Aubin et al., 2006). The tool was already described on p. 112, as we also used it to analyze the Bentham corpus (5.2.1). This tool works in English and French and we used it since we had successfully applied it in previous work (Mélanie et al., 2015). However, newer tools for keyphrase extraction exist, e.g. Keyphrase Digger by G. Moretti et al. (2015). It would be interesting to test such newer tools on the corpus as future work. The choice of a keyphrase extraction tool for the current implementation was not critical, since it was planned to manually inspect the keyphrases and create a filter for bad quality ones. The filter is a list containing uninteresting or ill-formed keyphrases found in Yatea's output for the corpus; items from this list are not displayed on the user interface.

6.4.2.2 Generic-domain entity linking

Entity Linking (EL), which was introduced in Chapter 1, is a technology for identifying references in a corpus to terms from a knowledge repository, in order to abstract away from variability in the way the terms are expressed in the corpus. For instance, *Marie Skłodowska* and *Mme Curie* can be identified by EL as referring to the same concept[14] in the DBpedia knowledge base by Auer et al. (2007). Another objective of EL is finding the correct referent for corpus expressions ambiguous across several referents. For example, *M. Curie* in a given context should be disambiguated as *Marie Curie*, taking *M.* as an initial. In other contexts, the same sequence should be disambiguated as *Pierre Curie*, taking *M.* as the French abbreviation for *Monsieur*.

The task of disambiguating mentions to "conceptual" terms, generally expressed by common nouns or noun phrases, is sometimes called *Wikification* in the literature, and the name *Entity Linking* is then reserved for the task of disambiguating named entities. These are lexical sequences belonging to specific categories like *organization*, *person* or *location*, which are generally expressed by proper nouns (p. 16). In this thesis I am using the names *Wikification* and *Entity Linking* interchangeably, as argued on p. 16.

Generic-domain EL refers to performing EL against general knowledge bases (KBs). For instance, DBpedia (Auer et al., 2007), which expresses the content of Wikipedia using semantic web formats. Other general repositories are Yago (Suchanek et al., 2007) and Babelnet (Navigli et al., 2012).

Entity Linking tools were surveyed in Chapter 1 (p. 18ff). For the ENB corpus, we need to identify KB terms that are expressed by common noun phrases, not exclusively by proper nouns. One of the publicly available

[14]The DBpedia terms for the examples in this paragraph can be browsed at `http://dbpedia.org/page/Marie_Curie` and `http://dbpedia.org/page/Pierre_Curie`

tools that obtains good results for KB-terms expressed by common nouns is DBpedia Spotlight (Mendes et al., 2011; Daiber et al., 2013), which annotates against the DBpedia knowledge base. We used the DBpedia Spotlight web service (with default settings) to annotate the ENB corpus.

The complete document text was sent to the web service, since Entity Linking needs the context around a term mention in order to disambiguate it. However, the results displayed on the user interface (6.5) need to be restricted to those terms whose mentions occur in the negotiation points of propositions. We cannot display terms from anywhere else in the sentence, because we want to use these terms as an overview of messages expressed by actors, not as an overview of the content of the corpus overall. Entity Linking results were postprocessed to limit the set of terms displayed on the UI to those terms mentioned within the proposition points.

Generic-domain entities can express general important notions in the corpus, but may miss the more specialized ones. This is why, in addition to entity linking to DBpedia, we also tagged proposition messages with a thesaurus specialized in climate topics.

6.4.2.3 Domain-specific thesaurus

To complement keyphrase extraction and generic entity linking, proposition messages were also annotated with a domain-specific thesaurus, specialized on energy and climate. A thesaurus is a type of controlled vocabulary. Thesaurus concepts are expressed by a preferred variant, and alternative formulations for the same concept are linked by a synonymy relation. Other relations usually specified in thesauri are broader terms (hypernyms) and narrower terms or hyponyms (Jing et al., 1994).

The specialized thesaurus we applied is the *Climate Thesaurus* (Bauer et al., 2011).[15] It covers clean energy and a set of climate change management practices known as "climate compatible development". As such, terms related to climate policy relevant for analyzing climate negotiations are part of the thesaurus.

A specialized public web service exists to annotate text with this thesaurus. It is called the *Climate Tagger API*.[16] To annotate the corpus, we ran requests against the API's `extract` service.[17] Among other information, the service returns Climate Thesaurus terms relevant for the text, besides a confidence

[15] `http://www.climatetagger.net/climate-thesaurus/`. The thesaurus has also been known as the *Reegle Thesaurus*.

[16] `http://api.climatetagger.net/` *API* stands for *Application Programming Interface*, i.e. a set of commands for communication between programs.

[17] `http://api.reegle.info/service/extract`

score between 0 and 100 for each term. We kept terms with a confidence score higher than 5.

To ensure that the terms annotated by the API belong in the point mention of the propositions we had extracted from the corpus, we actually sent to the web service the propositions' point mentions only, rather than the complete proposition or the complete sentence, unlike what we did for DBpedia Spotlight (6.4.2.2). Spotlight benefits from access to the complete sentence, since it needs context in order to disambiguate concept mentions. However, disambiguation is not necessary for annotating occurrences of a thesaurus' terms: The terms and term-variants in the thesaurus have an unambiguous reference, and finding them in the corpus proceeds by string matching.

The goal of using a domain-specific thesaurus was complementing not only generic entity linking, but also keyphrase extraction. Keyphrase extraction can give a detailed overview of a corpus' content, as it is corpus-driven and not affected by the coverage of corpus terms in an external knowledge resource. However, it tends to extract phrases of limited variety in terms of the part-of-speech sequences composing them. A domain-specific thesaurus will tag technical terms even if the part-of-speech pattern expressing them does not match patterns usually considered by keyphrase extraction. An example is the term *common but differentiated responsibilities*.[18] This refers to developed countries' greater role in having caused climate change as a result of industrialization, which gives them an increased responsibility in managing climate change compared to developing countries. Keyphrase extraction identifies a related term (*historical responsibilities*), but not the other term just mentioned, which contains the conjunction *but* and would be untypical as a keyphrase.

6.5 User Interface: Corpus Navigation via Enriched Propositions

The user interface (UI)[19] makes the *Earth Negotiations Bulletin* (ENB) corpus navigable, using full-text search, and via a structured search based on the annotations generated by the Natural Language Processing (NLP) backend for proposition extraction and enrichment, described in 6.4.

The UI allows researchers to examine statements by an actor, filtering them by predicate type (opposition, support or neutral reporting) and providing an overview of actors' messages, thanks to keyphrases or concepts extracted from them. The default view of the UI can be seen in Figure 6.3.

[18] http://www.reegle.info/glossary/3109
[19] http://apps.lattice.cnrs.fr/ie/uidev/

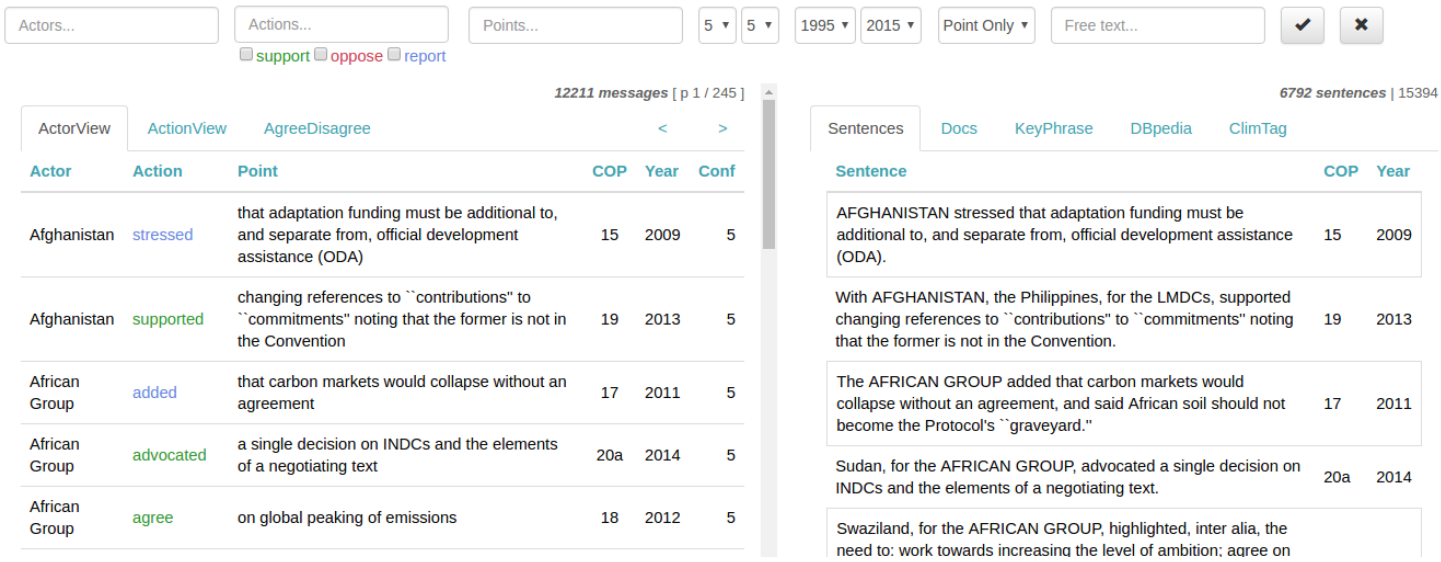


FIGURE 6.3 – Main View of the User Interface for the ENB Corpus. The left pane displays propositions, and each tab on the right pane displays different types of information. First, sentences and documents where the propositions matching a query have been extracted. Second, keyphrases, DBpedia concepts or Climate Thesaurus terms extracted from messages in propositions matching the query. The queries can be performed from the boxes at the top of the UI. On the left, boxes *Actors*, *Actions*, *Points* search in proposition elements. Predicate types can be selected from the checkboxes below the *Actions* box (*support*, *oppose*, *report*). The *Free text* box (far right) searches in sentences. Results can be filtered by confidence and by date, with the drop-downs in the middle of the top row.

This section describes the search and navigation workflows available on the interface. In 6.5.1, search based on proposition elements and on sentences is introduced. These search functions correspond to tabs *ActorView* and *ActionView* on the UI. In 6.5.2, a way to browse the corpus based on agreement and disagreement between actors is presented; on the UI, this workflow is available on the *AgreeDisagree* tab. Finally, some technical details about the UI implementation are provided in 6.5.3.

6.5.1 Search Workflows: Propositions, sentences and documents

The search workflows described in this subsection involve a query against proposition elements, against sentences, or a combination of both query types. For each type of query, the results returned, and details about result highlighting and sorting are discussed below.

6.5.1.1 Proposition queries

This refers to queries on the *Actors*, *Actions* or *Points* search boxes, and the checkboxes for predicate types *support*, *oppose*, *report* (top left in Fig. 6.3).

Expected usefulness: These queries can be performed in order to gain an overview of an actor's statements. They can help answer questions like *what are common issues in propositions where a given actor uses a verb of opposition?* Or *what predicates does a given actor use most?*.

Proposition panel results: Proposition queries return, on the **left panel**, propositions matching the query, as will be detailed below. The *Actions* box searches against proposition predicates; the checkboxes below it select a predicate type. The *Points* box searches in the message expressed by the actor in each proposition.[20] The search terms are considered using an `AND`-logic.

These queries have some limitations. They do not allow wildcards for query expansion or boolean operators. Improving on these limitations would be useful future work. The way search terms are currently processed is the following:[21]

- **Actors**: Propositions are returned whose actor name contains the query string. Besides, a mapping was created for common actor variants. For instance, query term *group* returns propositions for both *African Group* and *Group of 77*. And query *UK* (which is in the variant map) returns propositions for actor *United Kingdom*.
- **Actions**: Propositions are returned whose predicate starts like the query term. E.g. if we enter *add*, we'll retrieve propositions whose predicate is *added* or *adding*.
- **Points**: The query term (or the term + *s* or + *es*) must match exactly one of the words in the proposition point. E.g. *approach* matches propositions whose point contains *approaches* or *approach*. Hyphens count as a word boundary, i.e. *water* matches *water-related*.

The initial **sort order** for proposition results depends on two factors: First, the tab that is active on the proposition pane (*ActorView* or *ActionView*). Second, the search boxes that were used for the query.

In *ActorView* (default tab), the sort order is first *Actor*, then *Action*. In *ActionView*, the order is first *Action*, then *Actor*. The *ActionView* tab opens by clicking it, or when the query contains a term in the *Actions* box.

Besides the initial sort order, clicking the headings for each column in the proposition table sort the propositions according to that column. This can be useful to sort results by year, for example. It is the complete set of propositions in the results that gets sorted, not just the ones for the page currently displayed.

Right panel results: For proposition queries, on the right panel, each tab returns different types of information:

[20] Recall from 4.2.1 (p. 74) that a proposition is defined as a triple of shape ⟨*actor*, *predicate*, *point*⟩, where point is the message emitted by the actor via the predicate.

[21] Proposition queries run against a MySQL database using Django built-in field-lookup operators (`https://docs.djangoproject.com/en/1.10/ref/models/querysets/`). Unlike for the Sentence queries (6.5.1.2), for the Proposition queries we're not using a search index, which would make wildcard expansion and boolean operators easier. This could be improved as future work.

- **Sentences**: The sentences where propositions matching the query have been annotated. Terms entered in the *Actors* and *Actions* boxes are highlighted in the sentence if they can be exactly matched against its text. Terms in the *Points* box are not highlighted in the sentence.
- **Docs**: Clicking on this tab displays snippets for documents where the propositions matching the query have been annotated. If we click on one of the propositions on the left, the sentence containing it will be highlighted in the document text and scrolled into view. This is useful to provide users with the document context in which the sentence, and propositions extracted from it, need to be understood.
- **KeyPhrase**: Displays keyphrases extracted from the points (i.e. the messages) for propositions matching the query. The keyphrase extraction process was described in 6.4.2.1.
- **DBpedia**: Displays DBpedia concepts extracted from the points of propositions matching the query. The process of Entity Linking to DBpedia was described in 6.4.2.2.
- **ClimTag**: Displays Climate Thesaurus terms annotated in the points of propositions matching the query. The process of tagging against the Climate Thesaurus was described in 6.4.2.3.

The items returned on the right pane are clickable. Clicking an item displays on the left pane propositions corresponding to that item. *KeyPhrase* items are always highlighted in the proposition point. For ClimTag and DBpedia terms, highlighting only happens for terms which can be exactly matched against the proposition point.[22]

The set of sentences and documents displayed on the *Sentence* and *Docs* tabs corresponds to the propositions for the page currently displayed on the left pane. However, keyphrases, DBpedia and Climate Tagger (*ClimTag*) terms have been aggregated over the complete set of propositions matching the query, not just over the 50 propositions displayed on each page.

Result counts: As regards result counts, they are displayed on the top right angle of both the left panel and the right panel of the UI. On the left panel, the number of propositions matching the query is displayed, besides the current page and total number of pages. On the right panel, the counts displayed depend on the tab.

[22] Besides incomplete highlighting, there are two other imperfections in the functions of the proposition table returned on the left pane when a keyhprase, DBpedia concept or *ClimTag* term are clicked. First, the table headers are not sortable, unlike in the normal proposition table returned by Proposition queries (i.e. queries from the proposition search boxes, described in 6.5.1.1). Second, clicking a proposition in this table can only display the sentence where it was found; we cannot toggle between the *Sentence* and *Docs* tab on the right pane to see its sentence and document context, unlike for propositions returned via Proposition queries. The implementation was simplified for time reasons and could be improved as future work.

For the *Sentences* tab, depending on the type of query, one or two different result counts are provided. For Proposition queries (6.5.1.1), the single count shown indicates the number of sentences retrieved for propositions matching the query. For Sentence queries (6.5.1.2), two values are displayed. The smaller value corresponds to the number of sentences in the result set for which propositions matching the query have been extracted. The larger value stands for the number of sentences matching the query term in the *Free text* box, whether propositions have been extracted from them or not.

In the *Docs* tab, the number of documents with propositions matching the query is displayed.

Other remarks: Recall the method used by the proposition extraction workflow (4.4.3) to represent propositions in which some actors have explicitly opposed others, i.e. propositions in sentences containing sequences like "opposed by" (Fig. 4.1 on p. 175 shows an example). These opposing propositions were represented as a "virtual proposition" where the predicate is a negated version of the main verb's predicate, and the point is the same as the main proposition point.

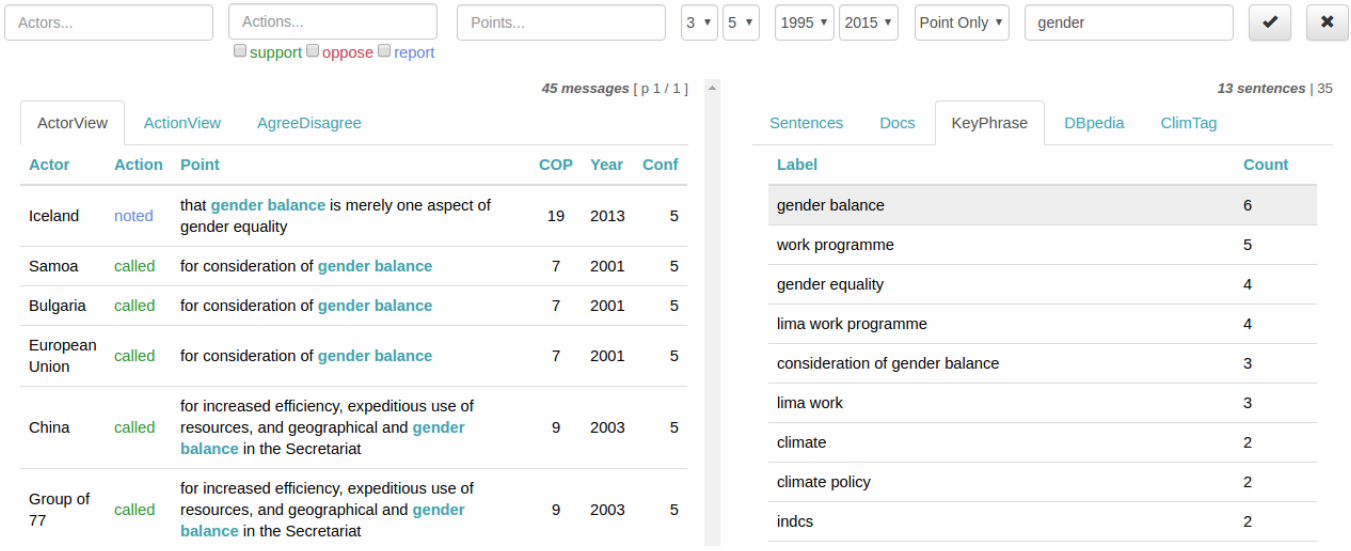
In Figure 4.1, a tilde sign (~) indicates the negated version of the predicate. However, in the system, this negated version of the predicate was represented by preceding it with *not*, just like a normal negation. This could create confusion between predicates that are originally negative in a sentence, and predicates preceded by *not* as a result of creating a "virtual proposition" to express opposing actors' views.

In practice, there is a way to tease apart both types of cases; this was verified by inspecting the proposition-extraction results. In the propositions extracted, predicates that are **originally negative** occur with an infinitive verb (*not* + *infinitive*). By contrast, **"negated virtual predicates"**, that result from applying a proposition creation rule for opposing actors, occur with a *past-tense, gerund, or nominal predicate*. In future work, predicates that are negated to indicate an opposition could be represented differently to actual negation, to eliminate any risk of confusion.

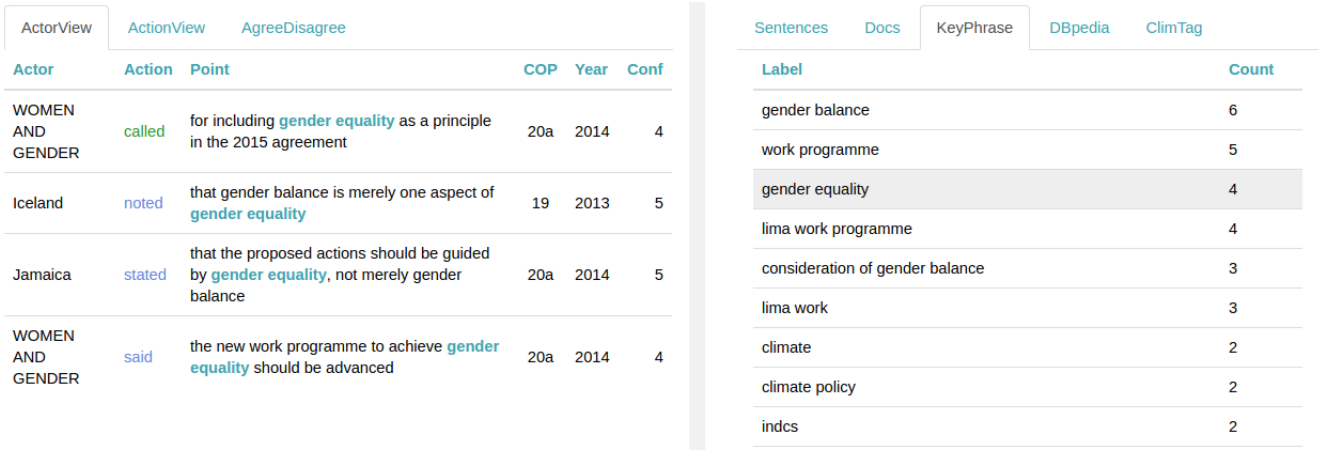
6.5.1.2 Sentence queries

Sentence queries are performed through the *Free text* box on the right of the top row of the UI. An example, using *gender* as the query term, is in Figure 6.4.

Expected usefulness: These queries can help examine which actors are represented in sentences mentioning a given search term, and what predicates they are using. Besides, the metadata tabs on the right pane (keyphrases etc.) provide an overview of the content of those actors' messages in sentences



A. Propositions matching *gender balance* are selected



B. Propositions matching *gender equality* are selected

FIGURE 6.4 – ENB UI: Sentence query. *Gender* was searched in the *Free text* box. The left pane shows propositions extracted from sentences where *gender* occurs. On the right pane, keyphrases extracted from the message of those propositions are displayed. In the top image, propositions matching *gender balance* are selected. In the bottom image, propositions matching *gender equality* are displayed. With the keyphrase-based overview of the messages, we see that certain countries, besides non-country actors, make a stronger statement than others, speaking of *equality* rather than *balance*. Proposition confidence is between 3 and 5, so that non-country actors like the *Women and Gender* group are also shown (see p. 84ff. for details on confidence scores).

where the term appears. E.g. *which actors participate in sentences where* gender *is mentioned? What concrete issues around* gender *are being addressed by whom?* Figure 6.4 illustrates this example.

Result description: Sentence queries run against the Solr search index where the corpus is stored, and boolean operators and wildcards can be used, following Solr syntax.[23]

The result set contains, on the **left pane**, propositions extracted from sentences matching a query term. On the **right pane**, the sentences themselves are displayed by default, with search terms highlighted. The metadata tabs contain items extracted from the messages of the propositions for sentences matching the query.

The **sort order** for proposition results for a Sentence query is as explained on p. 176 for Proposition queries. Regarding the sort-order of sentences, it corresponds to the sort-order of the propositions extracted from them. Sentences for which no propositions were extracted get sorted after sentences which have propositions.

Sentences matching the query will be clickable if propositions have been extracted for them. By clicking on the sentence, its propositions will be displayed on the left pane. The last sentences on the sentence list may not be clickable, if no propositions were extracted for them.

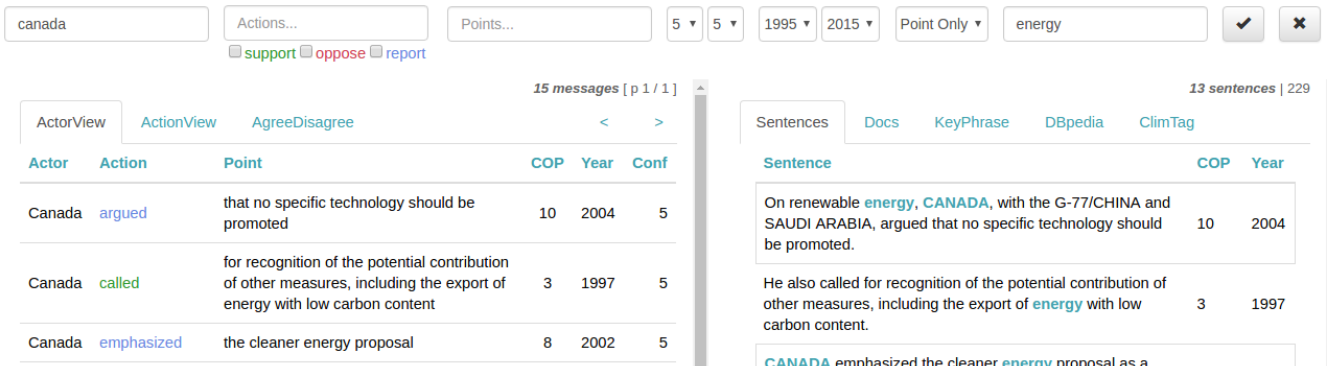
6.5.1.3 Combined Proposition and Sentence queries

Queries on the proposition boxes (*Actors*, *Actions*, *Points*, and the predicate-type checkboxes) can be combined with a query term on the *Free text* box.

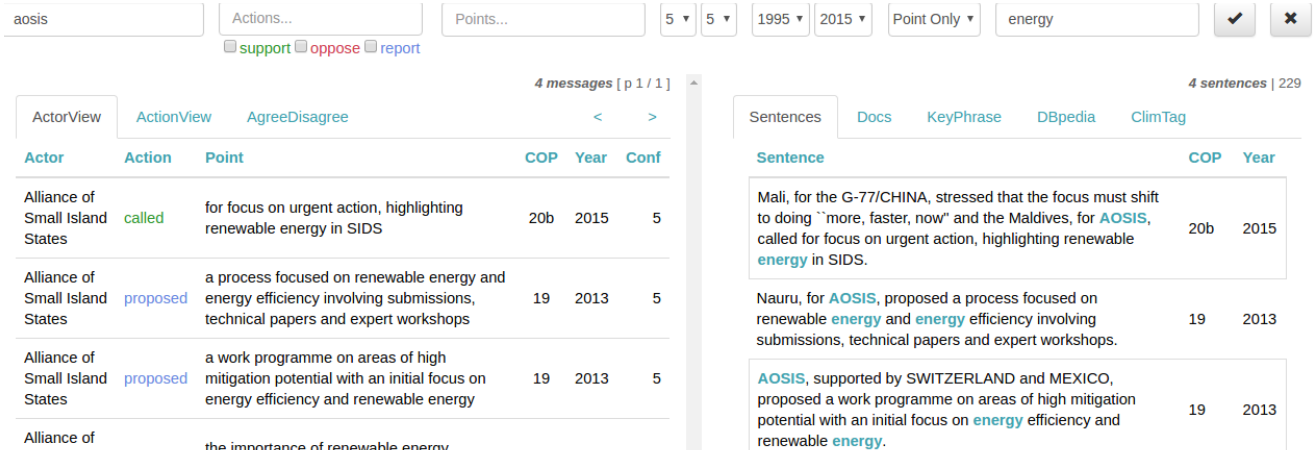
Expected usefulness: Since the term on the *Free text* box is searched in the whole sentence, combining a Sentence query with a Proposition query is useful to make up for possible errors in delimiting propositions' points. We can inspect visually the results to verify if the query term is in the point or not. And metadata (keyphrases and concepts) are always extracted from the proposition points only, so the overview we get for the proposition points in the result set does consist in terms that actors have addressed via the predicates specified in the proposition, not of terms mentioned elsewhere in the sentence.

Result description: Highlighting and the sort order for propositions and sentences, as well as sentence clickability, is as described on p. 180 for Sentence queries.

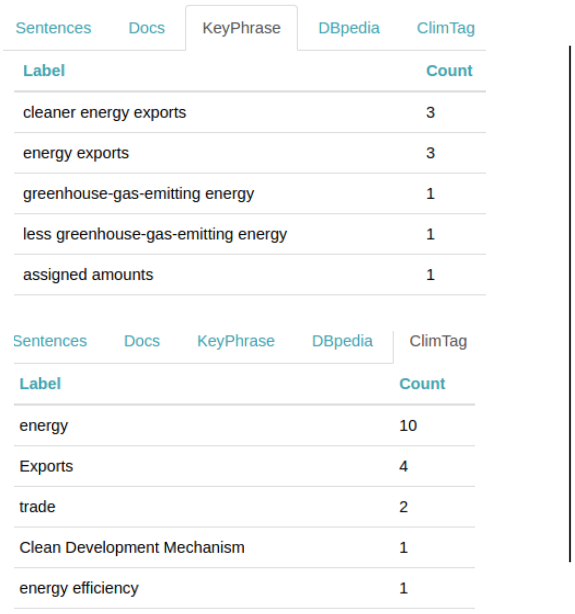
[23] For the boolean and wildcard searches allowed, see `https://lucene.apache.org/core/4_0_0/queryparser/org/apache/lucene/queryparser/classic/package-summary.html`. We tested boolean queries with *AND* and *OR*, wildcards and "fuzzy searches" based on edit distance. Some of the functions, like regular expression searches or field-specific searches are not possible on our interface.



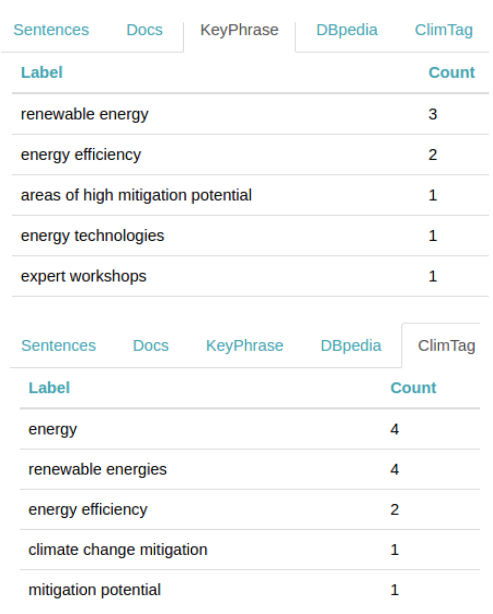
A1. *Canada* in *Actors* & *energy* in *Free Text*. Right pane displays *Sentences* tab



B1. *AOSIS* in *Actors* & *energy* in *Free Text*. Right pane displays *Sentences* tab



A2. *Canada* & *energy*:
Keyphrases (top) and Climate Thesaurus (bottom)



B2. *AOSIS* & *energy*:
Keyphrases (top) and Climate Thesaurus (bottom)

FIGURE 6.5 – ENB UI: Combined Proposition and Sentence queries help compare actors' negotiation topics. **A1** and **B1** show queries for propositions whose actor is Canada (A1) and AOSIS (B1) and where the sentence contains *energy*. **A2** and **B2** show the top-5 keyphrases and Climate Thesaurus terms extracted from the proposition points by each actor in those sentences (A2 for Canada, B2 for AOSIS). We see that both countries speak about *energy efficiency*, but only Canada speaks about *exports* of energy, perhaps because it has abundant renewable energy resources.

6.5.1.4 Confidence and date-range filters

Propositions are assigned a confidence score that indicates how informative and how complete the proposition is expected to be, as was explained in 4.4.4. Among other criteria, propositions with a very short point receive low scores. Propositions where the actor is not in the domain model are assigned lower scores than those where the actor is in the model. If pronominal anaphora resolution has applied to identify actors, this also lowers the score.

Confidence scores can be lowered to return a larger set of propositions for a query. This is useful according to the domain experts who evaluated the system (pp. 191, 192), in order to obtain propositions by less commonly studied actors who were not part of the model, and who had not been analyzed in previous studies on the corpus either.

The confidence score range is between 0 and 5. The minimum and maximum confidence for the propositions to display can be set with drop-downs on the top row on the UI, which can be seen on all UI figures in this chapter, e.g. Fig. 6.5. Unless the user selects a wider score range, only propositions with a confidence score of 5 are displayed.

6.5.2 Browsing for agreement and disagreement

The *AgreeDisagree* tab on the left pane of the UI allows us to choose a pair of actors, and a relation type (agreement or disagreement). Metadata (i.e. keyphrases, DBpedia concepts and Climate Tagger terms) extracted from proposition messages over which actors agreed or disagreed are displayed. Clicking on a metadata row will display the sentences for the propositions from where the metadata were extracted. Agreement and disagreement was determined on the basis of explicit occurrences of expressions indicating opposition (e.g. *opposed by*).

The *AgreeDisagree* tab as currently implemented has several shortcomings. First, there is no search on this tab. It would be useful to search for pairs of actors and see metadata extracted from the propositions in which they agree or disagree, rather than having to select the actors first. It would also be useful to be able to select more than one term for *Actor1* and *Actor2*; currently a single item can be selected. Another imperfection is that actor pairs for which there is no data can still be searched, returning a message indicating that no data was found. It would be easier for a user to find relevant information if selectable actor pairs are restricted to those for which agreement/disagreement information is available, e.g. by filtering the set of possible second actors in the pair once the user selects the first actor. These limitations could be improved as future work.

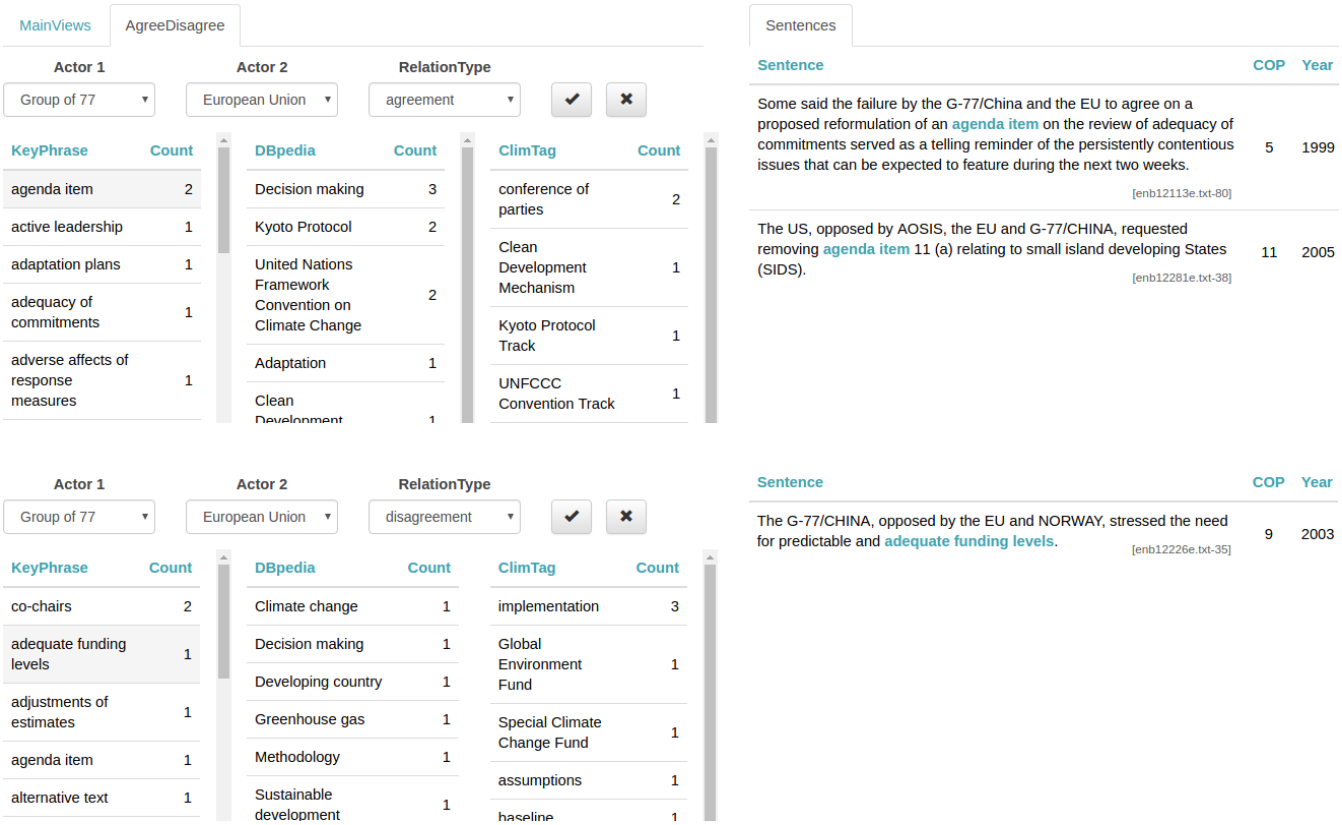


FIGURE 6.6 – Agree-Disagree View for actors *European Union* (developed countries) and *Group of 77* (developing countries). The left-pane columns show metadata extracted from proposition points for which those actors were in agreement **(top)** or disagreement **(bottom)**. Clicking on the metadata displays the sentences where the propositions were extracted. As regards disagreement, we see that funding is one of the issues where both actors have opposing views, as expressed by the keyphrase *adequate funding levels* and some of the ClimTag climate thesaurus concepts.

6.5.3 UI Implementation

Propositions and metadata were stored in a MySQL database. The corpus documents and sentences were indexed in a Solr search server.[24] Results from both data sources are merged thanks to common identifiers for documents and sentences in the database and in the search indices.

The interface was developed with Django,[25] a Python web development framework. This framework makes it possible to query the database using Object Relational Mapping (ORM). Thanks to this, database tables are represented as objects, and Python Object Oriented Programming features can be used to formulate queries, instead of writing raw SQL queries.

The front end uses the Bootstrap library,[26] which allows creating responsive sites, i.e. that adapt to device characteristics. If accessing the UI on a narrow screen or phone, the layout adapts.

[24] https://lucene.apache.org/solr/. The Solr server was already described on p. 117
[25] https://www.djangoproject.com/ Version 1.8.4 was used.
[26] http://getbootstrap.com Version 3.3.5 was used.

UI tabs are populated with AJAX requests, i.e. for faster responses, after a user action, only the components of the UI whose content needs to change are reloaded, instead of the complete interface.[27]

6.6 User Interface Evaluation with Domain-experts

This section reports on an evaluation task with three domain-experts. The section is structured as follows: After an introduction describing the evaluation goals and general characteristic of the approach (6.6.1), the evaluation hypotheses can be found in 6.6.2. The evaluation task is introduced in 6.6.3; basic data about the evaluation sessions (date, length, etc.) are also found there (p. 188). Finally, a summary of the evaluation results is presented, along with a discussion (6.6.4).

6.6.1 Scope and approach

The evaluation sets out to assess to what an extent the corpus navigation application we developed for the ENB corpus helps expert obtain an overview of the corpus, and whether it can help them gain new insights on it.

A qualitative evaluation was carried out, based on interviews of over one hour with three domain-experts familiar with the corpus. The experts were asked to come up with questions about the corpus, use the interface to obtain information relevant to their questions, and comment on the results. It was assessed whether the user interface (UI) helps them gain an overview as well as new insights on the corpus.

The assessment was based on verbal evidence (expert comments) or other behavioural evidence (queries and other operations performed on the UI). This type of broad qualitative assessment can be considered as an exploratory evaluation; some of the results remain open to interpretation. I see the evaluation results as an indication of the usefulness of the tool, rather than its definitive validation or "condemnation". In this respect, consider that user interface evaluation is a complex task. Khovanskaya et al. (2015) offer a though-provoking overview of methodological and epistemological challenges involved. Lieberman (2015) provides informal critical reflection.

A quantitative evaluation of the UI based on Human Computer Interaction notions, like *usability* or *user satisfaction* (e.g. Kelly, 2007; Al-Maskari et al., 2010) is out of scope in this thesis. Such an evaluation would involve defining tasks to perform with the UI, employing a somewhat larger sample of

[27] For instance, when clicking on a keyphrase to get all propositions that mention it, the proposition pane is reloaded to show the new information, but the keyphrase list is not computed again or reloaded. *AJAX* stands for *Asynchronous JavaScript and XML*.

domain-experts than we used. The proportion of tasks completed successfully could be measured, as well as task completion time; a questionnaire could measure user satisfaction. This would be relevant future work to complement the qualitative evaluation carried out here.

6.6.2 Hypotheses

The hypotheses and the types of evidence considered relevant to assess them are presented below. The hypothesis are as follows:

- **Hypothesis 1 (H1):** The information presented on the interface, the way it is presented, is useful to gain an **overview** of the behaviour of different **actors**, or the treatment of **issues** in climate negotiations as portrayed in the corpus.
- **Hypothesis 2 (H2)**: Interacting with the corpus via the UI can give the expert new ideas for research on the corpus or **new insights** on the corpus (i.e. information and possible interpretations the expert was not aware of).
- **Hypothesis 3 (H3)**: The **factual correctness**, and the **coverage**, of the information on the UI is sufficient for the expert's needs to perform research on this corpus.

Hypothesis 1 (H1) and 3 (H3) overlap somewhat: if there is no sufficient coverage of corpus information (H3), it's unlikely that the expert will gain an overview. I don't see this overlap as a weakness. There will just be some evidence that speaks to two hypotheses.
Regarding H3, recall that we had evaluated proposition extraction *intrinsically* (p. 87), obtaining an F1 of 0.69. The issue here is to get an indication whether this quality is enough for a researcher working with this corpus, or whether it "gets in the way".

Evidence relevant for the hypotheses: The hypotheses were assessed based on the experts' spontaneous comments about the UI results, and based on their behaviour with the UI (e.g. the queries they made and functions they used). Examples of experts' behaviours relevant to assess the hypotheses follow.

For any of the hypotheses, the expert's explicit verbal confirmation or denial of the hypothesis counts as evidence in favour or against the hypothesis. For instance, if the expert states that the UI is not useful to gain an overview of an issue, this goes against H1. For such statements, I am looking at *spontaneous* utterances, not statements in response to an experimenter's question that would elicit such confirmation or denial (see 6.6.3 below). I prefer spontaneous utterances for the following reason. As Khovanskaya et al. (2015) report, study participants often form an opinion about the intended

contribution of a study, and they believe that it is helpful for the research if the experiment provides evidence for their expected result. These beliefs can bias participants' behaviour accordingly. In this sense, asking participants for direct feedback about our hypotheses may bias them towards confirming them, more so than eliciting indirect feedback.

Besides explicit verbal confirmation or refutation of the hypotheses by the expert, other verbal or behavioural evidence considered is detailed below.

Hypothesis 1

- **Evidence agreeing:** The expert making generalizations about an actor or issue on the basis of the data he or she obtains through the UI queries.
- **Evidence against:** Expressions by the expert that the information is confusing, incomplete, biased, etc.

Hypothesis 2

- **Evidence agreeing:** The expert making comments along the lines of "I hadn't thought of this, but looking at this information now ...", or "based on this information, something to look more deeply at would be ..."
- **Evidence against:** I consider the absence of such comments evidence against H2.

Hypothesis 3

- **Evidence agreeing:** I consider a lack of recurrent comments by the expert about bad quality (or lack of coverage) of the extractions to speak in favour of H3. The reasoning is that, if extraction quality is not usable by an expert, this will surface in the evaluation; results for too many of the expert's queries will be unusable. Occasional comments about bad extractions are expected, since F1 was 0.69, see p. 87).
- **Evidence against:** Recurrent comments about bad extraction quality or about missing information (results that were expected but not found).

6.6.3 Evaluation Task

As stated, the task consisted in an interview with a domain-expert, who performed queries on the interface and commented on the results.

6.6.3.1 Structure of an evaluation session

The experimenter for all sessions was myself (I am also the main contributor to the interface). The sessions followed the steps below.

1. First I explained to the expert how the corpus was constructed (see 6.2), each of the UI's functions (see 6.5 above), and answered the expert's questions about this.

2. I then showed the expert some examples how the interface can be used to navigate the ENB corpus and answered the expert's questions about this (see 6.6.3.2).
3. I then gave instructions to the expert about the task (see 6.6.3.3).
4. Finally, the expert performed several queries on the interface. A discussion of the query results followed each of the queries. An overview of those discussions, paying attention how they relate to to the evaluation hypotheses, is in 6.6.4 below.

6.6.3.2 Examples shown to experts

The following examples were shown to experts to make them acquainted with the UI functions. Unless otherwise stated, the time range for all queries is 1995 to 2015, and the confidence score range is restricted to a score of 5 (the maximum confidence).

EXAMPLE 1

Search Box: `Free Text`
Search Term: *gender*
Results pointed out to expert: `Sentence` tab and `KeyPhrase` tab, pointing to the different actors mentioning *gender balance* and *gender equality* in their propositions (selecting those propositions by clicking on the keyphrase). A screenshot for this example is in Figure 6.4.

EXAMPLE 2

Search Boxes: `Actors` and `Free Text`.
Search Terms: *Canada* in `Actors`, and *energy* in `Free Text`. This was compared to *China* in `Actors`, also with *energy* in `Free Text`.
Results pointed out to expert: the fact that keyphrases for *Canada* differ from those extracted for *China*. Figure 6.5 shows a similar example.

EXAMPLE 3

Search options: In `AgreeDisagree` tab, actors *Group of 77* and *European Union* were selected, looking at relation types *agreement* and `disagreement`.
Results pointed out to expert: The different keyphrases or DBpedia and Climate Thesaurus concepts extracted for each relation type for the chosen actors.
It was also shown to the expert that clicking on a term in the terms lists displays sentences, which show (dis)agreement between the selected actors, and whose message contains that term. See Figure 6.6.

6.6.3.3 Instructions given to experts

The expert was asked to perform queries on the interface. The following two general types of queries were suggested to the expert:

(a) Queries to verify information he or she knows about the corpus (e.g. about the behaviour of a specific actor)

(b) Queries in order to obtain information they do not still know and would like to obtain from the corpus

Besides these suggestions, the expert was also told that they could use the interface whichever way they wanted.

The instructions were broad and open-ended. This was for two reasons. First, I judged such broad instructions sufficient to obtain data relevant for the hypotheses, which were also defined in broad terms. Second, I was interested in not directing experts' behaviour too much, and see what they would spontaneously do with the UI.

6.6.3.4 Evaluation data storage

Sessions 2 and 3 were audio-recorded, and I later transcribed the audio (non-verbatim); I also took notes during the session. For Session 1, I took notes during the session, writing a complete session-report after it. Since I have no audio for the Session 1, I sent the report to the expert shortly after, for her to confirm whether the report reflected the content of the session correctly. She made some modifications included in the session report in the appendix. See Appendix D for details and links to the audio as well as full session reports.

6.6.3.5 Basic Data about Evaluation Sessions

Basic data about the sessions are provided in Table 6.1. Some details about the domain-experts for each session are below. See Appendix D for more information.

	Session 1	**Session 2**	**Session 3**
Date	06/16/2016	06/24/2016	08/04/2016
Place	Paris SciencesPo	Paris SciencesPo	Skype
Duration	1h 15min	1h 20min	1h
Number of queries	13	20	6
Domain-Expert			
- Knowledge of the corpus	detailed	general	detailed
- Familiarity with corpus navigation tools	strong	strong	medium

TABLE 6.1 – Basic data about evaluation sessions

Kari de Pryck (Session 1): Researcher specialized in climate negotiations and the *Earth Negotiations Bulletin* (ENB) corpus, who has previously published work on these topics and is completing a Ph.D. at Sciences Po in Paris. She is also an assistant at the University of Geneva.

Tommaso Venturini (Session 2): Lecturer at the Digital Humanities Department at King's College (London) and associate researcher at the médialab

at Sciences Po (Paris), at interview time. Some of his research areas are Digital Methods and Controversy Mapping. He has led the EMAPS[28] and MEDEA[29] projects, which studied different aspects of adaptation to climate change.

Nicole de Paula (Session 3): Writer/Editor at the *Earth Negotiations Bulletin* or *ENB* (i.e. the publication that authors the texts analyzed on the interface). She holds a Ph.D. in Political Science/International Relations from Sciences Po Paris. One of her research areas is global environmental governance (climate change and biodiversity).

6.6.4 Results and discussion

First, a discussion of the sessions for each of the three domain experts is provided, taking into account how their behaviour speaks to each of the hypotheses above, plus strengths and weaknesses perceived by the experts.[30] An overall discussion of findings, paying attention to those same aspects, can be found at the end (6.6.4.4).

Full reports for the evaluation sessions, detailing all the queries performed by each expert, and the ensuing discussions, are available in Appendix D.

In the rest of this section, references to the passages in Appendix D on which each statement is based are provided (e.g. *D.1.3.3*).

6.6.4.1 Session 1 (Kari de Pryck)

Overview of corpus (H1): Two spontaneous comments by Kari speak to H1. **(a)** "the tool is a bit *Latourian*,[31] as it helps follow actors in time and regarding the subjects they discussed". **(b)** "you can find clear examples for the typical behaviour of this country" (D.1.3.3). Both comments suggest that the tool is useful to get an overview of an actor's behaviour and examples to document that behaviour.

Gain of insight (H2): Kari developed a new research idea while using the UI (D.1.3.1). Seeing sets of ⟨*actor*, *predicate*, *message*⟩ propositions for different actors, she realized that some actors preferentially emit procedure-related messages (about formal or legal aspects of the negotiation), whereas other actors mostly emit truly policy-related messages. She stated that studying actor profiles in terms of to what an extent they discuss formalities vs. actual policy issues would be a new topic for research. Seeing a systematic

[28] http://www.emapsproject.com/blog/

[29] http://projetmedea.hypotheses.org/

[30] There is of course some overlap between these notions, e.g. a user may see *potential for insight gain* (Hypothesis 2) as a strength of the tool.

[31] Callon, Latour and Law proposed a method of inquiry in social sciences called Actor-Network Theory or ANT (Law et al., 1999; Latour, 2005). The method describes actors' relations within a network as a way to analyze social phenomena.

breakdown of actors and their messages on the proposition-pane of the UI helped the expert come up with this idea.

Accuracy and coverage (H3): In terms of wrong results, Kari identified a wrong proposition extraction, out of 13 queries (D.1.4.2). This is within the expected error rate. In terms of coverage, when looking for information about Brazil's statements on forestry issues, it was necessary to play with several search terms and functions and use wildcards (e.g. `*forest* OR REDD` in the `FreeText` box) in order to get adequate coverage (D.1.2.12).

Strengths perceived by user: The expert appreciated the fact that the tool provides results at sentence and proposition level, dividing actors' utterances into components like *Actor*, *Predicate* and *Message* (D.1.3.3.a). She points out that this result-format allows you to find information about an actor faster than the *Climate Negotiations Browser* interface,[32] to which she was a contributor (Venturini et al., 2015), since in that tool you need to search again inside each hit for the query in order to find the sentences relevant for each actor. That tool was presented in the related work section of this chapter (6.3.3).

Weaknesses perceived by user: (a) Kari thought that annotating messages with DBpedia concepts is not informative enough for a domain-expert, who would find such concepts too general—she found the domain-specific Climate Thesaurus tags more informative (D.1.4.3). **(b)** Kari perceived coverage for Brazil's statements about forestry to be weak, which required her to reformulate the query with my help to get enough results (D.1.4.1).

I consider these perceived weaknesses to have easy workarounds: It would take little development to make DBpedia-concept display optional for a user, and the *Brazil/forestry* query can yield satisfactory results if expanded by the user with synonyms or wild-cards.

Other comments about the session: Kari regularly sorted results chronologically, to get evidence about the evolution of an actor's position. She also appreciated having access to the context for the sentences containing propositions (i.e. their location in the document), in order to understand them better (D.1.5.c).

6.6.4.2 Session 2 (Tommaso Venturini)

Overview of corpus (H1): Tommaso mentioned (D.2.2.14) that the level of detail he can get with the UI *(who said what, with access to the sentences documenting it)*, is useful and a good complement to the keyword co-occurrence methods his team has used on this corpus (Venturini et al., 2014). Nonetheless, he pointed out that more aggregations over the data would make the

[32] http://www.climatenegotiations.org/explore/#/?

tool better for corpus exploration. For instance, now it is possible to see ⟨*actor*, *predicate*, *message*⟩ propositions, getting frequency tables for keyphrases and entities extracted from the messages. However, Tommaso would like to see how many times a given actor used each predicate, and, for a given predicate, which are the actors that use it most. He argued that this would give a better overview of the corpus content (D.2.2.10).

Gain of insight (H2): Tommaso found it very interesting that the UI shows propositions for non-governmental groups, e.g. constituencies[33] like the Indigenous People's Organizations or Women and Gender (D.2.2.8). He did not expect to be able to find such groups, and they are not available on the interface his team developed (Venturini et al., 2015). Tommaso also suggested another way in which the information that feeds the UI can create new knowledge on the corpus: He said that, now that agreement and disagreement data between actors is available, he would like to use those data to create two country networks: the agreement network and the disagreement one, to see which countries behave like "allies" or "enemies" in the negotiations (D.2.2.12).

Accuracy and coverage (H3): The number of errors was within the expected range. Tommaso performed 20 queries and found three cases of potential errors (see D.2.2, queries 15, 8 and 12); he found it debatable whether the last two cases were errors or not. There were no comments about weak coverage. However, Tommaso noticed that complex predicates like *express concern* are now analyzed with *express* as the predicate and *concern* within the message (D.2.2.9). What is relevant to determine support and opposition patterns (one of the intended uses of the UI) is the actor's *concern* with whatever points are mentioned in the message, not the verb *express*. It would therefore be more informative to output a proposition where *concern* is the predicate. I was aware that the tool does not treat these complex predicates optimally, but what matters is that the expert pointed this out as something to improve on—I'm interested in weaknesses that the experts perceive as "getting in the way" for their work.

This is not a major limitation for two reasons. First, this affects less than 2% of propositions.[34] Second, as a workaround, propositions where the predicate is *express* can be filtered further by entering terms like `concern` in the `Points` search-box.

Strengths perceived by user: Other than positive spontaneous comments by Tommaso, mentioned above, he repeatedly expressed liking the interface,

[33] https://unfccc.int/files/parties_and_observers/ngo/application/pdf/constituency_2011_english.pdf

[34] There are 475 propositions (out of a total of 26,475 for the 0 to 5 confidence range), with *express* as the predicate and *concern*, *disappointment* or *alarm* in the message, i.e. 1.79% of the total.

but without providing details. Towards the end of the session, I asked him explicitly for reasons why he liked the interface. This is a departure from the evaluation protocol, which intended to assess the hypotheses not on the basis of explicit questions, but based on experts' comments on query results. I asked explicitly, and towards the end of the evaluation, in order to to better understand the value of the UI in the expert's perception.[35]

Tommaso's answer to this explicit question was that the search fields (`Actors`, `Actions`, `Messages`) model closely the types of information that a researcher wishing to study this corpus would like to search for, and it is helpful to be able to search by all of the criteria separately (D.2.3).

Weaknesses perceived by user: Tommaso stated missing more aggregation on the information extracted (D.2.2.10). Two concrete examples: If you extract 100 propositions where the actor is Canada, he would be interested in seeing how many times Canada has used what verb. Likewise, if we extract 100 propositions where the predicate is *complain*, he would like to know how many times each actor complained (D.2.3).

6.6.4.3 Session 3 (Nicole de Paula)

Overview of corpus (H1): I interpret the following statements by the expert as evidence that she finds the UI useful to get an overview of corpus actors and issues: **(a)** Nicole mentioned that she would use the UI to get an indication of which actors are most engaged with an issue (D.3.2.3). **(b)** She mentioned that, for people familiar with the corpus, it can help them recall the way the negotiation developed on a given Conference of the Parties (COP) (D.3.2.6).

Gain of insight (H2): Nicole said that having messages expressed by non-governmental groups like Indigenous Peoples or Women and Gender is valuable, since these groups usually push for interesting changes in climate-policy agenda, compared to countries (D.3.3.3.b).

Accuracy and coverage (H3): Nicole made no comments about inaccuracies or lack of coverage. She made some queries for search term *health*, observing that there were few results (e.g. for Brazil there were only two statements on health). However, as I later verified, if we perform a standard full-text-search against the whole corpus,[36] we will see that hits for health are mostly false positives, corresponding to sequences like *Minister of Health* as a speaker, not

[35] Along related lines, Khovanskaya et al., 2015, (p. 59) discuss cases where, in order to consider particularly relevant evidence for their research questions, they accepted evidence provided by participants during the debriefing conversation that followed the evaluation task

[36] This can be done with our standard Solr index for the corpus, see `http://apps.lattice.cnrs.fr/solr/enb12/browse?q=health&x=0&y=0&rows=100`

part of country statements on health. In this sense, for this example our tool provides more accuracy than a standard search, rather than less coverage.

Strengths perceived by user: Nicole stated preferring a tool like this, that returns results per author, to simply having to read all relevant ENB reports, for use cases where you want to see all information related to an author (D.3.3.1.c).

Weaknesses perceived by user: Nicole finds the tool of interest for a very specialized audience only (e.g. people who already carry out research on climate policy), as she feels that other people would not be interested in looking at the detailed behaviour of actors in the negotiation (D.3.4).

I do not see this as a weakness, since the tool was primarily intended for domain-experts.

Other comments about the session: Nicole stated finding the year-based search useful, in order to be able to restrict searches to time-periods she's interested in, and in order to get an idea of when issues matter (D.3.2.3). She also appreciated (D.3.2.4) having access to the context of a proposition (i.e. its sentence and its document).

6.6.4.4 Overall discussion of results

After feedback from each of the domain-experts has been examined above, a summary of the results is provided below.

Regarding whether the tool helps an expert obtain an **overview of the corpus (H1)**, Kari and Nicole's statements indicate that they find the tool helpful to follow actors and issues in the corpus, statically or across time. Tommaso however expressed that more aggregations over the proposition data would provide a better overview: e.g. not only extracting all propositions for an actor like *Canada* and showing counts for keyphrases and entities in its messages, but also displaying tables showing how many times the actor used each predicate.

In terms of **potential gain of insight (H2)** by experts using the tool, there were several examples suggesting this potential.

- Kari thought of a new topic for research by looking at the systematic articulation of corpus content into propositions, namely, that actors could be compared in terms of whether their messages are preferentially about formal/legal aspects, or about introducing policy items to address climate change issues.
- Tommaso was positively surprised to find propositions for non-governmental actors like Women and Gender or Indigenous Peoples' Organizations. I consider that these extractions can be seen as new knowledge

generated on the corpus, since these actors had not been considered in prior work on the corpus (Salway et al., 2014; Venturini et al., 2015). The particular interest of extracting messages by such actors was also pointed out by Nicole.

- Tommaso also stated that, looking at these data, he would now like to create two networks, one for disagreeing actors, and one for agreeing ones. Creating these networks is possible since the tool generates agreement/disagreement data, which was not previously available to him. Note however that Salway et al. (2014) extracted pairs of opposing and agreeing actors, but without identifying the objects of (dis)agreement, which we did indicate.

As regards **accuracy and coverage (H3)**, the amount of errors was well within the expected rate (F1 for proposition extraction was 0.69, see p. 88), and this error rate was not judged by experts as detrimental to their work. Some examples showed that having the option to search a query-term in the message emitted by an actor can yield more relevant results for the query than searching the term in the complete document (see p. 192 for an example related to search-term *health*).

Coverage was satisfactory. In cases where the expert suspected lack of coverage, it was possible to retrieve a larger result-set with wildcards or enriching the query with synonyms.

One expert noticed the imperfect treatment of complex predicates like *express concern*. Now, *express* is tagged as the predicate. However, tagging *concern* as the predicate would capture the actor's position better. A workaround was shown to the expert to deal with this, even if it would be useful to improve this as future work.

An **overall perceived strength** of the tool was that the ability to search for proposition elements makes it easier to find information related to an actor or issue than a search that returns unanalyzed document fragments. Regarding **perceived weaknesses**, an expert missed aggregations for predicates by actors and vice versa. **Other useful elements** of the tool were perceived to be the fact that actors are extracted dynamically even if they are not part of a predefined actor-list, and having access to the context for the propositions extracted (their sentence and document).

In **summary**, the domain-expert evaluation provided evidence that the interface can help experts gain an overview of actors and issues in the corpus. I consider perceived weaknesses regarding data aggregation a data-presentation issue that can be fixed with additional development, rather than a core flaw that would detract from the interest of the approach. Some of the queries performed by the experts resulted in new insight on the corpus

for the experts. In spite of the errors that need to be expected from any automatic linguistic analysis, the quality of the extractions was judged sufficient by the experts. Perceived lack of coverage for some queries was shown not to be such, or there were workarounds in order to retrieve sufficient results. For these reasons, I consider that the interface fulfills its goals satisfactorily.

6.7 Summary and Outlook

A detailed chapter summary is presented in 6.7.1. Future work possibilities were already mentioned in the chapter, and are recapped in 6.7.2.

6.7.1 Summary

The chapter presented an application[37] to navigate a subset the *Earth Negotiations Bulletin*,[38] covering negotiation reports for *Conference of the Parties* climate policy summits. As this is a corpus of international political negotiations, it is important to know not only what issues are being discussed, but also which participants are addressing which issue, and what their attitude towards the issue is, whether it is something they favour or something they are against. This type of information cannot be provided by the term cooccurrence methods we applied for analyzing the Bentham and PoliInformatics corpora in the other application cases in this thesis. For this type of analysis, we need to identify the reporting predicates relating an actor and the issues that the actor speaks about, in order to get an indication of the actor's attitude towards those issues.

To extract this type of relational information, the **proposition extraction** system which had been described in Chapter 4 (p. 78ff) was applied to the corpus. This annotates *propositions*, i.e. triples consisting of an actor, the messages it emits, and the reporting predicates (verbal or nominal) mediating between the actor and its messages.

Predicates can be of three types: support, opposition or neutral reporting. Different metadata were extracted from the propositions' messages: Keyphrases, DBpedia concepts, and terms from the Climate Thesaurus (a domain-specific thesaurus relevant for climate policy). The metadata are exploited to provide an overview of issues in propositions where actors use predicates of support, opposition, or general reporting.

A **user interface** displays all of the information above. Users can search for propositions emitted by a given actor, via a given predicate or predicate-type, and containing specific query terms in their message. The metadata are displayed next to propositions matching a query, as an overview of the

[37] http://apps.lattice.cnrs.fr/ie/uidev/
[38] http://enb.iisd.org/enb/vol12/

content of the propositions' messages. This can help answer questions like the following:

- What issues are mentioned in propositions where an actor (e.g. *China*) is using a predicate of opposition?
- What actors are mentioning a given issue (e.g. *human rights*) and using what verbs?
- What other keyphrases or thesaurus concepts appear in the context of the query terms, within actors' messages?

The interface also has a tab where two actors can be chosen, and issues in messages over which the actors agree or disagree are displayed, with access to the sentence where this agreement or disagreement was attested.

Figures 6.4, 6.5 and 6.6 show how the UI helps obtain information of the types just mentioned.

The interface's level of detail in analyzing actors' statements, including the predicates and issues related to them was not available in previous work on the corpus, which was reviewed in 6.3.

The **application's goals** were giving an overview of the corpus, helping to answer questions like the ones listed above, and ideally helping a domain-expert gain new insight on the material. For instance, by providing them with evidence they were not aware of, or by helping them develop ideas for research that they may not have thought of before.

A **domain-expert evaluation** was performed to assess the extent to which the application reaches the goals just mentioned. The interface was evaluated by three domain experts who have previously carried out research on this corpus.

In terms of obtaining a **corpus overview**, the experts appreciated the navigation possibilities provided by a differentiated search for each proposition element, and how the UI helps follow a given actor's behaviour in the corpus. However, for a more global overview, one of the experts pointed out that new aggregations of the results extracted would be helpful. E.g. besides displaying propositions where *Canada* is the actor, displaying how many times the actor uses each predicate. Or given a certain predicate (e.g. *rejected*), displaying counts for how many times each actor uses it (for *rejected*, this would give an indication of which countries are more confrontational). The data to create those aggregations could be gathered manually on the interface now, but this is time-consuming and as future work it would be an improvement to provide such aggregations.

As regards **potential for new insight**, two experts mentioned that an interesting result is how the interface shows propositions for less commonly

studied actors that do not correspond to a country or country group, e.g. the Indigenous People's Organizations or the Women and Gender group. The previous studies we reviewed for this corpus were not analyzing material on these actors. Other examples suggesting that new insight can be gained from the type of corpus representation provided on the interface are discussed on pp. 189 and 191 (under "Gain of insight").

Proposition extraction, which is the basis of the navigation workflows on the interface, had been evaluated intrinsically for exact match against a reference set, with an F1 score of 0.69. The quality of proposition extraction results was considered sufficient by the experts who tested the interface.

6.7.2 Outlook

Several improvements could be made as future work. Regarding proposition extraction, complex predicates like *express alarm* are now being analyzed with *express* as the predicate, whereas it would be more informative to identify the noun following the verb (e.g. *alarm*, *concern*) as the predicate.

In terms of the interface, a function to export the results of a query would be useful for users to post-process the results. It would also help them create their own aggregations of results beyond what the UI offers. Besides, as just mentioned above, additional result aggregations on the interface would also be helpful for users to get an overview of a result-set. The agreement-disagreement view could be improved by adding a search on it—now the content about which actors agree or disagree can be browsed, but not searched. Finally, search functions and highlighting could be improved in some cases, as was discussed on p. 176 and p. 177 (footnote 22).

The application presented in the chapter intended to be an example how Natural Language Processing analyses (involving in this case syntactic dependencies and semantic roles) help obtain information that complements cooccurrence-based methods, providing more structured results that would be unavailable with cooccurrence information only. Experts' feedback on the application suggested that it is relevant for their research needs on the corpus, and cases of new insight on the corpus derived from using the application were attested.

Conclusion

The thesis examined how several Natural Language Processing (NLP) technologies can help access relevant information in large textual corpora. Two technologies called Entity Linking (EL) and keyphrase extraction were applied in order to annotate actors and concepts in the corpora. Relation extraction methods were employed to determine how those actors and concepts are related to each other. The NLP annotations were integrated in corpus navigation applications, which combine full-text search, networks, and structured search based on the annotations. As the quality of NLP results varies according to the corpus, it was necessary to perform some development in order to adapt the tools to the corpora of application.

This conclusion discusses the following aspects: First, some generalizations on the domain-expert feedback for the application cases in the thesis are presented, along with a recap of possible future work to improve the applications. We start the discussion with this topic since it is the potential usefulness for an expert that justifies the work carried out, and conclusions based on these case studies are relevant for similar work on other corpora. Second, we turn to some remarks on the generic vs. corpus-specific character of the basic NLP technology applied or developed for the thesis' case-studies. Finally, as a large part of the thesis focused on applications, some "lessons-learned" comments are provided about implementation choices that have a beneficial impact in developing the type of work produced in the thesis.

Domain-expert Evaluation: Reproducing Available Knowledge and Gain of Insight

Case studies for three different corpora were carried out. Corpus navigation interfaces integrating NLP annotations were created. It was evaluated whether the information presented on these interfaces, and the navigation workflows they afford, provide a useful overview of the corpus according to domain-experts, and also, whether experts arrive at new insight on the corpus by using the applications. In following, the outcomes of domain-expert evaluation are outlined. Strengths as well as shortcomings of the applications are summarized, besides possible ways to improve them as future work.

Application 1: Bentham's Manuscripts

The corpus for the first case study (section 5.2) was the manuscripts of Jeremy Bentham. This is an 18th–19th century corpus in political philosophy, ethics and related topics. 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 towards DBpedia did not give satisfactory results, as the DBpedia concepts chosen by the Entity Linking (EL) tool were often anachronistic, i.e. modern terms not applicable to Bentham's writings. It was still possible to use EL outputs to annotate important terms in the corpus. However, rather than using the DBpedia concept labels assigned by EL to corpus mentions, labels chosen among the corpus mentions themselves were used (see p. 110).

As regards the application's usefulness for a corpus overview, and insight-gain potential for a domain-expert, two main results emerged from the domain-expert evaluation.

The corpus overview provided by the networks corresponds to the expert's knowledge of the corpus; 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, like *sinister interest* or *operative power*. The DBpedia term-mentions were found to represent basic elements of meaning (e.g. *interest* or *power*) underlying those characteristic Bentham notions, but not the precise expressions used by Bentham.

A way in which the expert saw the application as having potential to help gain new knowledge about the corpus was the following: The expert found keyphrase networks useful for finding alternative phrasings for a term. For instance, a core concept in Bentham's writings is *sinister interest*. The networks showed near-synonyms for this term: *self-regarding interest*, *interest of the subject* and *private interest*. 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. A distributional semantic model that could be used is word2vec (Mikolov et al., 2013). An implementation that allows for multi-token queries would be required since keyphrases (generally multi-token) were found to be more useful for specialized research on the corpus than the DBpedia concept mentions (largely single-token). Current versions of the *gensim* library (Řehůřek et al., 2010) could be used.

Application 2: PoliInformatics

The corpus for the second case study (section 5.3) was a subset of the PoliInformatics corpus, which contains materials about the American financial crisis of 2007–2008. A corpus navigation application was created, which exploits several Entity Linking tools to annotate actors and concepts in the corpus.

Social science researchers we have worked with had expressed a need for measures that would allow them to evaluate Entity Linking outputs, in order to select the ones that are more likely to represent important notions in the corpus and remove errors. Our application offers two measures that intend to fulfill that purpose. First, a confidence measure giving an estimate of how likely the annotation is correct. Second, a corpus-level coherence measure, indicating the extent to which a concept is thematically consistent with other concepts annotated in the corpus overall; the notion of coherence relies on common inlinks between concepts in the Wikipedia link graph. Besides the measures to aid manual selection, an automatic selection of annotations is also provided, based on the weighted voting method to combine Entity Linking outputs from Chapter 3.

No domain-expert evaluation was performed for the PoliInformatics application. However, the terms and result-quality measures available on the interface were inspected in order to assess whether they could help an expert obtain a corpus overview or select relevant terms to analyze the corpus with. A network for organizations mentioned in the corpus was also examined, to assess the corpus overview it provides. We summarize here the outcome of those exercises, as it speaks to the potential usefulness of the application for domain-experts.

Examples were shown where the measures of annotation quality and the automatic annotation selection help to tease apart correct Entity Linking outputs from incorrect ones. These examples were mostly for annotation types *Organization* and *Concept*. However, the methods developed have limitations. For annotations of type *Person* and *Location*, the methods did

not give adequate results. As future work, it would be relevant to create measures of annotation quality appropriate for all annotation types.

Another limitation of the application is that corpus actors are not annotated unless they are mentioned in the Wikipedia/DBpedia knowledge base (KB). Whereas the coverage of corpus organizations in the KB was acceptable, several people who are important in the corpus are absent from the KB. Possible future work to annotate person names independently of a KB was outlined (p. 158): Clustering coreferential person-names, as in Coll Ardanuy et al., (2016a; 2016b), or building a domain-specific KB and, as in Frontini et al. (2015), disambiguating against the generic and specific KBs simultaneously.

Application 3: The *Earth Negotiations Bulletin*

The final case study in the thesis (Chapter 6) was an application to navigate volume 12 of the *Earth Negotiations Bulletin* (ENB). This consists in reports on international climate policy conferences. Relation extraction methods based on semantic role labeling and syntactic dependency parsing were applied in order to automatically annotate propositions, i.e. triples for actors who make a statement in the negotiations, their message, and the reporting verb or noun relating the actor and its message. A user interface allows searching the corpus based on proposition elements. For instance, actors who have used verbs of opposition can be retrieved, along with the sentences where this was attested. Keyphrases and concepts from DBpedia and from a domain thesaurus are annotated in messages emitted by the actors, to provide an overview of issues addressed by each actor. The interface also displays which actors agreed or disagreed with which, and regarding what issues. An interesting point about this application is that the information it provides cannot be obtained with simple cooccurrence between actors and concepts; it requires identifying the relation between them. In this sense, the analysis is deeper than the methods used for the other case studies in the thesis.

A qualitative evaluation performed with three domain-experts showed that experts were able to obtain new insights on the corpus thanks to the application. More specifically, experts appreciated seeing statements by actors that are not commonly studied, and that had not been covered by prior work on the corpus (p. 166), e.g. indigenous peoples' organizations. Prior applications had covered a predefined list of countries and country groups. By contrast, in our pipeline, besides countries and their groups, any other actor that is the agent of a reporting verb is susceptible of being annotated as a speaker. One expert reported that seeing a detailed analysis of corpus sentences articulated as $\langle$*actor, predicate, message*$\rangle$ triples is useful to find dimensions along which to compare countries. She suggested that one such

dimension, that she had not previously thought of, is the extent to which actors' statements refer to formal and legal aspects of the negotiations or to actual climate change management measures. Finally, one of the experts suggested that the agreement/disagreement information on the interface be used to create an agreement and disagreement network with actors as nodes. One of the objectives of the application was precisely extracting information from a corpus that could be used to create corpus networks where the information encoded by network edges is more precise than co-occurrence. That expert's comment suggests the usefulness of the data provided by the application in this respect.

As regards useful improvements suggested by the domain-expert evaluation, an expert mentioned that it would be desirable to provide additional aggregations for the triples extracted. For instance, providing counts for how many times each actor uses each predicate and predicate type (i.e. support/opposition/neutral) and how many times each predicate is used by the different actors.

Another type of interesting future work could be to publish as linked data the corpus annotations, i.e. the ⟨*actor, predicate, message*⟩ triples, plus the keyphrases, DBpedia concepts and domain-specific thesaurus concepts. Publishing the complete corpus text as linked data may also be possible depending on its editors' interest to do so.

Generic and Corpus-specific NLP Developments

As was mentioned in the thesis, the efficacy of NLP tools varies according to the corpus. After having summarized the results of the domain-expert evaluations, and points to improve based on them, another aspect to address in the conclusion is to what an extent the solutions we developed to analyze the corpora in the thesis would generalize to new corpora.

In Chapter 3 a method was presented to combine the results of several Entity Linking/Wikification systems, in order to obtain combined results that outperform each individual tool's results. The combined results improved over individual systems on four publicly available reference-sets. However, it is open to question how well the method can generalize to other corpora, for several reasons. First, one of the most effective systems we combined was a publicly hosted web-service. This service was discontinued, and a local deployment takes significant effort. Other appropriate tools would need to be tested to attempt to reproduce the improvements shown with the original set of tools we combined. Second, the weighted-vote method for system combination relied on the following assumptions (p. 63): A user can select to annotate Knowledge-Base terms that are expressed by named

entities only, or both KB-terms expressed by named entities and those expressed by common-noun phrases. In the first case, the combined results should improve if systems performing best at corpora where named-entity mentions predominate are given more weight in the vote. In the second case, the combined results should improve if systems which perform well on corpora rich in common-noun annotations bear more weight. These assumptions are not unreasonable and the method worked as expected on several test-sets. However, the weights assigned to systems in the vote are based on the systems' performance on a small set of external reference corpora, and the weighted vote is not otherwise adapted to characteristics of the corpora to be annotated. Accordingly, it is open to question whether the method can generalize well. As future work to develop a system combination method, the machine learning literature discusses approaches like stacked generalization, where a higher-level classifier is created, which applies to the outputs of several systems with the goal of improving over those systems' individual predictions.

Besides a method to combine Entity Linking results, the other basic technology developed is a system to extract ⟨*actor, predicate, message*⟩ triples, which we called *propositions*. The system was presented in Chapter 4, and it was developed for exploiting it in the application in Chapter 6, to navigate volume 12 of the *Earth Negotiations Bulletin* (ENB) in a structured manner, according to these actors, predicates and messages. The proposition extraction system employs an NLP pipeline which provides semantic role labeling and syntactic dependencies among other information. The system applies rules on the NLP output to identify propositions. Some of the rules rely on very generic semantic roles (*agent* or *theme*) or syntactic functions (*subject*) and can generalize to other corpora; one of the reasons for using syntactic and semantic structure instead of simpler part-of-speech and lexical patterns was that by using structure, we can create more generic rules, that abstract away from variability in word-order. However, some other rules in the system were created to deal with specific aspects of the corpus and are not expected to be applicable to new corpora. Also, note that the ENB corpus shows limited syntactic variety, since ENB editors seek neutrality and want to avoid variations in style that may present certain negotiation actors more prominently than others. The system we developed is intended for extracting reporting events from texts with limited stylistic variety; a possible text type with those characteristics, besides other volumes of the *Earth Negotiations Bulletin*, could be parliamentary proceedings.

For a generic analysis of reporting events, we mentioned relevant systems in our literature review. E.g. *rsyntax* (Van Atteveldt et al., 2017), for news corpora, or PETRARCH (Schrodt et al., 2014), which annotates speech events

and others. The difference in the workflow we developed is that it deals with corpus-specific aspects that would not be treated optimally by pre-existing systems, e.g. addressing both verbal and nominal reporting predicates, and the detailed structure of some of the sentences typical for the ENB corpus. Besides the basic technology for proposition extraction, part of the interest of the work in this thesis comes from the corpus navigation interface integrating the proposition extraction results, which was shown to have suggested new research ideas to domain-experts, as mentioned on p. 202 and elaborated on in Chapter 6 (p. 193ff).

Lessons Learned regarding Implementation

Finally, in this conclusion I would like to recap on some implementation issues that it's beneficial to consider in order to create applications like the ones developed in the thesis; this is relevant as a large part of the thesis had an applied focus.

Some of the NLP tools applied were publicly hosted services. This is not ideal because if the service is discontinued, it may not be possible to reproduce its results. Besides, this does not allow a complete control over the tools' configuration, which could be changed by the team hosting the tools. It is a better choice to use tools that can be deployed locally with reasonable effort.

The domain-experts who gave feedback on the applications asked for a feature to export the results of all queries and data manipulation they performed on the interfaces. This was not implemented for time-reasons and would be relevant future work.

The Bentham navigation application (Ch. 5) used concept networks with some interactive features. For these interactive networks, a useful feature to implement as future work would be the following: The corpus contexts for terms represented in a network node could be accessible directly by clicking on the node. Similarly, clicking on a network edge could give access to the corpus elements where the information encoded in the edge is attested. These corpus elements would be the window within which nodes co-occur in co-occurrence networks, or the sentence where a relation between actors is expressed, or where an ⟨*actor, predicate, message*⟩ proposition was attested, if a network that exploits such information were to be created. Access to an annotation's context is possible in the current applications, but not from the networks directly: In the Bentham application, corpus contexts for a network-node are accessible from a search index rather than from the networks. In the ENB application, which does not provide networks but which outputs the information needed to create them, the sentences and document contexts in which a proposition was annotated are accessible from

the results panes. In both cases, experts appreciated having access to the corpus contexts. Accordingly, making the contexts directly available from the networks would be a useful enhancement.

Other possible implementation improvements, that were already mentioned when discussing the application cases in chapters 5 and 6, would be as follows:

Some of the information displayed on the interfaces (like actors or concepts) corresponds to a label that represents several corpus variants. At the moment, the variants are listed in external materials (for ENB and Bentham), or only available in the backend database (for PoliInformatics). Exposing this information directly on the interface would be useful.

Similarly, highlighting terms (and their variants) in their context of occurrence in the corpus is only partially implemented, and it would be useful to perfect that.

Final Remarks

In this conclusion, domain-experts' feedback was discussed first, as what justifies the work carried out in the thesis is its potential usefulness for researchers in a social sciences or humanities field. Two of the three applications were evaluated with domain-experts. In the Bentham corpus application, the domain-expert found that the corpus representations produced correspond to his knowledge of the corpus, but they did not generate new research ideas. In the *Earth Negotiations Bulletin* application, the systematic annotation, via relation extraction methods, of speakers and their messages, as well as agreement and disagreement between speakers, did result in experts finding new research ideas by using the application.

Generic NLP tools were applied, with additional developments to better capture corpus-specific characteristics. The limitations of annotating corpus concepts with general-domain knowledge bases were discussed. These limitations make it helpful to complement such analyses with domain-specific knowledge bases or data-driven methods like keyphrase extraction and distributional similarity models.

Implementation choices that have a positive impact on the usefulness of applications were mentioned: Relying on tools that can be deployed locally rather than on publicly hosted services, and providing access to the corpus context for annotations, possibly directly from interactive corpus networks if these have been produced.

We were fortunate to work with a variety of corpora about diverse topics, and speak to researchers from the related domains, who shared their feedback

about the corpus navigation applications we created, and on how applying Natural Language Processing technologies can contribute to their research on large textual corpora.

Appendix A

Term Lists for Concept-based Navigation

The term lists for the concept-based navigation applications are below. For the Bentham corpus (see Chapter 5, 5.2), Entity-Linking based lists are in A.1.1, and lists based on keyphrase extraction are in A.1.3.

For PoliInformatics (see Chapter 5, 5.3), A.2.1 shows a list for the DBpedia concepts used to create the corpus network in Figure 5.18, as well as their corpus mentions.

A.1 Bentham corpus 210
A.1.1 Entity Linking on the Bentham corpus 210
A.1.2 Entity Linking on Bentham: Deleted terms 212
A.1.3 Keyphrase Extraction on the Bentham corpus 213
A.2 PoliInformatics 216
A.2.1 Entity Linking on PoliInformatics 216
A.2.2 Co-occurrence table for PoliInformatics corpus network 220

A.1 Bentham corpus

A.1.1 Entity Linking on the Bentham corpus

For Entity Linking, the most frequent mention for each concept was taken as the concept label, rather than the DBpedia label (see Chapter 5, 5.2.3). The label is looked up in the text in addition to any variant and lookup is case-insensitive.

Minimum frequency to keep a variant was 100 and minimum annotation confidence was 0.1. These thresholds, plus manual deletion of some very general terms (p. 212) yielded a list of 257 terms. CorText was asked to create two networks: One with all of these terms, and another one with the 150 most frequent ones in the list.

In network creation, CorText filters weakly connected terms, so that the actual number of nodes in the networks was 233 and 141. Nodes filtered out by CorText are **greyed out** on the table. Terms with an asterisk **(*)** are part of both the 141 and 233-node networks (unless greyed out as filtered out by CorText).

#	Label	Variants	#	Label	Variants	#	Label	Variants
1	abuse	abuse	21	capital	capital	41	death	death
2	action	action	22*	case	case, cases	42*	decision	decision
3*	Acts	acts	23*	class	class	43*	defence	defence
4*	addition	addition	24*	Code	code	44*	defendant	defendant
5*	aggregate	aggregate	25	Codification	codification	45*	degree	degree
6*	Appeal	appeals, appeal	26	Common Law	common law	46*	demand	demand
7*	application	application	27*	community	community	47	democracy	democracy
8*	applied	applied	28	conception	conception	48	design	design
9*	aptitude	aptitude	29*	consideration	consideration	49	despotism	despotism
10*	argument	arguments, argument	30*	Constitution	constitutional, constitution	50	dignity	dignity
11*	art	art	31	contract	contract	51	discourse	discourse
12*	article	article	32	Corinthians	cor	52	Doctrine	doctrine
13	attention	attention	33*	corruption	corruption	53	dominion	dominion
14*	authority	authorities, authority	34	Cortes	cortes	54	duty	duty
15	belief	belief	35*	country	country, countries	55	Economy	economy
16*	benefit	benefit	36*	Court	court, courts	56*	Election	election
17*	Bentham	bentham	37	Court of Session	court of session	57	Elector	elector
18*	Bill	bill	38	crime	criminal, crime	58*	employed	employed
19*	body	body	39	Crown	crown	59*	England	england
20	bribery	bribery, bribe	40	damage	damage	60*	English	english

[continues on next page]

[continues from previous page]

#	Label	Variants
61	English law	english law
62	entities	entities, entity
63	Equity	equity
64*	evidence	proof, evidence
65	evil	evils, bad, evil
66	execution	execution
67*	exercised	exercise, exercised
68*	existence	existence, existing
69*	expected	expected, expectation
70*	experience	experience
71*	fact	fact
72	faculty	faculty
73	faith	faith
74	Fallacies	fallacies
75	fear	fear
76	fide	fide
77*	field	field
78*	force	force
79	foreign	foreign
80	fraud	fraud
81	free	free
82	function	function
83*	God	god
84*	good	good
85	goods	goods
86*	government	government
87*	hands	hands, hand
88*	happiness	happiness
89*	hope	hope
90	House of Commons	house of commons
91	House of Lords	house of lords
92*	human	human
93	idea	idea
94	income	income
95*	individual	individuals, individual
96	injury	injury
97	injustice	injustice
98	instrument	instrument
99	intellectual	intellectual
100*	interest	interest
101*	Jesus	jesus
102	John	john
103*	Judge	judge
104*	judgment	judgment
105*	judicature	judicatory, judicial, judicature
106	jurisdiction	jurisdiction
107	jurisprudential	jurisprudential
108*	Jury	juries, jury
109*	justice	justice
110*	King	monarch, king, monarchy
111	knowledge	knowledge
112*	labour	labour
113*	language	language
114*	law	law, legal, laws
115*	lawsuit	litigation, suit
116*	lawyers	lawyers, lawyer
117*	learned	learned
118	legislation	legislation
119	legislator	legislator
120	length	length
121*	Letter	letter
122	liable	liable
123	liberty	liberty
124*	life	life
125*	Logic	false, logic
126*	Lordship	lordship, lord
127	love	love
128	Luke	luke
129	majority	majority
130*	man	men, man
131	Mark	mark
132*	mass	mass
133	matter	matter
134*	measure	measure
135*	Member	member
136*	mind	mind, minds
137	Minister	minister
138	miracle	miracles, miracle
139*	money	pecuniary, money
140*	moral	moral
141	motion	motion
142	nation	nation, nations
143*	number	number
144*	object	object
145	obligation	obligation
146*	observation	observed, observations, observation
147*	office	offices, office
148*	official	official
149*	opinion	opinions, opinion
150	opposition	opposition
151*	ordinary	ordinary
152	pain	pain
153*	paper	paper
154	parliamentary	parliamentary
155*	Parliament	parliament
156	parties	parties
157*	party	party
158	patronage	patronage
159*	Paul	paul
160*	people	people
161	performed	performed
162*	personal	personal
163	person	person, persons
164	persuasion	persuasion
165	Peter	peter

[continues on next page]

[continues from previous page]

#	Label	Variants
166	plaintiff	plaintiff
167*	plan	plan
168*	pleasure	pleasure
169*	point	point
170*	political	political
171	population	population
172*	possession	possession
173*	power	power
174*	powers	powers
175	practical	practical
176*	practice	practice
177	prejudice	prejudice
178*	price	prices, price
179*	principle	principle
180	private	private
181*	probability	probable, probability
182*	procedure	procedure
183	productive	productive
184*	profit	profit
185	property	property, properties
186*	proportion	proportion
187*	proposition	proposition, propositions
188*	public	public
189*	punishment	punishment
190*	purpose	purpose
191	quality	quality
192*	quantity	amount, quantity
193*	question	question
198	Reform Bill	reform bill
199*	Reform	reform
200	regulations	regulations
201*	relation	relation
202*	religion	religion
203*	remedy	remedy
204*	representatives	legislative, representatives
205	reputation	reputation
206	rights	rights
207*	rulers	rulers
208*	rule	rule
209*	sacrifice	sacrifice
210	science	science
211*	seat	seat
212*	security	security
213	service	service
214*	share	share
215	silent	silent
216*	sinister	sinister
217	sin	sin
218	Sir	sir
219	social	social
220	society	society
221*	source	source
222	space	space
223*	Spain	spain
224	Spanish America	spanish america
225*	Spanish	spanish
230*	suffering	suffering, suffer
231	suffrage	suffrage
232	supreme	supreme
233*	tax	tax, taxes
234	testimony	testimony
235	theory	theory
236*	thought	thought
237	title	title
238	trade	trade
239	Tripoli	tripoli
240	trust	trust
241*	truth	truth
242*	understanding	understood, understand, understanding
243	universal suffrage	universal suffrage
244*	universal	universal
245	utility	utility
246*	view	view
247	virtue	virtue
248*	vote	vote, voting
249	war	war
250*	wealth	wealth
251	Whigs	whigs
252	wisdom	wisdom
253*	witness	witnesses, witness
254*	word	word
255*	worth	worth
256*	writing	written, writing
257*	year	year, years

A.1.2 Entity Linking on Bentham: Deleted terms

The set of terms passing the confidence and frequency thresholds was larger than 257. However, the following terms were deleted manually (the label is shown; in most cases, the variant-set was equal to the label):

account, *business*, *Ch* (i.e. an abbreviation for *chapter*), *character*, *import*, *instance*, *left*, *manner*, *mode*, *nature*, *note*, *place*, *respect*, *sense*, *set*, *shape*, *side*, *sort*, *system*, *speaking*, *term*, *time*, *times*.

The reason for deletion was that the terms are part of complex prepositions like *in respect of*, or otherwise too vague to contribute to the core meaning of the corpus, like *place* or *time*.

A.1.3 Keyphrase Extraction on the Bentham corpus

The lists document our keyphrase extraction results. The keyphrases were used to create network maps for the corpus. Keyphrases in the list have a minimum corpus frequency of 10. Ill-formed or irrelevant phrases had been previously filtered (see p. 112).

Maps of 133 and 240 nodes were created with the keyphrases. For the 133-node one, the most frequent ones from the 240 keyphrase list below were used. Items marked with an asterisk **(*)** are part of both maps.

Corpus lookup in order to create networks was case-insensitive.

#	Keyphrase	#	Keyphrase	#	Keyphrase	#	Keyphrase
1*	absolute monarchy	**16**	body of men	**31***	constituted authorities	**46**	distant dependencies
2*	act of parliament	**17***	body of the law	**32***	constitution	**47***	division of power
3	adam smith	**18***	body of the people	**33***	constitutional law	**48***	doctrine
4*	aggregate mass	**19**	breach of trust	**34**	constitutive power	**49***	efficient cause
5*	american united states	**20**	british constitution	**35**	corrupt dependence	**50**	efficient causes
6*	anglo-american united states	**21***	business of government	**36***	corruption	**51***	election
7*	annuity	**22***	case	**37***	corruptive influence	**52**	election district
8	appropriate active talent	**23**	case admitts	**38***	country	**53**	election districts
9*	appropriate aptitude	**24***	cause	**39***	court	**54***	elector
10*	appropriate moral aptitude	**25***	circumstantial evidence	**40**	court of justice	**55**	encrease of wealth
11*	arbitrary power	**26**	civil war	**41***	defendant	**56**	end of government
12	author of the acts	**27**	common interest	**42***	degree	**57**	end of justice
13*	bank paper	**28***	common sense	**43***	delay vexation	**58**	ends of judicature
14*	bentham esq	**29**	commons house	**44***	difficulty	**59**	english constitution
15*	bill	**30***	community	**45**	direct evidence	**60**	english jurisprudence

[continues on next page]

[continues from previous page]

#	Keyphrase	#	Keyphrase	#	Keyphrase	#	Keyphrase
61	english lawyers	**86***	greater number	**111**	judicial injustice	**136***	mixt monarchy
62	english practice	**87***	greatest happiness	**112***	judicial procedure	**137**	mode of procedure
63*	evidence	**88***	greatest number	**113***	jurisprudential law	**138***	mode of voting
64*	evil	**89**	hands of the judge	**114***	justice	**139***	monarch
65	exchequer bills	**90***	holy ghost	**115***	language	**140***	money
66	external instruments of felicity	**91***	house of commons	**116***	law	**141**	moral sanction
67	factitious causes	**92**	human beings	**117**	legislative power	**142**	national wealth
68*	factitious delay	**93**	human breast	**118***	legislator	**143**	natural causes
69*	factitious honor	**94**	human happiness	**119**	limited monarchy	**144***	natural procedure
70	factitious reward	**95***	human mind	**120**	line of conduct	**145**	natural system
71*	failure of justice	**96***	individual	**121***	litigation	**146***	nature of man
72	female sex	**97***	individual case	**122***	logic	**147***	nature of the case
73*	fictitious entities	**98**	individual instance	**123**	lord chancellor	**148***	new south wales
74	fictitious entity	**99**	individual occasion	**124***	lord president	**149**	non agenda
75*	fide appeals	**100***	influence	**125**	lords delegates	**150***	number
76	fide defendant	**101**	influence of understanding	**126***	majesty	**151**	number of individuals
77	fide suitor	**102***	influence of will	**127**	majority of the people	**152**	number of the members
78*	field of law	**103**	inner house	**128***	man	**153**	number of the persons
79*	field of legislation	**104**	instruments of felicity	**129**	marginal insertion	**154**	official establishment
80	forms of government	**105***	interest of the people	**130**	mass of money	**155***	open mode
81*	forthcomingness	**106**	interest of the ruling	**131***	matter of corruption	**156**	operative power
82	freedom of suffrage	**107**	interest of the subject	**132***	members	**157**	original draught
83	general interest	**108***	jeremy bentham	**133**	men of law	**158***	paul
84*	general rule	**109***	judge	**134**	military force	**159***	penitentiary house
85*	great britain	**110***	judicial establishment	**135***	mischief	**160***	person

[continues on next page]

[continues from previous page]

#	Keyphrase	#	Keyphrase	#	Keyphrase	#	Keyphrase
161*	plaintiff	**181**	public functionary	**201***	religion of jesus	**221**	standard of rectitude
162*	pleasure	**182***	public interest	**202**	review chamber	**222**	state of dependence
163	point of fact	**183**	public mind	**203***	rise of prices	**223***	statute law
164	political community	**184***	public opinion	**204***	rule of action	**224***	statutory law
165*	political power	**185***	public opinion tribunal	**205**	scotch law	**225**	substantive branch of the law
166	political state	**186**	public spirit	**206**	secrecy of suffrage	**226***	sum of money
167*	population	**187**	pure monarchy	**207***	secret mode	**227***	supreme operative
168*	power	**188***	quantity of money	**208***	securities	**228***	supreme operative power
169*	powers of government	**189**	quantity of time	**209***	security	**229**	system of pleading
170	presence of the judge	**190**	question of fact	**210***	self-regarding interest	**230***	system of procedure
171	principal fact	**191***	question of law	**211**	separate interest	**231***	theory
172*	principle of utility	**192***	radical reform	**212**	side of the cause	**232**	tothill fields
173	private interest	**193***	rate of interest	**213**	single hand	**233**	ultramarian provinces
174	probative force	**194**	real evidence	**214**	single individual	**234**	universality
175*	profit	**195**	real law	**215**	single person	**235**	unwritten law
176	proper end	**196**	real wealth	**216**	single word	**236**	vast majority
177	proper end of government	**197***	reform	**217**	sinister end	**237***	vices
178*	proportion	**198***	reign	**218***	sinister influence	**238***	westminster hall
179	public discussion	**199***	relation	**219**	social affection	**239***	work
180	public functionaries	**200***	religion	**220**	species of evidence	**240**	written evidence

A.2 PoliInformatics

This section documents the term lists used to create the PoliInformatics corpus network in Chapter 5. The corpus itself was introduced in Chapter 5 (p. 135).

The DBpedia concept labels for the nodes in that network (Figure 5.18), as well as the corpus mentions for those concepts, are shown in A.2.1.

Co-occurrences between concept-mentions within the same sentence are shown in A.2.2. These co-occurrences are the basis of the network arcs in Figure 5.18.

A.2.1 Entity Linking on PoliInformatics

The DBpedia concepts and the mentions on the basis of which those concepts were annotated in the corpus are below. The table is limited to concept type *organization*. Column *Freq* in the table corresponds to the corpus frequency for each DBpedia entity. The number in parentheses after each mention indicates the mention's frequency.

Greyed out items correspond to concepts that were identified as erroneous. Information about annotation quality on the UI (p. 157) suggested that these concepts are likely wrong, and this was verified manually via corpus searches and by looking up the concept's definition in Wikipedia. The top network in Fig. 5.18 was created using all the concepts (whether correct or not). The bottom network of Fig. 5.18 represents the network once erroneous concepts had been deleted.

#	DBpedia label	Freq	Mentions
1	ABN AMRO	3	ABN Amro (1), ABN AMRO (1), ABN AMRO Bank N.V (1)
2	Ally Financial	6	GMAC (3), Ally Financial (2), GMAC Financial Services (1)
3	American International Group	115	AIG (110), AIG's (4), Aig (1)
4	Bank of America	49	Bank of America (44), BofA (2), B of A (2), bankofamerica.com (1)
5	Bank of America Home Loans	68	Countrywide (56), Countrywide Financial Corporation (9), Bank of America Home Loans (2), Countrywide Financial (1)
6	Bear Stearns	209	Bear Stearns (189), Bear Stearns Asset Management (19), Bear (1)
7	BlackRock	10	BlackRock (10)
8	Citigroup	58	Citigroup (40), Citi (13), Citicorp (3), citi (2)
9	Congressional Research Service reports	134	Report (86), report (48)
10	Deutsche Bahn	8	DB (8)
11	Deutsche Bank	1008	Deutsche Bank (955), DB (22), Deutsche (13), Bank (8), Deutsche Bank Securities (3), DEUTSCHE BANK (2), Deutsche Bank Securities, Inc (2), Deutsche Bank AG (1), deutsche (1), Deutsche Bank Securities Inc (1)
12	EBay	13	International (13)
13	Enron	39	Enron (28), Enron Corporation (7), Enron's (3), ask: Why (1)
14	FICO	62	FICO (58), Fair Isaac Corporation (2), Fair Isaac (2)
15	Fannie Mae	287	Fannie Mae (248), Fannie (22), Federal National Mortgage Association (12), fanniemae (4), Mortgage (1)

[continues on next page]

#	DBpedia label	Freq	Mentions
16	Federal Deposit Insurance Corporation	1300	FDIC (1162), Deposit Insurance Fund (67), Federal Deposit Insurance Corporation (39), federally insured (7), fdic (6), federal deposit insurance (4), Federal (3), federally insured bank (3), inception (3), Corporation (2), deposit insurance fund (2), FDIC-insured (1), Banking Act (1)
17	Federal Housing Administration	53	FHA (46), Federal Housing Administration (4), Federal (2), federal housing (1)
18	Federal Reserve System	174	Federal Reserve (125), Fed (24), Federal Reserve System (5), Federal Reserve Board (4), Federal Reserve Bank (3), The Federal Reserve (3), Federal Reserve's (2), Federal Reserve Board of Governors (2), FED (2), Federal Reserve board (1), fed (1), Reserve (1), FederalReserve (1)
19	Financial Industry Regulatory Authority	78	FINRA (63), finra (9), Financial Industry Regulatory Authority (6)
20	Fitch Group	50	Fitch (38), Fitch Ratings (12)
21	Freddie Mac	302	Freddie Mac (205), Freddie (83), Federal Home Loan Mortgage Corporation (12), Corporation (2)
22	Gemstone Publishing	91	Gemstone (91)
23	Goldman Sachs	2379	Goldman (890), Goldman Sachs (652), GS (392), Abacus (147), Abacus 2007-AC1 (120), Fabrice Tourre (64), ABACUS (42), Goldman Sachs International (16), Goldman, Sachs & Co (10), ABACUS 2007-AC1 (8), Goldman, Sachs (7), Abacus 2007- AC1 (4), Goldman Sachs Group Inc (3), Goldman Sachs Group (3), Goldman Sachs Group, Inc (3), GOLDMAN SACHS (2), Goldman, Sachs & Co. (2), Goldman Sachs & Co (2), Goldman Sachs Group, Inc. (2), ABACUS 2007- AC1 (2), goldmansachs (1), ABACUS-2007- AC1 (1), Sachs (1), ABACUS 2007 AC1 (1), goldman-sachs (1), abacus (1), Goldman Sachs Group Inc. (1), AIG bailout (1)
24	HM Treasury	4	Treasury (4)
25	HRG Engineering Company	16	Hrg (16)
26	Independent agencies of the United States government	234	agencies (189), agency (44), Administrator (1)
27	JPMorgan Chase	300	JPM (174), JPMorgan Chase (68), Chase (36), JPMorgan (13), J.P. Morgan Chase (6), JPMorgan Chase & Co (2), J.P. Morgan (1)
28	Law enforcement agency	12	enforcement (10), law enforcement agencies (2)
29	Lehman Brothers	268	Lehman (215), Lehman Brothers (50), lehman (3)

[continues on next page]

#	DBpedia label	Freq	Mentions
30	McGraw-Hill	15	McGraw-Hill Companies (6), The McGraw-Hill Companies (2), McGraw-Hill (2), The McGraw-Hill Companies, Inc (1), McGraw-Hill Companies, Inc. (1), mcgraw (1), mcgraw+hill (1), McGraw-Hill Companies, Inc (1)
31	Merrill Lynch	45	Merrill Lynch (41), Merrill Lynch, Pierce, Fenner & Smith (2), Merrill Lynch & Co (1), Merrill Lynch & Co. (1)
32	Moody's	585	Moody's (551), Moody's Investors Service (12), Baa2 (10), credit ratings (3), Moody's Investor Services (3), Investors (2), Moody's Corporation (2), Moody's Investor Service (1), Moody's Investors (1)
33	Morgan Stanley	233	Morgan Stanley (227), Morgan Stanley's (5), Stanley (1)
34	Nationally recognized statistical rating organization	38	NRSRO (13), Nationally Recognized Statistical Rating Organizations (10), NRSROs (10), Credit Rating Agency Reform Act of 2006 (3), nrsro (2)
35	New Century	101	New Century (87), New Century Financial Corporation (9), New Century Financial (5)
36	Nielsen ratings	350	ratings (350)
37	Office of Thrift Supervision	376	OTS (310), Office of Thrift Supervision (61), ots (3), OFFICE OF THRIFT SUPERVISION (2)
38	Office of the Comptroller of the Currency	89	Office of the Comptroller of the Currency (37), Comptroller of the Currency (33), OCC (17), occ (2)
39	OneWest Bank	76	IndyMac (60), IndyMac Bank (14), Indy Mac (1), Indymac (1)
40	Orion International	3	Orion (3)
41	PNC Financial Services	8	PNC Mortgage (4), PNC (3), PNC Financial Services (1)
42	Portfolio.com	63	portfolio (63)
43	Reuters	6	Reuters (4), Reuters.com (2)
44	Schneider Electric	58	Schneider (58)
45	Shooto	5	Shu (5)
46	Standard & Poor's	535	S&P (488), Standard & Poor's (37), Standard and Poor's (6), Standard & Poor (2), Standard and Poor (2)
47	The Wall Street Journal	30	Wall Street Journal (22), The Wall Street Journal (4), WSJ (2), the Wall Street Journal (1), wsj.com (1)
48	The Washington Post	20	Post (6), Washington Post (6), post (4), washingtonpost.com (4)
49	Total S.A.	40	total (38), Total (2)
50	U.S. Bancorp	16	Star Bank (10), U.S. bank (2), U.S. Bancorp (2), U.S. Bank (2)
51	U.S. Securities and Exchange Commission	838	SEC (667), commission (40), Securities and Exchange Commission (38), Commission (34), sec.gov (24), SEC's (24), U.S. Securities and Exchange Commission (4), Sec (3), us." SEC (1), Commission, members (1), sec (1), Securities and Exchange Commission's (1)
52	UBS	61	UBS (60), ubs (1)

[continues on next page]

#	DBpedia label	Freq	Mentions
53	USPG	10	SPG (10)
54	United States Department of Veterans Affairs	15	VA (8), Department of Veterans Affairs (4), Veterans Administration (2), Veterans Affairs (1)
55	United States Department of the Treasury	114	Treasury (64), Department of the Treasury (21), Treasury Department (9), U.S. Department of the Treasury (7), Department of Treasury (5), U.S. Treasury Department (3), U.S. treasury (2), treasury (2), U.S. (1)
56	United States Senate Committee on Banking, Housing, and Urban Affairs	18	Senate Committee on Banking, Housing, and Urban Affairs (4), Banking, Housing, and Urban Affairs (2), Senate Banking Committee (2), Committee on Banking, Housing, and Urban Affairs (2), Senate Committee on Banking, Housing and Urban Affairs (2), Banking Committee (1), Banking, Housing and Urban Affairs (1), U.S. Senate Committee on Banking, Housing, and Urban Affairs (1), Senate Committee on Banking and Currency (1), Committee on Banking, Housing and Urban Affairs (1), U.S. Senate Committee on Banking, Housing and Urban Affairs (1)
57	United States Senate Committee on Homeland Security and Governmental Affairs	25	Senate Committee on Governmental Affairs (8), Committee on Governmental Affairs (8), U.S. Senate Committee on Governmental Affairs (6), Committee on Homeland Security and Governmental Affairs (2), Homeland Security and Governmental Affairs (1)
58	United States federal courts	65	federal (40), Federal (23), federal Court (1), federal court (1)
59	Wachovia	18	Wachovia (14), Wachovia Bank (4)
60	Washington Mutual	3707	WaMu (2745), Washington Mutual (674), Washington Mutual Bank (121), Long Beach Mortgage (78), WAMU (30), Washington Mutual Inc (30), Washington Mutual Inc. (13), Long Beach mortgage (3), Washington Mutual Savings Bank (3), Wamu (3), Washington Mutual Bank, FSB (2), Washington Mutual, Inc. (1), Mortgage (1), Washington Mutual, Inc (1), WASHINGTON MUTUAL BANK (1), wamu (1)
61	Wells Fargo	24	Wells Fargo (22), Wells Fargo bank (2)

A.2.2 Co-occurrence table for PoliInformatics corpus network

The list of co-occurrences between mentions to organizations in the corpus is below. These co-occurrences are the basis of the network arcs in Figure 5.18. The minimum co-occurrence frequency considered was 4.

In greyed out arcs, at least one of the concepts (either the source or the target) was identified as erroneous (see p. 216 for details). The top network in Fig. 5.18 was created using all the arcs (whether correct or not). The bottom network of Fig. 5.18 represents the network once arcs containing erroneous concepts had been deleted.

These co-occurrences could be imported into the network analysis tool Gephi as an "edge-table" to reproduce the networks in Figure 5.18. To the same end, the layout parameters used to create the networks are listed on p. 224.

Source	Target	Weight
Moody's	Standard & Poor's	148
Washington Mutual	Federal Deposit Insurance Corporation	128
JPMorgan Chase	Washington Mutual	110
Washington Mutual	Office of Thrift Supervision	93
Freddie Mac	Washington Mutual	74
Moody's	Nielsen ratings	71
Standard & Poor's	Nielsen ratings	63
Fannie Mae	Washington Mutual	60
Fannie Mae	Freddie Mac	53
Office of Thrift Supervision	Federal Deposit Insurance Corporation	44..
Freddie Mac	Fannie Mae	42
Standard & Poor's	Moody's	36
Washington Mutual	Freddie Mac	35
Moody's	U.S. Securities and Exchange Commission	30
Total S.A.	Washington Mutual	25
Federal Reserve System	Federal Deposit Insurance Corporation	23
Shooto	Goldman Sachs	23
Goldman Sachs	Lehman Brothers	23
Office of Thrift Supervision	Washington Mutual	21
Washington Mutual	United States Department of the Treasury	19
Standard & Poor's	U.S. Securities and Exchange Commission	18
Nielsen ratings	U.S. Securities and Exchange Commission	18
Moody's	U.S. Securities and Exchange Commission	18
Deutsche Bank	Gemstone Publishing	17

[continues on next page]

Source	Target	Weight
U.S. Securities and Exchange Commission	Congressional Research Service reports	14
Fitch Group	Standard & Poor's	12
U.S. Securities and Exchange Commission	Nationally recognized statistical rating organization	12
United States federal courts	Federal Deposit Insurance Corporation	11
Goldman Sachs	U.S. Securities and Exchange Commission	11
JPMorgan Chase	Morgan Stanley	11
U.S. Securities and Exchange Commission	Independent agencies of the United States government	11
Citigroup	JPMorgan Chase	11
OneWest Bank	Office of Thrift Supervision	10
JPMorgan Chase	Federal Deposit Insurance Corporation	10
Nielsen ratings	Independent agencies of the United States government	10
OneWest Bank	Federal Deposit Insurance Corporation	10
Federal Reserve System	U.S. Securities and Exchange Commission	9
OneWest Bank	Washington Mutual	9
Federal Reserve System	Office of the Comptroller of the Currency	9
Fitch Group	Moody's	9
Goldman Sachs	Democratic Party (United States)	9
EBay	Goldman Sachs	8
Washington Mutual	Schneider Electric	8
Moody's	Standard & Poor's	8
Deutsche Bank	U.S. Securities and Exchange Commission	8
Deutsche Bank	Goldman Sachs	8
Fannie Mae	Federal Reserve System	8
Washington Mutual	Fannie Mae	8
Time (magazine)	U.S. Securities and Exchange Commission	8
JPMorgan Chase	Federal Reserve System	8
Office of the Comptroller of the Currency	Federal Deposit Insurance Corporation	8
JPMorgan Chase	Bank of America	8
Moody's	Deutsche Bank	8
Moody's	Congressional Research Service reports	8

[continues on next page]

Source	Target	Weight
Nielsen ratings	Nationally recognized statistical rating organization	6
USPG	Goldman Sachs	6
JPMorgan Chase	Office of Thrift Supervision	6
Bear Stearns	JPMorgan Chase	6
Goldman Sachs	JPMorgan Chase	6
Office of Thrift Supervision	Independent agencies of the United States government	6
JPMorgan Chase	Merrill Lynch	6
Nationally recognized statistical rating organization	Independent agencies of the United States government	6
Hoover's	Goldman Sachs	6
Reuters	Goldman Sachs	6
Goldman Sachs	Morgan Stanley	6
Standard & Poor's	Congressional Research Service reports	6
Federal Deposit Insurance Corporation	Office of Thrift Supervision	6
Washington Mutual	Goldman Sachs	6
United States federal courts	U.S. Securities and Exchange Commission	6
HRG Engineering Company	United States Senate Committee on Banking, Housing, and Urban Affairs	6
Deutsche Bahn	Deutsche Bank	6
Federal Deposit Insurance Corporation	Office of the Comptroller of the Currency	6
Goldman Sachs	American International Group	6
The Washington Post	The Wall Street Journal	6
Freddie Mac	Federal Housing Administration	6
Freddie Mac	Bear Stearns	6
Standard & Poor's	Washington Mutual	5
Moody's	Morgan Stanley	5
Freddie Mac	Federal Deposit Insurance Corporation	5
Deutsche Bank	Moody's	5
PNC Financial Services	Washington Mutual	5
Standard & Poor's	Federal Reserve System	5
Moody's	Time (magazine)	5
BlackRock	Standard & Poor's	5

[continues on next page]

Source	Target	Weight	Source	Target	Weight
Wachovia	JPMorgan Chase	5	U.S. Securities and Exchange Commission	United States Senate Committee on Homeland Security and Governmental Affairs	4
Deutsche Bank	Congressional Research Service reports	5	Fannie Mae	JPMorgan Chase	4
Goldman Sachs	Standard & Poor's	5	JPMorgan Chase	U.S. Securities and Exchange Commission	4
U.S. Securities and Exchange Commission	United States Senate Committee on Banking, Housing, and Urban Affairs	5	Moody's	McGraw-Hill	4
Bear Stearns	Morgan Stanley	5	Ally Financial	PNC Financial Services	4
Moody's	United States Senate Committee on Banking, Housing, and Urban Affairs	4	Bank of America	Bank of America Home Loans	4
Office of the Comptroller of the Currency	Office of Thrift Supervision	4	McGraw-Hill	Moody's	4
Orion International	Moody's	4	Standard & Poor's	Moody's	4
Moody's	Washington Mutual	4	Moody's	United States Senate Committee on Banking, Housing, and Urban Affairs	4
Office of Thrift Supervision	United States federal courts	4	JPMorgan Chase	American International Group	4
Merrill Lynch	Bank of America	4	Law enforcement agency	Federal Deposit Insurance Corporation	4
Wells Fargo	Federal Reserve System	4	FICO	Washington Mutual	4

PoliInformatics Network: Layout Parameters

The network described by the arcs listed above (p. 220) was spatialized in Gephi, using the *Force Atlas* layout. This type of layout was discussed on p. 116. The *Label Adjust* layout was also used in order to rearrange nodes whose labels overlap with other nodes' labels. *Label Adjust* does not otherwise modify the structure of the network.

Force Atlas	
Parameter	*Value*
Inertia	0.1
Repulsion	80000
Attraction	5
Maximum displacement	10
Auto stabilize	yes
Gravity	600
Attraction distribution	no
Adjust by sizes	yes
Speed	1

Label Adjust	
Parameter	*Value*
Speed	1
Include node size	yes

Layout parameters for network in Fig. 5.18, representing organizations in the PoliInformatics corpus, using Gephi's *Force Atlas* and *Label Adjust* algorithms.

Appendix B

Domain Model for Relation-based Navigation

The domain model for proposition extraction on the 12th volume in the Earth Negotiations Bulletin corpus[1] is reproduced below.[2] The corpus covers climate negotiations. The model consists in actors which participate in the negotiations, and predicates (verbs or nouns) whereby those actors emit their messages.

B.1 Actors 227
B.2 Predicates 231
B.2.1 Verbal predicates 232
B.2.2 Nominal predicates 232

B.1 Actors

Actors in the model are divided into three types: countries, groups of countries, and "generic actors", which refer to to generic roles at the conferences like *the chair*, *the delegates*, etc. A source for participant countries and groups is the UNFCC site.[3]

For **countries and groups**, the list below shows actor **labels** and their textual **variants**. The label generally corresponds to the DBpedia name for the actor, once the space is replaced by an underscore. On the User Interface (p. 174), generally the label is displayed. The label and variants were searched in a case-insensitive manner in the relevant semantic roles attached to speech predicates, following the procedure in Chapter 4 (p. 81). **Generic actors** are shown at the end (p. 231).

The model also divides countries into two types according to the climate agreements they have signed, but this attribute is not exploited in the thesis applications and is not shown on the list below.

Actor Label variants

Afghanistan afghanistan

Albania albania

Algeria algeria

Angola angola

Antigua and Barbuda antigua and barbuda

Argentina argentina

Armenia armenia

Australia australia

Austria austria

Azerbaijan Democratic Republic azerbaijan, azerbaijan democratic republic

Bahrain bahrain

Bangladesh bangladesh

Barbados barbados

[1] http://enb.iisd.org/enb/vol12/
[2] It is also online at https://sites.google.com/site/nlp4climate/domain-model
[3] United Nations Framework Convention on Climate Change: http://unfccc.int/parties_and_observers/items/2704.php

Belarus belarus

Belgium belgium

Belize belize

Benin benin

Bhutan bhutan

Bolivia bolivia

Bosnia and Herzegovina bosnia, bosnia and herzegovina

Botswana botswana

Brazil brazil

Brunei brunei, brunei darussalam

Bulgaria bulgaria

Burkina Faso burkina faso

Burundi burundi

Cambodia cambodia

Cameroon cameroon

Canada canada

Cape Verde cabo verde, cape verde

Central African Republic central african republic

Chad chad

Chile chile

China china

Colombia colombia

Comoros comoros

Cook Islands cook islands

Costa Rica costa rica

Côte d'Ivoire cote d'ivoire, côte d'ivoire

Croatia croatia

Cuba cuba

Cyprus cyprus

Cyprus cyprus

Czech Republic czech republic

Democratic Republic of the Congo democratic republic of the congo

Denmark denmark

Djibouti djibouti

Dominica dominica

Dominican Republic dominican republic

East Timor east timor, timor-leste

Ecuador ecuador

Egypt egypt

El Salvador el salvador

Equatorial Guinea equatorial guinea

Eritrea eritrea

Estonia estonia

Ethiopia ethiopia

Fiji fiji

Finland finland

France france

Gabon gabon

Georgia (country) georgia

Germany germany

Ghana ghana

Greece greece

Grenada grenada

Guatemala guatemala

Guinea-Bissau guinea-bissau

Guyana guyana

Haiti haiti

Honduras honduras

Hungary hungary

Iceland iceland

India india

Indonesia indonesia

Iran iran

Iraq iraq

Israel israel

Italy italy

Jamaica jamaica

Japan japan

Jordan jordan

Kazakhstan kazakhstan

Kazakhstan kazakhstan

Kenya kenya

Kiribati kiribati

Kuwait kuwait

Kyrgyzstan kyrgyzstan

Laos lao people's democratic republic, laos

Latvia latvia

Lebanon lebanon

Lesotho lesotho

Liberia liberia

Libya libya

Liechtenstein liechtenstein

Lithuania lithuania

Luxembourg luxembourg

Madagascar madagascar

Malawi malawi

Malaysia malaysia

Maldives maldives

Mali mali

Malta malta

Malta malta

Marshall Islands marshall islands

Mauritania mauritania

Mauritius mauritius

Mexico mexico

Micronesia micronesia

Moldova moldova, republic of moldova

Monaco monaco

Mongolia mongolia

Montenegro montenegro

Morocco morocco

Mozambique mozambique

Myanmar myanmar

Namibia namibia

Nauru nauru

Nepal nepal

Netherlands netherlands

New Zealand new zealand

Nicaragua nicaragua

Nigeria nigeria

Niger niger

Niue niue

North Korea democratic people's republic of korea, north korea

Norway norway

Oman oman

Pakistan pakistan

Palau palau

Panama panama

Papua New Guinea papua new guinea

Paraguay paraguay

Peru peru

Philippines philippines

Poland poland

Portugal portugal

Qatar qatar

Republic of Ireland ireland, republic of ireland

Republic of Macedonia macedonia, republic of macedonia

Republic of the Congo republic of the congo

Romania romania

Russia russia, russian federation

Rwanda rwanda

Saint Kitts and Nevis kitts and nevis, saint kitts and nevis

Saint Lucia saint lucia

Saint Vincent and the Grenadines saint vincent and grenadines, saint vincent and the grenadines

Samoa american samoa, samoa

San Marino san marino

São Tomé and Príncipe sao tomé and principe, são tomé and príncipe

Saudi Arabia saudi arabia

Senegal senegal

Serbia serbia

Seychelles seychelles

Sierra Leone sierra leone

Singapore singapore

Slovakia slovakia

Slovenia slovenia

Solomon Islands solomon islands

Somalia somalia

South Africa south africa

South Korea republic of korea, south korea

Spain spain

Sri Lanka sri lanka

Sudan sudan

Suriname suriname

Swaziland swaziland

Sweden sweden

Switzerland switzerland

Syria syria, syrian arab republic

Tajikistan tajikistan

Tanzania tanzania, united republic of tanzania

Thailand thailand

The Bahamas bahamas, the bahamas

The Gambia gambia, the gambia

Togo togo

Tonga tonga

Trinidad and Tobago trinidad and tobago

Tunisia tunisia

Turkey turkey

Turkmenistan turkmenistan

Tuvalu tuvalu

Uganda uganda

Ukraine ukraine

United Arab Emirates united arab emirates

United Kingdom the uk, united kingdom

United States the us, united states, united states of america

Uruguay uruguay

Uzbekistan uzbekistan

Vanuatu vanuatu

Venezuela venezuela

Vietnam viet nam, vietnam

Yemen yemen

Zambia zambia

Zimbabwe zimbabwe

GROUPS

African Group africa group, african group, the african group

ALBA alba, alliance of bolivarian states for the peoples of our america, bolivarian alliance, bolivarian alliance for the peoples of our america, bolivarian states for the peoples of our america

Alliance of Small Island States alliance of small island states, aosis

Asian Group asia group, asian group

Caribean Community caribean community, caricom, the caribean community, the caricom states

Central America Group central america group

Central Asia, Caucasus, Albania and Moldova cacam; central asia, caucasus, albania and moldova

Coalition for Rainforest Nations cfrn

Eastern European group eastern european group, eeg, group of eastern european countries

European Union eu, european union, the eu

Environmental Integrity Group eig, environmental integrity group

Group of 77 g -77, g 77, g- 77, g-77, g77, group of 77

Independent association of Latin America and the Caribbean ailac, alliance of independent latin american and caribbean states, association of independent latin american and caribbean states, independent alliance of latin america and the caribbean, independent association of latin america and the caribbean

Latin American and Caribbean Group group of latin american and caribbean countries, group of latin american and caribbean states, grulac, latin american and caribbean group

League of Arab States arab group, arab league, arab states group, group of arab states, league of arab states

Least Developed Country ldc, ldcs, least developed country, the ldc, the ldcs

Like Minded Group group of like-minded, like minded group, like-minded group, lmdc, lmdcs, the lmdc, the lmdcs

Organisation for Economic Co-operation and Development oecd, organisation for economic co-operation and development, organization for economic cooperation and development

Organization of Petroleum Exporting Countries opec, organization of petroleum exporting countries

Umbrella Group umbrella group

Valdivia Group valdivia group

Western European and Others Group weog, western european and others group

FORMER COUNTRIES

Yugoslavia yugoslavia

Serbia and Montenegro serbia and montenegro

GENERIC ACTORS

The triggers for generic actors are *committee*, *delegate*, *party*, *chair*, *participant* and their plurals.

B.2 Predicates

Both verbal and nominal predicates are considered. Predicate attributes are the predicate type: support (e.g. the noun *preference*), opposition (e.g. the verb *oppose*), or general reporting (e.g. the verb *state*).

All tokens in the text are lemmatized, and lowercase matching is performed against the domain predicate lemmas.

The predicate lemmas are listed overleaf.

B.2.1 Verbal predicates

General reporting	suggest, acknowledge, add, address, admit, allude, announce, answer, argue, ask, attribute, brief, circulate, cite, claim, clarify, comment, compare, conclude, confirm, consider, corroborate, debate, declare, demand, demonstrate, describe, differentiate, discuss, distinguish, elaborate, emphasize, enquire, estimate, explain, express, highlight, identify, indicate, inform, inquire, insist, label, learn, listen, maintain, manifest, mention, moderate, negotiate, note, notice, notify, observe, offer, perceive, ponder, portray, present, propose, reaffirm, realize, recognize, reconsider, redefine, re-evaluate, refer, reformulate, reiterate, remind, repeat, reply, report, request, respond, restate, reveal, review, revise, say, scrutinize, specify, state, stress, summarize, suppose, swear, synthesize, tell, underline, underscore, understand, utter, voice
Opposition	accuse, attack, alert, apologize, banish, ban, blame, caution, complain, condemn, conflict, contradict, criticize, decline, decry, deny, deplore, disagree, discourage, dispute, distance, doubt, fail, fear, forbid, frustrate, lament, object, obstruct, oppose, praise, question, refuse, refute, regret, reject, resist, threaten, validate, veto, withdraw
Support	accept, adopt, advocate, agree, allege, allow, appeal, applaud, approve, authorize, boast, call, commend, concede, congratulate, defend, encourage, endorse, favor, follow, further, guarantee, justify, laud, lobby, obey, plead, permit, pledge, prefer, promise, promote, rectify, recommend, re-emphasize, support, urge, vindicate, welcome, wish

B.2.2 Nominal predicates

General reporting	address, admission, allusion, announcement, answer, argument, attribution, circulation, citation, claim, clarification, comment, comparison, conclusion, confirmation, consideration, corroboration, debate, declaration, demand, demonstration, description, differentiation, discussion, distinction, elaboration, enquiry, estimate, estimation, explanation, expression, identification, indication, information, inquiry, inquiry, insistence, label, manifestation, mention, moderation, negotiation, note, notice, notification, observation, offer, perception, portrayal, presentation, proposal, question, reaffirmation, realization, recognition, reconsideration, redefinition, re-evaluation, reformulation, reiteration, repetition, reply, report, request, response, restatement, revelation, review, revision, scrutiny, statement, summary, supposition, synthesis, utterance
Opposition	accusation, alert, apology, attack, banishing, ban, blame, complaint, condemnation, conflict, contradiction, criticism, decline, denial, disagreement, discouragement, dispute, distance, doubt, failure, fear, frustration, lament, objection, opposition, praise, prohibition, questioning, refusal, refutation, regret, rejection, resistance, validation, veto, withdrawal
Support	acceptance, adoption, agreement, allegation, appeal, approval, bid, authorization, call, commendation, compliance, concession, congratulations, defence, defense, emphasis, encouragement, endorsement, guarantee, justification, permission, pledge, preference, promise, promotion, recommendation, rectification, suggestion, support, vindication, welcoming, wish

Appendix C

Test-Sets for Intrinsic Evaluation

Besides evaluations with domain-experts, reference sets were used for intrinsic evaluation, in two cases:

1. Entity Linking System Combination method in Chapter 3 (Tables 3.1 and 3.2)
2. Proposition Extraction method in Chapter 4 (Table 4.4)

The test-sets take too much space to print, and I have therefore made them available on a website:

https://sites.google.com/site/thesisrf/

The site explains the data format. Besides the test-sets, system results, and steps to reproduce the evaluation are also provided on the site.

Appendix D

Domain-Expert Evaluation Reports

This appendix contains evaluation-reports for the the domain-experts evaluation sessions. The evaluation procedure was explained in subsection 6.6.3. The reports consist in a non-verbatim transcript of the sessions, enriched with some explanations.

The reports are organized as follows (with small variations in each report):

- Basic data about expert and evaluation session: Location and time of the session, expert's identity etc.
- Queries run by expert
  - **Query**: Search function and search-terms used
  - **Results**: Aspect of the results focused on by the expert
  - **Enabling function**: *Only documented in Session 1.* It refers to a UI function other than the original query (sorting, filtering) that made the expert's manipulation possible.
  - **Expert comment** or **Discussion**: Expert's remarks after each query, or discussions between the expert and the experimenter.
- Other comments by expert *(if applies)*
- Weaknesses of the tool according to the expert
- Comments about expert's use of the tool *(if applies)*
- Incidences *(if applies)*

Unless otherwise stated, the time range for a query is 1995–2015, and the confidence range is 5–5 (i.e. confidence 5 only).

Audio files for Sessions 2 and 3 are publicly available at this link.[1] There is no audio for Session 1, but the evaluation report included here was validated by the expert (see 6.6.3.4).

The experimenter was myself (i.e. the main contributor to the UI).

[1] https://drive.google.com/drive/folders/0B41tv-I-4xMJTW43MFhEekFtejQ

D.1 Session 1

D.1.1 Basic session data 239
D.1.2 Queries run by expert 239
D.1.2.1 Query 1 239
D.1.2.2 Query 2 240
D.1.2.3 Query 3 240
D.1.2.4 Query 4 240
D.1.2.5 Query 5 240
D.1.2.6 Query 6 240
D.1.2.7 Query 7 240
D.1.2.8 Query 8 241
D.1.2.9 Query 9 241
D.1.2.10 Query 10 241
D.1.2.11 Query 11 241
D.1.2.12 Query 12 242
D.1.2.13 Query 13 242
D.1.3 Other comments by the expert 242
D.1.3.1 New research idea 242
D.1.3.2 Clear examples for an actor 242
D.1.3.3 General comments about the tool 242
D.1.4 Weaknesses pointed out by the expert 243
D.1.4.1 Weakness 1 243
D.1.4.2 Weakness 2 243
D.1.4.3 Weakness 3 243
D.1.5 Comments on expert's use of the UI 244

D.1.1 Basic session data

Expert: The expert, Kari de Pryck, is a researcher specialized in climate negotiations and the ENB corpus, has previously published work on these topics and is completing a PhD thesis at Sciences Po in Paris. She is also an assistant at the University of Geneva.

Time and place: June 16, 2016, at the expert's institution (Sciences Po), using `https://apps.lattice.cnrs.fr/ie/uidev/`. I also had access to a local version, which was equivalent to the online one in terms of the functions looked at by the expert.

Duration: The session took around 1 hour 15 minutes.

UI Versions: UI version was commit *a1785da*[2] (online) and commit *c37a4fe* (local)

Incidences: Started evaluation with the online UI at commit *97bd37a*. But an error was found in sentence highlighting, which crashed the `Docs` pane when trying to show a sentence's context. This was fixed on the spot on the online version, and the fix was committed as *a1785da* after the evaluation. The local version never showed that bug.

D.1.2 Queries run by expert

D.1.2.1 Query 1

QUERY: *African Group* in `Actors` box.
RESULTS: The results looked at by the expert were the content of the propositions.
EXPERT COMMENT: It's useful to see when the group started participating in the climate conferences (COPs) as a group (1998).
ENABLING FUNCTION: Sorting the propositions by `Year`.

[2]Commit hashes correspond to a private code repository but are noted here to document the version univocally.

D.1.2.2 Query 2

QUERY: *loss and damage* in the `Points` box
RESULTS: She focused on the proposition pane.
EXPERT COMMENT: Makes sense that actors are majoritarily underdeveloped countries.

D.1.2.3 Query 3

QUERY: *Saudi Arabia* in `Actors`, *`oppose`* checkbox.
RESULTS: The 42 propositions returned.
EXPERT COMMENT: This is a good query to make, because this country tends to oppose negotiation issues regularly.
ENABLING FUNCTION: Being able to filter the propositions by predicate types (*`oppose`*, *`support`*, *`report`*). Proposition counts.
OTHER OBSERVATIONS: I discussed with the expert that maybe we should compare the ratio of opposing vs. supporting propositions for this actor to the *oppose/support* ratio in propositions for another actor whose behaviour is not characterized by opposition. This resulted in some related queries (below).

D.1.2.4 Query 4

QUERY: *Saudi Arabia* in `Actors`, *`support`* checkbox (performed following our discussion about the results for Query 3)
RESULTS: The 55 propositions extracted.
EXPERT COMMENT: She mentioned that the ratio between opposition and support propositions (3:2) agrees with her knowledge that this actor tends to oppose a lot.
ENABLING FUNCTION: Being able to filter the propositions by predicate types (*`oppose`*, *`support`*, *`report`*), proposition counts.

D.1.2.5 Query 5

QUERY: *AOSIS* in `Actors`, *`support`* checkbox, and later *`oppose`* checkbox (performed following our discussion about the results for Query 3)
RESULTS: 86 supporting propositions, 17 opposing ones.
EXPERT COMMENT: Kari says that seeing the ratios for this actor agrees with her previous comment that Saudi Arabia has a strong tendency to oppose negotiation points, compared to other actors like AOSIS, who expresses much more support than opposition.
ENABLING FUNCTION: Being able to filter the propositions by predicate types (*`oppose`*, *`support`*, *`report`*), proposition counts.

D.1.2.6 Query 6

QUERY: *adaptation and mitigation* in the `Points` box.
RESULTS: Kari focused on the propositions extracted.
EXPERT COMMENT: Kari's comment was that it is useful to look at when this exact phrase starts, because at the beginning of climate negotiations, talking about adaptation was perceived as conceding failure to mitigate climate change, and only later on it was accepted that adaptation was unavoidable. So it is "political" to use both terms together. We see that the phrase is not mentioned until 2001, year of the publication of the IPCC TAR[3] which put adaptation on the international agenda.
ENABLING FUNCTION: Sorting propositions by *`Year`*.

D.1.2.7 Query 7

QUERY: In `AgreeDisagree` tab, *Saudi Arabia* and *the US*, looking at relation types *`agreement`* and *`disagreement`*.

[3]Third Assessment Report (TAR) by the Intergovernmental Panel on Climate Change (IPCC), `https://www.ipcc.ch/ipccreports/tar/`

RESULTS: Kari focused on the terms extracted to characterize points the actors agree or disagree about (i.e. the `KeyPhrase`, `DBpedia` and `ClimTag` tabs).

EXPERT COMMENT: She explained that this is a good query to make, because she would expect these actors to disagree often with each other. She mentioned that agreement takes place over procedural issues (evidenced by keyphrases like *agenda items* or *committee rules*), whereas disagreement takes place over more "political" issues that have implications for the countries' economy. For instance, *adverse effects of response measures*, which refers to the adverse effects of limiting fossil fuels for Saudi Arabia's economy.

D.1.2.8 Query 8

QUERY: *REDD* in the `Points` box, and *Brazil* in the `Actors` box.

RESULTS: She focused on the propositions. There was only one result.

EXPERT COMMENT: She found it surprising that there was only one result for Brazil talking about *REDD*.

OTHER OBSERVATIONS:

- The context for this query is that I asked the expert to search some of the central terms in the clusters of Figure 5 of the Venturini et al., 2014 article on ENB (clusters for *REDD* and *LULUCF*[4] at this link), to see how she assessed the information she got from the UI about those terms (the expert is a co-author of that article).
- The expert's comment about this was that it would be useful to know who the actors who mentioned the terms in the clusters are. That reading the outputs of the UI would complement the information provided by a network like those in her article, and that it would have been a good way to validate the network (easier than reading the full ENB issues, which is the way they validated the networks, working in a group of several people, when they wrote their article).
- To address the issue that there was only one result, I asked the expert what the term *REDD* means. It's an acronym for *Reducing Emissions from Deforestation and Forest Degradation*. I suggested looking for terms like *forest* in the `Points` or `Free Text` boxes, or wildcard searches like `*forest*` in the `Free Text` box. This resulted in queries 9 through 12 below.

D.1.2.9 Query 9

QUERY: *REDD* in the `Free Text` box, *Brazil* in `Actors` box.

RESULTS: Kari focused on the propositions. There were two results.

EXPERT COMMENT: Kari found it surprising that there were only two propositions for *Brazil* and *REDD*.

OTHER OBSERVATIONS: I suggested other queries that may return more results (queries 10 through 12 below).

D.1.2.10 Query 10

QUERY: *forest* in `Points` box, *Brazil* in `Actors` box.

RESULTS: Kari focused on the one proposition extracted.

EXPERT COMMENT: Kari found it surprising that there was only one proposition.

OTHER OBSERVATIONS: I suggested other queries that may return more results (queries 11 and 12 below).

D.1.2.11 Query 11

QUERY: *forest* in `Free Text` box, *Brazil* in `Actors` box.

RESULTS: Kari focused on the one proposition extracted.

EXPERT COMMENT: Kari found it surprising that there was only one proposition.

OTHER OBSERVATIONS: I suggested another query that may return more results (Query 12 below).

[4] *REDD* stands for "Reducing Emissions from Deforestation and Forest Degradation" and *LULUCF* means "Land use, land-use change and forestry".

D.1.2.12 Query 12

QUERY: **forest* OR REDD* in `Free Text` box, *Brazil* in `Actors`.
RESULTS: The expert looked at the 14 propositions extracted.
EXPERT COMMENT: Her comment was that the results are reasonable now with this new query.
OTHER OBSERVATIONS: This query is the successful attempt to get relevant results after queries 8 through 11 above, which did not retrieve enough information. See weaknesses discussion below.

D.1.2.13 Query 13

QUERY: *sustainable development* in the `Points` box.
RESULTS: She focused on the propositions.
EXPERT COMMENT: She said that she expected developing countries to be very present in the results, which was verified. However, developed countries were also found (like Australia), which might suggest their support for developing countries' claims in her opinion.

D.1.3 Other comments by the expert

This refers to comments that the expert made as a reflection on the tests she was carrying out, and may be based on her experience with several queries.

D.1.3.1 New research idea

Kari mentioned that some of the propositions referred to procedural content, e.g. a country's statement that they will submit a draft on some issue. She first said that these propositions are not very useful. However, she later reflected that a new research idea would be to compare country profiles in terms of how many of their interventions are procedural and how many contain real negotiation content. Because some countries' delegations are very "legally-oriented', with many lawyers (e.g. Brazil), and their interventions seek to control the legal aspects of the negotiation. Whereas other countries focus more on actions to undertake (e.g. AOSIS, who was one of the groups that promoted an adaptation agenda).

D.1.3.2 Clear examples for an actor

Kari found it gratifying to come across clear examples of what she knew about actor Saudi Arabia's negotiation behaviour. Reading through the propositions extracted for actor *Saudi Arabia* mentioning *energy* in the messages, she found two propositions, for two different COPs, that clearly illustrate some of this actor's concerns in her opinion:

Saudi Arabia	reminded	Parties that the UNFCCC is not an energy convention
Saudi Arabia	Stressing	that the UNFCCC is not an energy convention

Kari explained that this illustrates how this actor is concerned about measures against fossil fuels (a source of energy), since this type of fuels is an important part of their economy.

D.1.3.3 General comments about the tool

The expert made the following general comments:

(a) The tool provides a clear navigation compared to ClimateDebateExplorer (Venturini et al., 2015) because you don't have to search for the individual sentence where an actor said something; the sentences are already available.

(b) The tool is a bit "Latourian" because it helps to follow actors (in time and in the subjects they discussed).

(c) Kari first stated that "you need a research question" to exploit the tool for. However, following some more testing, she also found the tool "useful for both exploration and validation of hypotheses".

D.1.4 Weaknesses pointed out by the expert

Issues that the expert considered weaknesses are documented here. A possible solution and action-items aimed at improving on those issues are listed where relevant.

D.1.4.1 Weakness 1

Description: It was unclear why we got so few propositions for *Brazil* in the `Actor` box and the term *REDD* or the term *forest* in the `Points` box.

Possible solution: This need not be a weakness in the sense that, by querying (in the `Free Text` box) for *REDD or *forest**, the results for *Brazil* did agree with the expert's intuition. The results included mentions to terms like *afforestation* or *deforestation*, that, while relevant to the notion of *forest*, do not match a query for *forest* itself.

The general idea is that results may not provide enough coverage if the user does not employ the right search terms for the corpus. Searches automatically expanded with related terms or synonyms would be helpful, but the tool does not currently offer them.

Searching in the `Free Text` box helps retrieve more results than searching in the `Points` box. For now you then need to verify whether the search terms are indeed contained in the Point or not.

Action Items:

- Currently, terms searched in the `Free Text` box are not highlighted in the points column (the only thorough highlight in the points column is when you click on an item on the `KeyPhrase` tab). It would help the users to highlight the `Free Text` search terms in the points; this would allow for quick verification by a user if the term is in the point or not.
- Currently, the `Points` box does not allow for wildcard searches or other search operators (only the `Free Text` box does). This could be implemented to help retrieve more propositions.

D.1.4.2 Weakness 2

Description: One case of a wrong proposition extraction was found when looking at query results (The sentence involved several predicates).

Possible solution: Not really a solution, but we can make users aware that around 30% of the propositions can be expected to be wrong. Since the user can click on the proposition to have access to the sentence, the user is still able to identify the error instead of letting it go unnoticed and taking the extraction as correct at face value.

Action Items: Creating an *About* page accessible from the UI where this is explained.

D.1.4.3 Weakness 3

Description: Expert was surprised that we're tagging the text with Wikipedia items (via entity linking to DBpedia), in the sense that she does not consider Wikipedia content reliable.

Also, she finds that, for someone who is familiar with the subject, DBpedia concepts are not informative (that they may be useful for people not familiar with the corpus). She finds the Reegle Thesaurus (Bauer et al., 2011)[5] concepts more informative.

Possible solution: To me as the experimenter, this is not a weakness but a different way to describe the corpus, that could be useful to users that are not proficient with the material (as opposed to domain experts, who are indeed the target audience for the UI).

Action items: None is planned as I don't consider this a weakness.

[5]The thesaurus terms were tagged with the Climatetagger API, `http://www.climatetagger.net/climate-tagger-api/`

D.1.5 Comments on expert's use of the UI

(a) Chronological sorting: the expert regularly sorted results chronologically. She pointed out that this was useful to compare an actor's position across time.

(b) Confidence scores: Propositions have been tagged with a confidence score: e.g. 5 (the maximum) for very complete propositions where the actor is a negotiation group or country, no anaphora resolution has been applied, and where the message looks sound (e.g. it does not equal uninformative contents like *it* or *this idea*). Then propositions get lower scores if the actors correspond to other groups, e.g. NGOs, or if the messages look uninformative as just described. The expert appreciated having access to actors that are not countries or groups by lowering the confidence score.

(c) Use of `Docs` panel to get context for a proposition: The expert appreciated being able to locate the sentence containing a proposition inside its document, in order to better understand the meaning of the proposition and its sentence (this can be done by clicking on the proposition while the `Docs` panel is active).

(d) Lack of use of keyphrases or concepts: The expert did not generally use the right pane tabs for keyphrases, DBpedia concepts or Climatetagger thesaurus concepts, that had been extracted from the proposition's messages.

The expert was then asked explicitly to look at those panes and judge the content: Her opinion was favourable to the domain-specific Climatetagger concepts (Bauer et al., 2011), but disliked the idea of using DBpedia concepts since she considers the information too general.

Reasons why the expert did not use the summarized, "vertical" reading allowed for by a list of keyphrases or concepts may be the following: She is very familiar with the material and has no problem reading the corpus sentences, even when they contain technical, specialized terms and acronyms. She does not really need the more "diluted" version of the content provided by a list of keyphrases/thesaurus concepts. Relatedly, she does not need to verify definitions for a term by clicking on the link to the thesaurus or to the DBpedia entry.

A way to verify whether familiarity with the material decreases the use of the lists of keyphrases or concepts would be to test the UI with users who are not domain-experts. However, this is out of the scope of the thesis.

D.2 Session 2

D.2.1 Basic session data 245
D.2.2 Queries run by expert 246
D.2.2.1 Query 1 246
D.2.2.2 Query 2 246
D.2.2.3 Query 3 246
D.2.2.4 Query 4 246
D.2.2.5 Query 5 247
D.2.2.6 Query 6 247
D.2.2.7 Query 7 247
D.2.2.8 Query 8 248
D.2.2.9 Query 9 248
D.2.2.10 Query 10 248
D.2.2.11 Query 11 249
D.2.2.12 Query 12 249
D.2.2.13 Query 13 250
D.2.2.14 Query 14 250
D.2.2.15 Query 15 250
D.2.2.16 Query 16 251
D.2.2.17 Query 17 251
D.2.2.18 Query 18 251
D.2.2.19 Query 19 252
D.2.2.20 Query 20 252
D.2.3 Other comments by the expert 252
D.2.4 Weaknesses pointed out by the expert 253
D.2.4.1 Weakness 1 253
D.2.4.2 Weakness 2 253
D.2.5 Comments on expert's use of the UI 254

D.2.1 Basic session data

Expert: Tommaso Venturini is a lecturer at the Digital Humanities Department at King's College (London) and an associate researcher at the médialab of Sciences Po (Paris). His research areas are Digital Methods, Controversy Mapping and Social Modernization. He has led the EMAPS[6] and MEDEA[7] projects, which studied different aspects of adaptation to climate change.

Time and place: June 24, 2016 at one of the expert's institutions (Sciences Po Paris), using the UI's public address.[8] (I also had access to a local version, which was equivalent to the online one in terms of the functions looked at by the expert).

Duration: The session took around 1 hour 20 minutes.

UI Versions: UI version was commit *c1d36ce*[2] (online) and commit *d1fc766* (local). The differences between these versions and the versions used in the UI evaluation on June 16 (with KdP) are minor "cosmetic" fixes that cannot be expected to have a major impact on the UI's usefulness.

Incidences: Sentence highlighting inside the document did not work with a sentence containing the character *ń* (for the *Poznań* COP). This is due to JSON escaping imperfections in the highlighting code I wrote.

[6] http://www.emapsproject.com/blog/
[7] http://projetmedea.hypotheses.org/
[8] https://apps.lattice.cnrs.fr/ie/uidev/

D.2.2 Queries run by expert

D.2.2.1 Query 1

QUERY: It was a blank query.
RESULTS: Looking at the verbs used on the default initial results.
DISCUSSION:

Tommaso asked why in the `Action` column we can see a past-tense verb-form like *added* and also an infinitive or present-tense like *agree*.

The answer is that the verb-form used in the original sentence is displayed on the UI, even if the NLP workflow lemmatizes verb-forms, so that the predicate list used in the workflow only needs to contain infinitives.

D.2.2.2 Query 2

QUERY: `ActionView` tab, reverse-sort on `Action`
RESULTS: Focused on predicate *urge*.
DISCUSSION:

Tommaso points out that it would be nice to have access at a single glance to the whole list of verbs used by actors in the corpus, in order to get more ideas what to search in the corpus.

I answered that a list of the verbs analyzed is currently available on a "companion website" to the papers published around the interface and its NLP pipeline.[9]

Also, that it is possible to look at types of verbs by using the checkboxes for `support`, `oppose`, `report`.

D.2.2.3 Query 3

QUERY: *urge* in `Actions` box.
RESULTS: focused on DBpedia concepts tagged in the messages for the 420 propositions containing *urge* as a predicate.
DISCUSSION: We discussed three topics:

Topic 1

Tommaso points out that with a predicate like *urge*, the analysis is imperfect, since you could have actors or bodies in the DBpedia concept that were urged to do something, rather than being part of what was requested to do.

The answer is that it is true that the **argument structure** of a verb like *urge* is not perfectly captured. This is a limitation for the analysis of statements with that verb, but does not affect the validity of extractions for the majority of verbs.

Topic 2

Tommaso also pointed out that it would be interesting to have access to the complete **list of actors** in the corpus (for the same reason why he asked for the list of predicates).

The answer to this is that they list of actors is also available on the companion site.[9] An `About` page could be created on the UI in order to make these lists directly accessible to users.

Topic 3

There was also a clarification question whether the **DBpedia concepts** are extracted from the propositions' points or not limited to the points (i.e. from the whole sentence). The answer is that they are only extracted from the points.

D.2.2.4 Query 4

QUERY: The verb *attack* in the `Actions` box.
RESULTS: 0 propositions.
DISCUSSION:

[9] https://sites.google.com/site/nlp4climate/domain-model/predicates

Tommaso performed this query after seeing that the verb *attack* was on the list of predicates given on the companion site, since he considers this verb interesting, because it would be a very strong statement.

Seeing that there are no results, I clarified that the verbs on the site are not the verbs found in the corpus, but the ones that were searched for (against the output of Semantic Role Labeling and Dependency parsing).

Tommaso then asked that, since I was using a predefined list of predicates, not a list truly emerging from the corpus, how I knew that I was not missing predicates present in the corpus but absent from my list.

The answer is that I cannot be sure. However, the list of predicates, based on VerbNet (Kipper-Schuler, 2005) and NomBank (Meyers et al., 2004), contains ca. 200 verbs and more than 150 nominal predicates, and I consider this number of predicates likely to provide a good coverage.

Tommaso suggested that, on the list in the companion site, I specify how many times each predicate was actually found in the corpus.

D.2.2.5 Query 5

QUERY: *frustrate* in `Actions` box.
RESULTS: no propositions extracted.
DISCUSSION: Tommaso searched for this since it is a connotated (not neutral) verb, so it would be interesting if it was found in the ENB corpus, which claims to adopt a neutral tone. No results were found, but this was to be expected.

D.2.2.6 Query 6

QUERY: *blame* in `Actions` box, at varying confidence ranges.
RESULTS: No propositions were extracted at confidence 5, but one result was extracted when confidence is 0.
ENABLING FUNCTION: Confidence scores and sentence-lookup in the document.
DISCUSSION:

Topic 1

The reason for this query is the same as queries 4 and 5 above, it is an interesting **verb** since it is **not neutral** and the corpus is supposed to use a neutral tone.

At confidence 5, there were no results.

However, if we decrease confidence to 3, there is a proposition, with an indefinite subject (*Some blamed ...*). Such propositions with a non explicit subject are typical for the *in the corridors* section of ENB. We verified that the sentence containing the proposition is part of an *in the corridors* section by looking for the sentence in the document using the `Docs` tab.

Topic 2

Tommaso pointed out that it was not clear to him which were the **dropdown** menus for selecting the **confidence range**.

There are however tooltips on all the UI elements. Labels had been avoided not to clutter the UI. I thought that users would hover over the elements and notice the tooltip, the previous expert had not complained about this.

D.2.2.7 Query 7

QUERY: selecting all opposition predicates by using the *oppose* checkbox.
RESULTS: the expert clicked on a proposition to see its sentence in the document.
DISCUSSION: Tommaso performed this query to understand how the proposition-lookup in the document works.

There was an incidence here since one of the propositions he searched was not being highlighted because of a bug related to escaping rare characters with diacritics (like *ń* in the word *Poznań*).

D.2.2.8 Query 8

QUERY: Playing with different confidence-score ranges.
RESULTS: Tommaso focused on different proposition elements (actors, points) that were being displayed for each range
DISCUSSION:

Tommaso finds an actor involving indigenous people very interesting (he didn't expect to be able to find such actors), which gave rise to query 9 below.

He found actor *market-based mechanisms* strange.

I explained that this is not an error. Syntactically it is OK for this noun-phrase to be extracted as an actor. Moreover, the actor who is the real subject of the reporting event (*International Forum of Indigenous Peoples on Climate Change*) was also extracted. These are the two propositions extracted:

(a) market-based mechanisms, threaten, rights to land and culture
(b) The INTERNATIONAL FORUM OF INDIGENOUS PEOPLES ON CLIMATE CHANGE, expressed, concern with market-based mechanisms, which threaten rights to land and culture

D.2.2.9 Query 9

QUERY: *indigenous people* in `Actors` box
RESULTS: looking at propositions extracted
DISCUSSION:

Topic 1

Tommaso found that an **actor** involving ***indigenous peoples*** was very interesting, and he didn't think it was possible to extract propositions where the actor is not an explicit country or country group. He finds these actors interesting to look at and he did not expect he would be able to do so. He would also be interested in having a list where, unlike in the list I gave in the companion site,[9] these types of alternative actors are mentioned (not just the countries and country groups).

Topic 2

Tommaso pointed out that one of the propositions, whose sentence contains the expression *express concern*, is imperfect. My system extracted ⟨*INDIGENOUS PEOPLE*, *expressed*, *concern over REDD*⟩. However, what's interested as a predicate for a researcher is the notion of *concern*, not the notion of *expressing*.

My comment was that, indeed, *express concern* contains a **complex predicate**. The semantic weight of the verb phrase is not carried by the verb, but by one of its complements. The system could be improved so that such predicates are treated appropriately. For now, a workaround would be to search for *express* in the `Action` box and *concern* in the `Points` box. Tommaso applied this workaround in query 10 below.

D.2.2.10 Query 10

QUERY: *express* in `Actions`, *concern* in `Points`.
RESULTS: focused on the keyphrases displayed in the `KeyPhrase` tab for the propositions' messages

DISCUSSION:

Topic 1

This query was to test the workaround I suggested above to get issues people have expressed concern over, given that *concern* is not currently extracted as a predicate.

Tommaso finds that the ***flexibility mechanism*** keyphrase is interesting, since the fact that people are concerned over this shows that the issue is controversial. This issue is known to him as controversial, since some countries were concerned that allowing for what was known as *flexibility mechanisms* may weaken certain countries' efforts to reduce emissions.

Topic 2

Tommaso also pointed out that a **more detailed aggregation of the data** would be helpful. For instance, which actor voices the most concern? (I.e. which is the one that occurs most often with the expression *express concern*). Now we would have to copy the propositions page by page to a spreadsheet and get the counts that way.

He pointed out that additional aggregations would be useful. Not just the propositions, but also some more counts, so that we know how many times each actor used which predicates.

My answer was that it would take time to develop the UI workflow and that this is a UI issue, not an NLP issue. That perhaps the quickest way to get to those counts faster would be to implement an `Export` button to export to CSV the results of any UI query—After all, the UI expresses in HTML the JSON responses for a query provided by the Django web-services in the UI's backend; a CSV response could also be created.

D.2.2.11 Query 11

QUERY: *AOSIS* in the `Actors` box.
RESULTS: Looking informally at the distribution of predicates for this actor.
DISCUSSION: Tommaso made this query to support his point that having counts on the UI for actor-predicate combinations would be useful. For instance, what does AOSIS mostly do? Looking down the proposition list, we see that they *welcome*, they *voice*, they *urge* a lot . . . But it would be useful to have the counts.

My answer was as above. For now one UI function that can help is to sort by the `Action` column. A solution to develop in the short-term solution could be an `export` button that returns results in delimited format—the expert would then count the results with spreadsheet software based on the export.

D.2.2.12 Query 12

QUERY: `AgreeDisagree` tab, with actors *China* and *The US*, and relation type `Agree`.
RESULTS: Focused on the keyphrases.
DISCUSSION: Several comments took place.

Topic 1

The first comment was about the following sentence:

> Many parties noted agreement on the continuation of the Convention principles, with: CHINA stressing CBDR; BARBADOS and NORWAY highlighting the precautionary principle; and the US suggesting that principles need to evolve to reflect changing circumstances and capabilities. [enb12561e.html-66]

Tommaso finds that it's an error to extract that China and the US agree over the issues mentioned in the sentence. If the US stresses that "principles need to evolve", this is a sign that China and the US are actually disagreeing over the Convention principles.

My answer was that the disagreement, if it exists, is subtle, and not detectable at the level of modeling we're using. Our tool detects actors in certain positions, relative to an agreement or disagreement predicate, and the sentence was modeled accordingly. Tommaso agreed that the disagreement is subtle.

Regarding the keyphrases displayed in the `KeyPhrase` tab, Tommaso's comment was that some keyphrases are contained in longer keyphrases. He does not think this is a big problem, but asked why this happens. The answer is that keyphrases have not been deduplicated to retain only the longest ones.

Topic 2

Tommaso also suggested that we display **counts** for the times a country is in agreement or disagreement with another country. That in order to get a global view, it would be useful to have a different type of aggregation of results, where you see at a glance which actors agree with each other, which are the "enemies" or "allies" of an actor (i.e. the ones that disagree or agree with that actor).

I also think that those counts would be very useful, and this had already been identified as future work, it had also been pointed out by someone who had looked at the interface informally.

Topic 3

Tommaso suggested that, since now we have this agreement-disagreement information available, the data be used to create two **country networks**, the agreement network and the disagreement one. To see which countries agree or disagree with which most often.

My answer is that that is exactly one of the applications we had in mind when we implemented this tool: being able to keep creating networks, but being able to give an explicit meaning (i.e. agreement or disagreement) to the network's edges, a meaning more informative than "the nodes co-occur".

D.2.2.13 Query 13

QUERY: *vulnerability* in the `Points` box.
RESULTS: Tommaso focused on the 57 propositions extracted.
DISCUSSION: Tommaso made this query since I asked him to make concrete queries to test the UI's results against his own knowledge of the corpus. This was the first query he made in relation with that. There were no comments, we moved to other queries.

D.2.2.14 Query 14

QUERY: *forest* in `Points`.
RESULTS: looking at the 82 propositions extracted.
DISCUSSION: This query was made in response to my request to try look for information from the cooccurrence networks in the Venturini et al., 2014 article, that the expert is an author for. (This request was also made to Kari de Pryck, a co-author of that article, in the previous evaluation session). Tommaso said he sees no problem with the propositions extracted and asked me what I wanted to know regarding this query.

I said that the idea was for him to compare the information the UI gives, compared to the information they used in their article.

Tommaso's comment was the following: In the UI you have access to the evidence regarding each country, including the sentences containing the evidence. It is stronger than the method in the article in terms of the level of detail you can access; this level of detail was missing in the article, the UI provides exactly what was missing in the approach used for that article. However, the UI lacks a bit of results aggregation.

D.2.2.15 Query 15

QUERY: Based on one of the propositions extracted for Query 14, Tommaso was trying to figure out the meaning of the acronym *MRV* using the UI.
RESULTS: Tried to find it in the `ClimTag` tab (Reegle Thesaurus as tagged with the Climatetagger API).[5]
DISCUSSION:

Topic 1

Tommaso looked for the **term's definition** in the concepts in the `ClimTag` tab, since these concepts are a good place to find technical definitions, as they come from a domain-specific thesaurus.

Tommaso did not find the definition, since the acronym itself was not there. However by looking at the concepts extracted, it was possible to find one whose initials match the acronym: *Measurement, reporting and verification*.

Topic 2

Tommaso also focused on a sentence that contains an error, since one of the propositions had not been extracted:

> Bolivia, for the G-77/CHINA, said MRV of support is also being discussed in the ADP and called for: coherence and coordination; clarity on the level of financial support to developing countries; guidance on the third forum of the SCF; and finance for forests. [enb12612e.html-7]

My answer was that this is a sentence with multiple predicates, and that whereas proposition ⟨*Bolivia, said, MRV of support is also being discussed in the ADP*⟩ had been extracted, proposition ⟨*Bolivia, called for, coherence and coordination [. . .] finance for forests*⟩ had not been extracted. Such errors are part of the approx. 30% error rate that was found when assessing the tool against a manually annotated test-set.

D.2.2.16 Query 16

QUERY: *Bolivia* in `Actors`, *forest* in `Points`

RESULTS: Clicked on the first sentence and asked why there are two propositions on the proposition pane

DISCUSSION:

The reason why there are two propositions extracted is that, unlike in Query 15 above, here both of the sentence's propositions were actually extracted. Besides, this sentence is an example how anaphora resolution can help extract relevant information—the anaphora resolution module found that *He* refers to *Bolivia*.

> He cautioned against perverse incentives to cut forests, and said that for many countries forest conservation would be the main way to participate in efforts to stabilize the global climate. [enb12156e.txt-86]

The propositions extracted were ⟨*[He=>Bolivia], cautioned, against perverse incentives to cut forests*⟩ and ⟨*[He=>Bolivia], said, that for many countries forest conservation would be the main way to participate in efforts to stabilize the global climate*⟩.

D.2.2.17 Query 17

QUERY: *Bangladesh* in `Actors`.

RESULTS: Looking at the propositions.

DISCUSSION: This query was used by Tommaso to illustrate his point that aggregating the extractions in a more thorough way would be useful. For instance: *What predicates are used with* Bangladesh*, and how many times each?*

Now it would be possible for an expert to obtain this information from the UI, but it requires manual work on the expert's part, who would need to count him or herself.

D.2.2.18 Query 18

QUERY: *vulnerability* in `Points`.

RESULTS: Looking at the distribution of propositions per year.

DISCUSSION: The reason for looking at this was that I asked Tommaso to look for concrete topics that he has some familiarity with, in order to see if findings from their 2014 article are reproduced.

He said that *vulnerability and impacts* is a topic that should come up later in the negotiations, perhaps with a peak in 2014.

This was not verified with the interface, there are several peaks for the keywords *vulnerability* and *impacts*. Even with the Climate Debate Explorer interface[32] (created by Tommaso and other people) you see several peaks (1997, 1998, 2004, 2007, 2014). So perhaps Tommaso's expectation is not accurate—he did state that his knowledge of the corpus is not very detailed.

D.2.2.19 Query 19

QUERY: *Venezuela* in `Actors` box, and verbs of different types (`oppose`, `support`, `report`) with the related checkboxes.

RESULTS: Looking at number of opposition predicates for each type. The results were: 16 opposition predicates, 18 support ones, 71 report predicates (i.e. neutral reporting).

DISCUSSION: Tommaso's expectation was that Venezuela would have more propositions expressing opposition than for the other predicate types.

This was not confirmed by the numbers on the interface. However, if we include *express concern* and *express alarm* as opposition verbs, the numbers would be compatible with Tommaso's expectation. This suggests that not treating complex predicates can pose limitations (in spite of the small proportion of such predicates in the corpus (see 6.6.4.2)).

D.2.2.20 Query 20

QUERY: *transparency* in `Points`, `oppose` in `Actions` checkboxes.

RESULTS: Looking at propositions and the sentences they were extracted from.

DISCUSSION: Tommaso looked at this since he finds it more interesting to look at opposition verbs; more interesting to look at issues people are disagreeing over. So he decided to look at *transparency* to see if it's a controversial issue or not.

Tommaso finds the results interesting in the sense that they confirm his knowledge of some of the facts in the corpus: One of the first sentences on the UI whose proposition point contains *transparency* is from COP 15 (Copenhagen). The sentence contains an opposition predicate. He finds this interesting since lack of transparency is one of the reasons why the COP was considered to have failed (a text was passed that parties had not looked at previously).

D.2.3 Other comments by the expert

In terms of spontaneous comments, Tommaso said the interface could be interesting for the team that writes the ENB, and that it could be a tool interesting for negotiating parties to use. He suggested either adding an *About* page with explanations how to use the UI's different functions, or a video showing some examples.

The evaluation protocol intended to elicit comments about the UI from experts, and then on the basis of those comments establish whether they are getting an overview on the data, whether they are getting new insight, and whether the quality of the extractions is sufficient, in terms of factual correctness and coverage. Asking the expert explicitly what he/she found useful or not useful was not the original intent.

However, even if this expert repeatedly stated that he found the UI "very cool", he kept pointing out possible improvements, and the imperfections he found were clearer to me than the strengths he saw. For this reason, I asked him explicitly, what he found so "cool" about the tool. The answer was the following:

- The search fields (`Actors`, `Actions`, `Points`) model closely the types of information that a researcher wishing to study this corpus would like to search for, and it is helpful to be able to search for each criterion separately.
- When you already have a question to examine, you find information for it fast (e.g. when you already know what actor you do research on).

What he misses is more aggregation on the information extracted. For instance, if you don't know what actor you're interested in, currently there is no easy access to the full list of actors in the corpus: making the list easily available would be an improvement. Two more concrete examples: if you extract 100 propositions where the actor is Canada, he would be interested in seeing how many times Canada has used what verb. Likewise, if we extract 100 propositions where the predicate introducing the message is complain, he would like to know how many times each actor complained.

This information can be computed by the user from the information that feeds the UI, but the way to do so is currently not convenient for the users (they need to copy-paste the information on a spreadsheet, which is not convenient if there are many pages of information).

Tommaso also stated finding the tool better as a *research* tool than as an *exploration* tool.

D.2.4 Weaknesses pointed out by the expert

Issues that the expert considered weaknesses are documented here. A possible solution and action-items aimed at improving on those issues are listed where relevant.

D.2.4.1 Weakness 1

Description: Tommaso finds it very useful that you can go down to the level of the sentence, and that you know who said what. However, he would find it useful to have a more detailed aggregation of results based on the proposition extraction. Two examples:

- Once you have all propositions where *Bangladesh* is the actor, how many times was each verb mentioned?
- In the `AgreeDisagree` view, it would be useful to list how many times each actor was in agreement or disagreement with another one.

Possible solution: Tommaso suggested adding tabs to the interface where such aggregation takes place. E.g. adding tabs on the right where you get predicate counts for the `ActorView` tab, or actor counts for the `ActionView` tab, and so on.

As an alternative, a faster way I can think of in order to approximate the functionality described by Tommaso would be an `export` button to export query results. Then the user would have to perform whatever aggregations with spreadsheet software or the like.

Action items: Whereas I agree that such aggregations would be useful, no modifications are planned for now in the interest of time.

D.2.4.2 Weakness 2

Description: Predicates like *express concern* are currently treated considering *express* as the predicate and *concern* as part of the message. Moreover, *express* is treated as a neutral reporting verb, whereas in reality *express concern* indicates a negative attitude on the actor's part. This can affect counts for reporting vs. opposition predicates.

Here are some quantitative data that I obtained later about this phenomenon:

- There are 860 propositions where *express* has been extracted as the predicate, for the confidence range 0 to 5.
- Among those 860, there are 475 propositions (55% of the 860) that contain *concern* (400 propositions), *disappointment* (71) or *alarm* (4).
- Those 475 propositions amount to 1.8% of the total 26,465 propositions in the confidence range 0 to 5.

Possible solution: These complex predicates could be treated appropriately by modifying the proposition extraction workflow. A possible modification would be looking for the element that bears semantic weight (*concern* etc.) in the points of propositions whose predicate is *express*. Another option

would be to exploit the NomBank annotations for nouns like *concern*—the SRL module does provide such annotations, but I have not examined their quality.

Action items: No action item is planned so far in the interest of time. A workaround to find propositions with complex predicates is to enter *express* in the `Actions` search-box and *concern*, *disappointment* etc. in the `Points` box.

D.2.5 Comments on expert's use of the UI

This section provides some comments on how the expert approached the evaluation task and his general behaviour with the UI.

Tommaso asked detailed questions about how to use the interface and about the results of many of the queries he ran. A discussion followed many of the 20 queries he ran (see above). Sometimes he focused on a specific element of the results and, based on that, he asked for clarification or for more information about the tool's behaviour.
He also decided to run additional queries as a result of our discussion of some queries.

Note that Tommaso does not have detailed knowledge of this corpus specifically, unlike the expert in Session 1. This is one of the reasons why his way to evaluate the interface did not primarily consist in checking corpus facts or corpus topics known to him with the interface, to see how the UI confirms or complements his knowledge of the corpus. Rather, he used his general knowledge how to address corpora covering controversial issues, as well as how to analyze the way they're covered (which actors are represented, which tone is employed etc.).

Tommaso regularly made comments that pertain to usability aspects of the interface (e.g. the comments regarding result aggregation). I clarified that I needed the evaluation to concentrate on the information provided by the pipeline and displayed on the UI, not so much on usability.

D.3 Session 3

D.3.1 Basic session data 255
D.3.2 Queries run by expert 255
D.3.2.1 Query 1 255
D.3.2.2 Query 2 255
D.3.2.3 Query 3 256
D.3.2.4 Query 4 256
D.3.2.5 Query 5 257
D.3.2.6 Query 6 257
D.3.3 Other comments by the expert 257
D.3.3.1 General comments about the tool 257
D.3.3.2 Comments about the evaluation procedure 258
D.3.3.3 Other comments 258
D.3.4 Weaknesses pointed out by the expert 258
D.3.5 Comments on expert's use of the UI 258

D.3.1 Basic session data

Expert: Nicole de Paula is a Writer/Editor at the Earth Negotiations Bulletin (i.e. the publication that authors the text analyzed on the interface). She holds a Ph.D. in Political Science/International Relations from Sciences Po Paris and is a non-resident fellow at the Center for Transatlantic Relations (CTR) at Johns Hopkins University (SAIS), Washington D.C. Her research focuses on global environmental governance (climate change and biodiversity), multilateral trade negotiations, international organizations and foreign policies of Brazil, European Union and United States.

Time and place: August 4, 2016 on Skype, using the UI's public address.[10] (I also had access to a local version, which was equivalent to the online one in terms of the functions looked at by the expert). It was an audio call and I shared my screen with the expert.

Duration: Around 1 hour.

UI Versions: Like in Session 2, the UI version was commit *c1d36ce* (online) and commit *d1fc766* (local). The differences between these versions and the versions used in Session 1 are minor "cosmetic" fixes that cannot be expected to have a major impact on the UI's usefulness.

Incidences: There were no incidences resulting from the tool itself. However, doing the session over Skype made communication more difficult than doing it in person. The expert pointed out that the resolution of the screen shared with her over Skype was too low for comfortable reading, I had to zoom in to show her the example queries. She had normal access to the UI at the UI's public address.[10]

Note that being prepared for using a second screen sharing service in case image quality is not good enough with Skype could be helpful.

D.3.2 Queries run by expert

The discussion for queries 1 through 3 is shared for the three queries and reported with Query 3.

D.3.2.1 Query 1

QUERY: *health* in `Points` box and *Brazil* in `Actors` box.
RESULTS: The expert wanted to look at the propositions. Two results were extracted.
DISCUSSION: see Query 3.

D.3.2.2 Query 2

QUERY: *health* in `Points` box and *U.S.* in `Actors` box.

[10] https://apps.lattice.cnrs.fr/ie/uidev/

RESULTS: The expert wanted to look at the propositions. Two results were extracted.
DISCUSSION: see Query 3.

D.3.2.3 Query 3

QUERY: *health* in `Points` box and *Canada* in `Actors` box.
RESULTS: The expert wanted to look at the propositions. No results were extracted.
DISCUSSION FOR QUERIES 1 THROUGH 3:

Topic 1: General use of the UI

Nicole's first comment was that there did not seem to be a lot of results.

She explained that she was trying to see if health had been included in the discussions somehow.

She stated that she did not have a specific urge to know anything at the moment, so it was not easy for her to look for information out of a real research context.

She mentioned that, nevertheless, she could see that, if she wanted to see the countries that are more involved with health issues, she could use the UI to get some indication about that. That the UI can be useful if you're doing research and want to know what a specific actor was saying about a topic.

She also mentioned that the years could be useful, in this sense: If you know that in a specific year of the negotiation something was problematic, you can adjust the time range to look at that.

Topic 2: Sorting by year

I pointed out that you can sort by year, which shows that most messages about health come from 2014.

Nicole's comment was that this is interesting, because health is an issue she's working on and that fact gives her an idea when the debate on health gained force, and you can see it's a very recent date.

Topic 3: Obtaining a larger result-set

I suggested that we could get more results by allowing for lower confidences and by searching in the complete sentence, not just the proposition points.

This gave rise to new queries, below.

D.3.2.4 Query 4

QUERY: *health* in `Free Text` box.
RESULTS: Nicole looked at the propositions extracted.
DISCUSSION:

Nicole's first comment for the `Free Text` result was that the first three propositions are not very connected to the issue of health. You see *health* on the sentences on the right, but not in the proposition point, so it's good that at least you have the sentence to see the broader context (see Figure D.1).

FIGURE D.1 – Query for *health* in `Free Text` box

My response was that this is to be expected since we are using the whole sentence now as the search scope, not just the negotiation point.

I pointed out that the sentence from which each proposition has been extracted can be isolated on the right panel by clicking each proposition when the `Sentence` tab is active on the right. Similarly, if you activate the `Docs` tab, then clicking a proposition highlights and scrolls into view the sentence containing the proposition, inside its document.

Nicole said that it's useful to have the whole sentence, and that it's very useful to see the sentence in the document, to have more context.

D.3.2.5 Query 5

QUERY: *European Union* in `Actors`.
RESULTS: Looking at the 943 propositions for that actor.
DISCUSSION:

Nicole's first comment was that there are a lot of results.

I mentioned that a first way to look at the results more systematically would be to sort by *Year* or by *COP*. To restrict results, she decided instead to filter for a specific year, 2009. See the next query.

D.3.2.6 Query 6

QUERY: *European Union* in `Actors`, filtered to year range 2009–2009.
RESULTS: Looking at the propositions.
DISCUSSION:

Nicole points out that she's choosing this year because she knows that it was a difficult year (the Copenhagen COP).

She says that you see some of the disagreement coming through, you see the tension, given propositions like ⟨*European Union, concluding, why a legally-binding agreement had been omitted from the text*⟩ or ⟨*European Union, expressed, concern that this rewards countries that overshoot first commitment period targets*⟩.

She said that the tool can help researchers who know COPs recall more easily the way the negotiation developed.

D.3.3 Other comments by the expert

This section contains comments that the expert did not make in response to only one of the queries she performed, but as general reflections.

D.3.3.1 General comments about the tool

(a) Nicole said that regardless of the queries to perform on the UI, an example of a practical application of a tool like this is to find countries who support an issue, for practical purposes like applying for funding with them, for initiatives regarding that issue. Some example issues she mentioned were biodiversity or health.

(b) Another practical application she can think of is for academic work; it can help you find evidence more easily.

(c) Nicole also mentioned that if she had to do research, she would download the full documents (COP summaries) and read them.

When I asked her why she would not use other tools, she said that she did not know about tools like this interface or the médialab's interface[32] before, so she is not used to other tools.

Nicole agreed however that if you want to look for information about a specific actor it is better if you have a tool like this one, that allows performing that type of search directly (better than reading all the documents and looking for actor-related information manually).

D.3.3.2 Comments about the evaluation procedure

Nicole stated finding the task a bit artificial. She said that it would be more natural for her to use the interface if she were in the middle of a research and had questions about which she wanted to find information.

My response to this is that the other experts were not feeling that way. Perhaps they are more used to this type of evaluation where the expert makes up a research scenario even if they don't need the results for their own work.

Nicole also mentioned that she's not used to employing this type of tool, so she finds it hard to point out what's missing from a type of tool she does not normally use.

D.3.3.3 Other comments

(a) When shown the example query for *China/US* `agreement` vs. `disagreement`, she pointed out that you can see that the amount of disagreement is higher. She agreed with my comment that this was to be expected for this country pair.

(b) When I explained to her that lower confidence scores allow propositions containing less typical actors, like *Indigenous Peoples Group*, or *Women and Gender* to appear in the propositions, she said that this is very useful, since it is precisely these types of actors (not countries or country groupings) that tend to push for new negotiation themes. Countries, conversely, do not always introduce interesting changes in the agenda.

D.3.4 Weaknesses pointed out by the expert

This section refers to issues that the expert considered weaknesses.

She finds the tool to be applicable for a very specialized audience only (e.g. people who already carry out research on climate policy). Because in her opinion, a more general audience is not going to care enough about this type of information to interact with it.

My response to this is that, since the target audience is indeed domain experts, this need not be seen as a weakness.

D.3.5 Comments on expert's use of the UI

The expert stated that she found the evaluation task somewhat artificial in the sense that she was not using the UI to look for information as part of her current research. This may be a resaon why she made less queries than the other domain-experts.

Appendix E

Publications Related to the Thesis

International Peer-Reviewed Journals

Lauscher, Anne, Federico Nanni, **Pablo Ruiz Fabo**, and Simone Ponzetto. (2016). Entities as Topic Labels: Combining Entity Linking and Labeled LDA to Improve Topic Interpretability and Evaluability. *IJCoL, Italian Journal of Computational Linguistics, 2(2)*: 67-87. *Special Issue on Digital Humanities and Computational Linguistics.* http://www.ai-lc.it/IJCoL/v2n2/4-lauscher_et_al.pdf

International Peer-Reviewed Conference Proceedings

Main Session

Ruiz Fabo, Pablo, Clément Plancq, and Thierry Poibeau. (2016). More than Word Cooccurrence: Exploring Support and Opposition in International Climate Negotiations with Semantic Parsing. In *Proceedings of LREC, Tenth International Conference on Language Resources and Evaluation*, pp. 1902-1907. European Language Resources Association (ELRA). Portorož, Slovenia. http://www.lrec-conf.org/proceedings/lrec2016/pdf/636_Paper.pdf

Ruiz Fabo, Pablo and Thierry Poibeau (2015). Combining Open Source Annotators for Entity Linking through Weighted Voting. In *Proceedings of *SEM 2015. Fourth Joint Conference on Lexical and Computational Semantics.* Association for Computational Linguistics. Denver, U.S. http://www.aclweb.org/anthology/S15-1025.pdf

Demonstrations and Workshops

Ruiz Fabo, Pablo, Thierry Poibeau, and Frédérique Mélanie (2015). ELCO3: Entity Linking with Corpus Coherence Combining Open Source Annotators. In *Proceedings of NAACL 2015: Demonstrations*. Denver, U.S. http://www.aclweb.org/anthology/N15-3010.pdf

Ruiz Fabo, Pablo, and Thierry Poibeau (2015). EL92: Entity Linking Combining Open Source Annotators via Weighted Voting. In *Proceedings of SemEval 2015, the 9th International Workshop on Semantic Evaluation*, 355-359. Denver, U.S. http://www.aclweb.org/anthology/S15-2060.pdf

[continues overleaf]

International Peer-Reviewed Workshop Coordination

Poibeau, Thierry, Melissa Terras, **Pablo Ruiz Fabo**, Steven Gray, Glenn Roe. (2015). Workshop Visualizing Data for Digital humanities: Producing Semantic Maps with Information extracted from Corpora and other Media. In *Digital Humanities Conference (DH 2015)*. https://hal.archives-ouvertes.fr/hal-01173965

International Peer-Reviewed Conference Abstracts

Nanni, Federico and **Pablo Ruiz Fabo**. (2016). Entities as topic labels: Improving topic interpretability and evaluability combining Entity Linking and Labeled LDA. In *Digital Humanities Conference (DH 2016)*. Kraków, Poland. https://hal.archives-ouvertes.fr/hal-01483336

Tieberghien, Estelle, Frédérique Mélanie, **Pablo Ruiz Fabo**, Thierry Poibeau, Tim Causer and Melissa Terras (2016). Mapping the Bentham Corpus. In *Digital Humanities Conference (DH 2016)*. Kraków, Poland. https://hal.archives-ouvertes.fr/hal-01378029

Ruiz Fabo, Pablo, Clément Plancq, and Thierry Poibeau. (2016). Climate Negotiation Analysis. In *Digital Humanities Conference (DH 2016)*. Kraków, Poland. https://hal.archives-ouvertes.fr/hal-01423299

Poibeau, Thierry and **Pablo Ruiz Fabo**. (2015). Generating Navigable Semantic Maps from Social Sciences Corpora. In *Digital Humanities Conference (DH 2015)*. Sydney, Australia. https://hal.archives-ouvertes.fr/hal-01173963

Navigation en corpus fondée sur les concepts et les relations : Applications du Traitement automatique des langues aux Humanités numériques

Résumé en français de la thèse

1 Technologies TAL pour la navigation en corpus 264
2 Étude de cas : Les manuscrits de Jeremy Bentham 265
2.1 Description du corpus 266
2.2 Travaux précédents sur le corpus 269
2.3 Cartographie du corpus 271
2.3.1 Résolution référentielle des mentions et extraction de termes 272
2.3.2 Regroupement de termes par proximité sémantique 276
2.3.3 Visualisation du réseau 277
2.4 Interface de navigation 277
2.5 Évaluation par des experts du domaine 283
2.6 Conclusions et perspectives 287
3 Étude de cas : Le *Bulletin des Négociations de la Terre (ENB)* 288
3.1 Description du corpus 288
3.2 Travaux précédents sur le corpus 289
3.3 Extraction de relations 290
3.3.1 Chaîne de traitements TAL 291
3.3.2 Modèle du domaine 293
3.3.3 Règles d'extraction 294
3.4 Interface de navigation 296
3.5 Évaluation par les experts du domaine 299
3.6 Conclusions et perspectives 301
4 Conclusions 301

1 Technologies TAL pour la navigation en corpus

Ce résumé a pour but de permettre à des personnes non anglophones de comprendre les aspects essentiels de la thèse.

La recherche en Sciences humaines et sociales (SHS) repose souvent sur de grandes masses de données textuelles, qu'il serait impossible de lire en détail. Le Traitement automatique des langues (TAL) peut repérer des informations pertinentes parmi cette masse textuelle pour fournir des aperçus utiles aux experts du domaine en question.

Dans cette thèse, nous avons appliqué des technologies de TAL qui permettent d'annoter les acteurs et les concepts majeurs dans un corpus. Les relations entre ces acteurs et les concepts sont ensuite identifiées, également grâce à des technologies de TAL. Un des enjeux de la thèse était d'aller au delà des aperçus simplement fondés sur la cooccurrence de termes (ex. les cartes réseaux), pour fournir des analyses où la nature de la relation entre ces termes est catégorisée sur une base linguistique. Par exemple, deux acteurs mentionnés dans le corpus sont-ils d'accord sur un sujet donné, ou sont-ils plutôt en désaccord ?

Trois études de cas ont été menées, sur des corpus différents. Pour ce résumé, nous avons sélectionné deux études de cas, qui illustrent les enjeux de la thèse et les points forts ainsi que les limites des approches choisies. Le troisième cas (corpus PoliInformatics) est laissé de côté car il met en avant des techniques d'analyse relativement similaires aux deux cas dont il sera question ici.

La première étude de cas dans ce résumé (section 2) correspond à un corpus inédit de manuscrits de Jeremy Bentham (1748–1832), le philosophe et réformateur anglais. Ce corpus aborde plusieurs sujets en éthique, philosophie morale, politique, etc. et il est trop vaste pour être lu exhaustivement par un expert. Il faut donc identifier les concepts majeurs du corpus et nous avons utilisé deux technologies complémentaires pour cela. La première technologie est le liage d'entités (plus connu sous le nom d'*Entity Linking*). La deuxième technologie est l'extraction de concepts, c'est-à-dire en général des mots clés (Keyphrase extraction). Les concepts ainsi identifiés sont ensuite visualisés sous forme de cartes-réseaux interactives, qui montrent des ensembles thématiques dans lesquels les concepts du corpus s'articulent,ainsi que l'évolution de ces ensembles au cours du temps.

La deuxième étude de cas (section 3) consiste en une analyse de textes du *Bulletin des Négociations de la Terre*, appelé *Earth Negotiations Bulletin* (*ENB*) dans sa version anglophone (c'est la version anglophone qui a été étudiée dans la thèse. Le 12^{e} volume du *ENB* contient des rapports détaillant les

interventions de chaque pays participant aux sommets internationaux sur la politique climatique, appelés *COP* (Conference of the Parties), comme la COP 21, qui a eu lieu à Paris fin 2015. Il s'agit d'un corpus de négociations politiques. Par conséquent, il est pertinent de connaître non seulement quels sujets ont été abordés lors des négociations, mais aussi quel participant a abordé quel sujet, et avec quelle attitude, par exemple soutien ou opposant à une mesure donnée. Des techniques d'extraction automatique des relations à base sémantique et syntaxique ont été appliquées au corpus pour déterminer la position des différents participants sur les sujets discutés dans les négociations. Tous les renseignements extraits peuvent être consultés via une interface utilisateur, qui permet d'effectuer des recherches par participant, par sujet, ou par relation (c'est-à-dire support, opposition, accord ou désaccord).

Les applications de navigation en corpus ont été évaluées par des experts du domaine, avec des résultats satisfaisants, comme il sera détaillé ci-dessous. L'interface liée au corpus ENB a en particulier été jugée par les experts comme ayant un potentiel immédiat, permettant de dessiner des pistes de recherche nouvelles et un intérêt applicatif évident.

Dans la thèse, un état de l'art a été dressé pour les technologies de base exploitées, à savoir le liage d'entités et les méthodes d'extraction de relations entre entités (cf. la partie I de la thèse). Les développements effectués autour de ces technologies pour mieux adapter les outils aux corpus visés ont également été décrits en détail (partie II), en fournissant une évaluation intrinsèque (c'est-à-dire quantitative, par rapport à un référentiel élaboré à la main). Dans ce résumé, ces éléments ont été omis.

Les deux études de cas citées ci-dessus sont exposées dans les sections suivantes. Les corpus analysés et les applications développées seront décrits. Un descriptif plus détaillé est disponible dans la thèse complète mais les aspects essentiels de l'état de l'art et l'intérêt pour les Humanités numériques des technologies appliquées seront évoqués ici, ainsi qu'une description des technologies suffisante pour comprendre les applications développées.

2 Étude de cas : Les manuscrits de Jeremy Bentham

Cette section présente notre application pour naviguer dans les manuscrits de Jeremy Bentham, ainsi que les principales technologies de TAL utilisées. D'abord, nous donnons des renseignements sur l'effort de création de corpus mené par University College London, qui nous a fourni les versions numériques du texte des manuscrits. En 2.1, le corpus est décrit, avec des détails sur l'échantillon que nous avons sélectionné pour nos analyses. Nous donnons ensuite en 2.2 un aperçu des outils préexistants pour naviguer

dans ce corpus et des travaux précédents portant sur l'analyse du contenu des transcriptions. En 2.3, la procédure de cartographie du corpus est présentée, incluant la détection de concepts basés sur l'Entity Linking et sur l'extraction de termes, ainsi que la création de cartes réseaux sur la base de ces extractions. Notre interface utilisateur pour naviguer dans le corpus à travers des cartes interactives et un index de recherche est présentée en 2.4. Enfin, l'évaluation de l'application par des experts du domaine est décrite dans 2.5.

2.1 Description du corpus

Le corpus, ainsi que l'échantillon que nous avons analysé, sont décrits ici.

2.1.1 Caractéristiques du corpus

Jeremy Bentham (1748–1832) était un philosophe et réformateur anglais, connu comme le fondateur de l'utilitarisme, une doctrine philosophique qui propose que la valeur éthique d'une action est d'autant plus grande qu'elle favorise le plus grand bonheur du plus grand nombre.[1] Bentham a écrit sur une grande variété de sujets, incluant l'économie politique, la religion et la moralité sexuelle. Certaines des idées de Bentham auraient pu tomber sous le coup de la loi à son époque, et ce type de contenu est resté inédit de son vivant et longtemps après encore. Bentham a produit plus de 60 000 folios de manuscrits, inédits ou non, grâce auxquels nous pouvons connaître son point de vue sur une énorme quantité de sujets très différents. Le *Bentham Project*,[2] de University College London (UCL) vise produire une nouvelle édition des œuvres complètes de Bentham (Bentham, 1968 – ongoing), tenant compte de l'intégralité des manuscrits. Les chercheurs du Bentham Project ont initialement transcrit une partie des manuscrits et catalogué le corpus, en ajoutant des métadonnées comme la date d'écriture présumée, le type de document et autres (voir ci-dessous). Depuis 2010, les manuscrits sont en train d'être numérisés et transcrits par des bénévoles grâce à une initiative de *crowdsourcing*[3] appelée *Transcribe Bentham* (Causer et al., 2014b), également coordonnée par UCL. Nous avons eu accès à un grand nombre de ces transcriptions dans le cadre de notre collaboration avec UCL, et l'interface utilisateur que nous avons développée permet de naviguer dans un sous-ensemble de ces transcriptions.

[1] *The greatest happiness of the greatest number* dans la formulation anglaise.
[2] https://www.ucl.ac.uk/Bentham-Project
[3] Parfois appelé en français *myriadisation*. Le crowdsourcing consiste au partage collaboratif d'une tâche en ligne par un grand nombre de participants.

Le corpus est encodé en format XML-TEI.[4] Les chercheurs du Bentham Project ont parcouru chaque folio pour déterminer le type de chaque document. La vaste majorité du corpus consiste en des versions temporaires ou brouillons de travaux en cours de Bentham. L'intérêt de ces brouillons (appelés *text sheet*) est important quand on sait que Bentham a généralement détruit les versions temporaires des textes publiés. Par conséquent, les brouillons conservés représentent des œuvres non publiées, ou des textes écartés des versions publiées de ses travaux. À part les brouillons, le corpus contient un petit ensemble de documents prêts à la publication (les *fair copies*), des lettres reçues ou envoyées par Bentham, et des *collectanea*, c'est-à-dire des textes copiés par les assistants de Bentham à partir de journaux ou d'autres sources, afin qu'il puisse les citer.

Outre le type de document, les chercheurs du Bentham Project ont produit plusieurs métadonnées pour chaque folio.[5] Il s'agit entre autres de la date réelle ou estimée de la composition du manuscrit, des titres et sous-titres de celui-ci et, pour la correspondance, de l'expéditeur et du destinataire.

Concernant les renseignements encodés dans le corpus à travers le balisage TEI, ce sont tout d'abord des informations sur le processus d'écriture, comme les ajouts et les suppressions (texte barré) qui ont été notés. Des éléments structurels comme les titres et notes marginales sont également annotés. D'autres balises TEI ont été utilisées pour indiquer des passages en langue étrangère ou le texte illisible.[6] Nous n'avons pas exploité cette information dans nos analyses, mais il serait utile de le faire, afin de pouvoir par exemple restreindre les recherches uniquement aux passages ajoutés ou supprimés, pour mieux comprendre le processus éditorial suivi par Bentham. La figure 1 montre un exemple de manuscrit et les informations annotées dans sa transcription TEI, à travers l'affichage HTML.

La plupart des documents du corpus sont en anglais, mais certains sont en français ou contiennent de longs passages en latin.

D'après le site Web du projet, 17 513 folios avaient été transcrits en janvier 2017.[7] Chaque folio est divisé dans plusieurs pages, et chaque fichier XML-TEI correspond à une page.

[4] Le format TEI est décrit sur http://www.tei-c.org/index.xml

[5] Ces métadonnées sont décrites en détail sur http://www.benthampapers.ucl.ac.uk/search.aspx?formtype=advanced.

[6] Causer et al. (2012, p. 123) ainsi que le guide d'annotation du projet Bentham (http://www.transcribe-bentham.da.ulcc.ac.uk/td/Help:Transcription_Input_Form#Core_Guidelines) donnent plus de détails.

[7] Le site de Transcribe Bentham montre l'état d'avancement courant : http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham

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'.

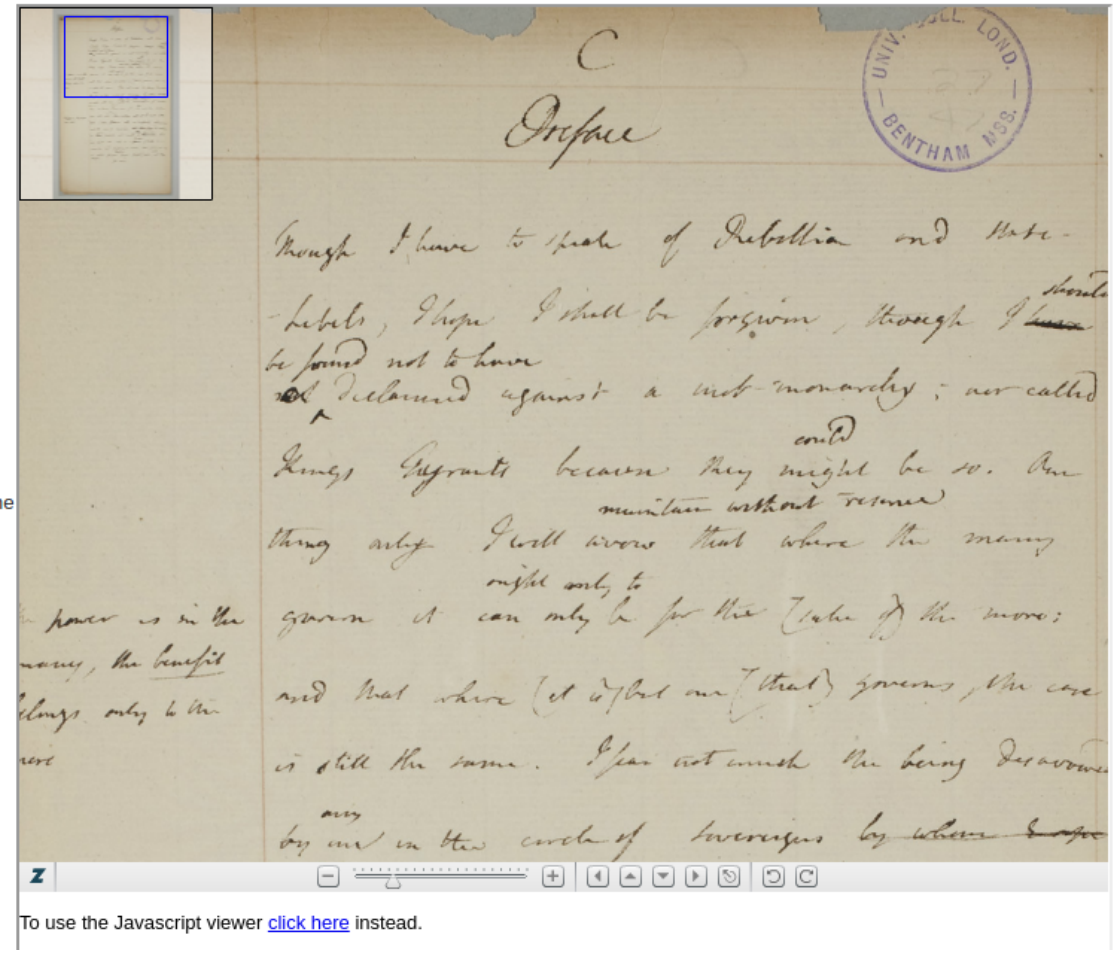


FIGURE 1 – Interface du projet Transcribe Bentham d'UCL. À droite on voit un manuscript numérisé par UCL, avec plusieurs éléments caractéristiques encodées en TEI par les transcripteurs bénévoles du projet, comme des ajouts et suppressions de texte et des notes marginales. Le panneau de gauche montre le rendu HTML de ces annotations TEI : le texte omis est barré en HTML, les ajouts sont suscrits, et les notes marginales se trouvent dans des boîtes. La capture d'écran correspond à http://www.transcribe-bentham.da.ulcc.ac.uk/td/JB/027/047/001

2.1.2 Échantillon analysé

Dans le cadre d'une collaboration entre le LATTICE et UCLDH (le Centre d'Humanités numériques d'UCL), nous avons eu accès à environ 30 000 pages de texte transcrit, pour effectuer des analyses automatiques. Pour nos analyses, nous n'avons utilisé qu'environ 55% de ces pages, pour les raisons suivantes.

Au début du projet nous n'avions pas accès aux métadonnées décrites à la p. 267, qui incluent les dates des manuscrits. Comme nous voulions analyser l'évolution temporelle du contenu du corpus, il nous était nécessaire d'attribuer une année à chaque fichier. L'heuristique simple que nous avions utilisée à cette fin était de considérer que la première séquence de quatre chiffres dans le document représentait son année de production s'il s'agissait d'une année comprise entre la naissance et décès de Bentham. Les années ainsi estimées sont en forte corrélation ($r = 0,976$)[8] avec les années réelles identifiées par le Bentham Project, auxquelles nous avons eu accès récemment. Cependant, l'heuristique n'était pas applicable à environ 44% des documents que nous avions reçus, car ils ne contenaient aucune séquence de quatre chiffres.

Nous avons également écarté les documents dont la langue principale n'était pas l'anglais, à l'aide d'un outil d'identification de la langue, appelé LINGUA-IDENTIFY, disponible sous forme d'un module Perl.[9] Environ 400 fichiers ont été identifiés comme n'étant pas en anglais.

Après avoir éliminé les fichiers dont notre heuristique n'a pas réussi à trouver l'année, ainsi que les fichiers non anglais, l'échantillon retenu continent 16 618 pages, c'est-à-dire 55,53% des documents qui nous avaient été envoyés. Un effet secondaire de notre heuristique de datation est que notre échantillon contient principalement des documents à partir de 1800, lorsque Bentham a commencé à régulièrement dater ses manuscrits. En conséquence, nos analyses d'évolution du contenu (p. 279) donnent des résultats plus clairs et plus pertinents à partir de 1800.

2.2 Travaux précédents sur le corpus

Avant de donner quelques exemples de travaux basés sur les transcriptions produites par Transcribe Bentham, il convient de rappeler que des années d'effort ont été nécessaires pour la création du corpus par Transcribe Bentham, ses volontaires participant à l'initiative de crowdsourcing, et par le Bentham Project lui-même (dont les chercheurs ont initialement catalogué le

[8] Coefficient r de Pearson tel que défini dans https://docs.scipy.org/doc/numpy-1.10.1/reference/generated/numpy.corrcoef.html.

[9] http://search.cpan.org/~ambs/Lingua-Identify-0.56/lib/Lingua/Identify.pm#langof

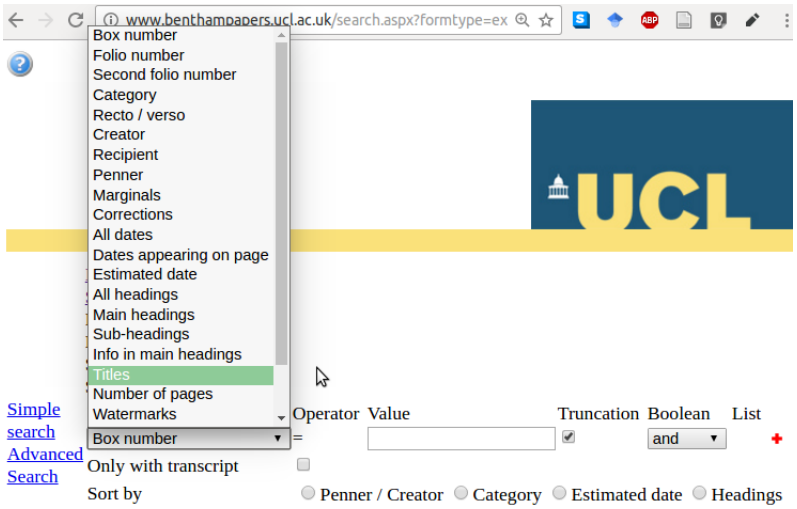


FIGURE 2 – Base de données *Bentham Papers Database* d'UCL, une plate-forme antérieure à Transcribe Bentham, pour la recherche par métadonnées sur les manuscrits de Bentham. La capture d'écran montre l'interface de recherche avec un menu déroulant. Plusieurs champs de métadonnées disponibles sont visibles.
http://www.benthampapers.ucl.ac.uk/

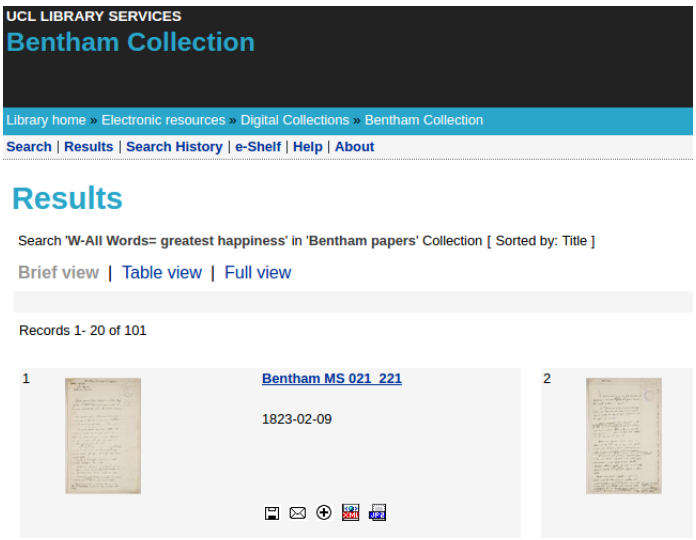


FIGURE 3 – Les bibliothèques d'UCL ont créé une interface qui permet la recherche par métadonnées ainsi que la recherche sur le texte intégral des manuscrits déjà transcrits. L'interface intègre maintenant les transcriptions résultant de l'initiative Transcribe Bentham. En réponse à une requête, l'interface retourne l'image des manuscrits correspondants, ainsi que leur transcription TEI si elle et disponible. La capture d'écran montre le premier résultat pour la requête *greatest happiness* (« le plus grand bonheur »), un concept caractéristique des écrits de Bentham.
https://www.ucl.ac.uk/library/digital-collections/collections/bentham

corpus, et transcrivaient les manuscrits avant la mise en place du crowdsourcing). Plusieurs études analysent ce processus (Causer et al., 2012 ; Causer et al., 2014a ; b). Les problèmes abordés comprennent la plate-forme de transcription, les choix de codage TEI, des aspects méthodologiques et d'évaluation de la qualité. Plusieurs articles discutent aussi de la participation d'un public non-spécialisé dans la valorisation du patrimoine culturel et dans la création de ressources utiles pour le travail de recherche académique.

Outre ces travaux sur la création du corpus, il existe deux plates-formes qui offrent certaines fonctions de navigation dans le corpus Bentham. La base de données *Bentham Papers* permet de rechercher dans les champs de métadonnées (p. 267). Elle retourne la liste de documents dont les métadonnées correspondent à la requête, en fournissant un lien vers la transcription si elle est disponible (figure 2).[10] Les bibliothèques d'UCL ont créé une plate-forme où, en plus des métadonnées, le texte intégral des transcriptions disponibles peut être recherché. Le système retourne alors des liens vers l'image du document, et vers sa transcription TEI si elle est disponible (figure 3).[11]

En ce qui concerne la recherche sur Bentham prenant en compte les transcriptions elles-mêmes, un résultat majeur est la nouvelle édition en cours des œuvres complètes (Bentham, 1968 – ongoing), qui est effectuée par une équipe dirigée par le Professeur Schofield de University College London. Le contenu des transcriptions est exploité comme une source additionnelle pour le commentaire critique et travail d'édition savante. Causer et al. (2014a, section 4) mentionnent plusieurs exemples de contenu inconnu avant la transcription du corpus, dans des domaines fondamentaux de la pensée de Bentham. Le Bentham Project et Transcribe Bentham sont auteurs d'un grand nombre de productions significatives. Dans ce paragraphe je n'ai mentionné que les contributions les plus importantes immédiatement basées sur les transcriptions des manuscrits. Un compte rendu systématique des réalisations de ces projets est disponible sur leurs sites respectifs.[12]

2.3 Cartographie du corpus

Les analyses du corpus décrites ci-dessus impliquent une lecture détaillée des transcriptions, afin de chercher de nouveaux faits qui puissent élargir nos connaissances sur la pensée de Bentham. De leur côté, les technologies que nous avons utilisées ont comme objectif de fournir de nouvelles sources de réflexion, en fournissant un aperçu, sous la forme d'un réseau ou d'une carte, du contenu du corpus. À notre connaissance, il n'existe pas d'analyses

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

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

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

automatiques du corpus Bentham jusqu'à présent, ce qui accroît l'intérêt de l'expérience décrite ici.

Pour créer des cartes de navigation, le corpus est analysé avec des outils de Traitement automatique des langues et de visualisation de graphes, selon trois étapes essentielles : nous procédons tout d'abord à une extraction de termes permettant de modéliser le corpus, à travers les méthodes décrites dans la sous-section 2.3.1. Ensuite, comme expliqué sous 2.3.2, un regroupement de ces expressions est effectué sur la base de leur contexte d'occurrence (c'est-à-dire en ayant recours à un calcul de similarité distributionnelle entre termes). Enfin, étant donné que le processus permet de calculer des distances sémantiques entre termes, le corpus peut être visualisé comme un réseau d'expressions sémantiquement apparentées (grâce à des algorithmes de « spatialisation » qui tiennent compte des distances entre termes). Le réseau ainsi obtenu peut être considéré comme une carte du corpus.

2.3.1 Résolution référentielle des mentions et extraction de termes

La première étape de la cartographie du corpus est l'extraction lexicale. À cette fin, nous avons utilisé deux technologies. La première est le liage d'entités, plus connu sous le nom anglais d'*Entity Linking* (*EL*), terme que nous utiliserons ici. La deuxième technologie que nous avons appliquée est l'extraction de termes (*keyphrase extraction*).

Dans la suite, les technologies seront décrites, ainsi que les outils que nous avons utilisés, leur paramétrage, et la procédure suivie pour sélectionner un ensemble d'expressions permettant de modéliser le corpus sur la base des résultats de chacune de ces deux technologies.

2.3.1.1 Entity Linking

L'Entity Linking (EL) comporte généralement deux étapes. La première consiste à trouver dans un corpus des mentions des concepts répertoriés dans une base de connaissances (BC), c'est-à-dire un référentiel de concepts, qu'il soit spécifique à un domaine ou générique et applicable à plusieurs domaines.[13] Un exemple de base de connaissances générique serait Wikipédia, ou sa version structurée selon les standards du Web sémantique appelée DBpedia.[14] Les mentions (c'est-à-dire les séquences lexicales du corpus) jugées susceptibles de représenter un concept de la BC sont ensuite annotées avec le concept pertinent. Ceci est utile pour relier des passages faisant référence au même concept de la BC, en dépit de la variabilité des expressions utilisées

[13] Cette définition de la tâche d'Entity Linking est parfois appelée *Wikification* dans la littérature ; nous utilisons les deux termes de façon indistincte, comme justifié à la p. 16 de la thèse.

[14] Les formats web sémantique sont des standards publiques facilitant l'échange d'information entre des applications informatiques, en spécifiant un ensemble de classes d'éléments, ainsi que leurs attributs et relations possibles pour modéliser un domaine de connaissances.

pour y référer dans le corpus. Par exemple, les mentions *amount* et *quantity* seront liées (« mappées ») au concept *Quantity* dans DBpedia.[15] On considère qu'une « annotation » correspond au couple formé par un concept et une mention textuelle.

Outil appliqué : Pour l'Entity Linking, nous avons utilisé DBPEDIA SPOTLIGHT (Daiber et al., 2013 ; Mendes et al., 2011). Cet outil emploie DBpedia (Auer et al., 2007) comme base de connaissances et, comme on l'a déjà vu, le contenu de DBpedia est extrait de Wikipédia. Une question qui se pose alors est celle de savoir si Wikipédia, en tant qu'encyclopédie de domaine général créée au 21^{e} siècle, est une source de connaissances pertinente pour analyser des textes spécialisés des 18^{e} et 19^{e} siècles. Il est raisonnable d'anticiper que la couverture des termes spécialisés du corpus sera très imparfaite. Néanmoins, l'étiquetage de corpus avec des concepts DBpedia est un domaine de recherche très actif en TAL, et nous voulions vérifier son applicabilité à ce corpus très particulier en comparaison des recherches généralement menées en TAL.

Les aspects essentiels de l'algorithme de DBPEDIA SPOTLIGHT sont les suivants. D'abord, l'outil identifie les mentions (c'est-à-dire des séquences lexicales susceptibles de faire référence à un concept de DBpedia), ainsi que les concepts DBpedia qui peuvent être considérés comme des candidats-références pour chaque mention. Le repérage de mentions repose sur un dictionnaire pré-défini qui met en correspondance des séquences lexicales avec des pages DBpedia, sur la base d'une liste contenant les titres des pages Wikipédia et les textes des ancres des liens Wikipédia.[16] Ensuite, l'outil compare le contexte autour des mentions dans le corpus avec les « vecteurs contextuels » de chaque candidat-référence. Le vecteur contextuel d'un concept est défini comme la concaténation de tous les paragraphes dans Wikipédia qui contiennent des mentions du concept (Mendes et al., 2011, p. 3). La similitude entre le contexte d'une mention dans le corpus et le vecteur contextuel de chaque candidat-référence est calculée. Pour le calcul, les mots du contexte sont pondérés en fonction de leur pouvoir discriminant pour différencier des candidats (c'est-à-dire que les mots trouvés dans les vecteurs contextuels d'un petit nombre de candidats ont plus de pouvoir discriminant et auront plus de poids dans le calcul que les mots trouvés dans un grand nombre de candidats). Le candidat-référence dont la similarité avec le contexte de la mention est la plus élevée est retenu, si le score de similarité est supérieur à un seuil configurable. Le score de similarité (entre

[15] Tant *amount* que *quantity* peuvent être traduits par *quantité* en français. Les données sur un concept contenu dans DBpedia peuvent être consultées en préfixant avec *http ://dbpedia.org/page/* l'étiquette du concept, par exemple `http://dbpedia.org/page/Quantity` pour le concept *Quantity*

[16] L'ancre est la séquence textuelle qui est affichée pour indiquer une lien, c.à.d. la séquence cliquable.

autres facteurs) est également utilisé pour produire un score de confiance pour l'annotation. Le score de confiance fournit une estimation de la validité de l'annotation.

Sélection des annotations : Une « annotation » dans ce contexte désigne un couple formé par une mention textuelle et le concept DBpedia qui lui a été assigné par le module d'Entity Linking. Seules les annotations dont le score de confiance était supérieure à 0,1 ont été retenues. En outre, nous ne retenons que les annotations ayant au moins 100 occurrences dans le corpus. Ces seuils (confiance de 0,1 et fréquence minimale de 100) ont été déterminés empiriquement, notamment parce que ces chiffres permettaient d'obtenir un nombre d'annotations satisfaisant.

Une liste de 258 annotations a été obtenue avec les seuils ci-dessus. Chaque annotation peut avoir une ou plusieurs variantes textuelles. Par exemple, le concept *Judiciairy*[17] a été assigné par SPOTLIGHT aux mentions *judicatory*, *judicial* et *judicature*, mais le concept *Doctrine* a été utilisé pour annoter les occurrences d'une seule variante textuelle (*Doctrine*).

La première étape après l'obtention de cette liste initiale de 258 annotations a été une **vérification manuelle** des annotations, c'est-à-dire à la fois des mentions et des concepts DBpedia qui leur ont été assignés. Cet examen a révélé plusieurs erreurs. On constate de fréquents **anachronismes**, quand une mention a été annotée avec un concept DBpedia postérieur à l'existence de Bentham. Voici par exemple deux exemples de ce type d'erreur : la mention *quantum* a été annotée comme le concept de physique *Quantum*,[18] et la mention *application*, dans environ 25% des cas, a été annotée comme *Application_software*, c.à.d « logiciel applicatif ». Les anachronismes sont faciles à repérer et à supprimer de la liste des termes avant de créer des cartes du corpus. Certaines **autres erreurs** sont plus difficiles à identifier, car il est nécessaire d'examiner les contextes d'occurrence des mentions dans le corpus pour déterminer qu'il s'agit d'erreurs. Par exemple, la mention *execution* (« exécution ») est utilisée dans le corpus dans le sens de « application d'une décision judiciaire », alors que cette mention a été annoté par le module d'Entity Linking avec le concept DBpedia *Capital_Punishment*, c'est-à-dire « peine capitale ». Si nous avions accepté cette annotation, nous aurions produit une représentation erronée du contenu du corpus, car tous les contextes

[17] Les traductions des expressions dans ce paragraphe sont comme suit : Judiciary (`http://dbpedia.org/page/Judiciary`) correspond à *pouvour judiciaire*. L'adjective *judicial* est traduit par *judiciaire*, et *judicatory* peut être considéré comme un nom pour dire *pouvoir judiciare*. *Doctrine* correspond au même mot en français, c'est-à-dire un ensemble de croyances et de principes traduisant une conception de la société particulière.

[18] Nous rappelons que les concepts DBpedia mentionnées dans la thèse peuvent être trouvés en préfixant `http://dbpedia.org/page/` à l'étiquette du concept, donc ici `http://dbpedia.org/page/Quantum`.

où le mot *execution* apparaît seraient considérés comme des contextes où la peine de mort est discutée, ce qui est bien évidemment erroné.

Pour éviter de telles erreurs, au lieu d'étiqueter les nœuds dans les cartes du corpus avec le concept DBpedia qui représente l'ensemble des variantes textuelles agrégées dans DBpedia, les nœuds ont été étiquetés avec la variante la plus fréquente de cet ensemble.

Outre la modification juste décrite, les résultats originaux de SPOTLIGHT ont été filtrés manuellement pour éliminer les résultats non informatifs. Rappelons que les annotations dont la confiance était inférieure à 0,1 et les variantes dont la fréquence dans le corpus était inférieure à 100 avaient été supprimées automatiquement. Une liste de 258 couples de type ⟨{*ensemble de mentions*}, *étiquette*⟩ avait été ainsi obtenue. Dans ces couples, certains éléments expriment une signification générale qui était peu susceptible de représenter des notions importantes dans le corpus, par exemple, des mentions ou des étiquettes comme *time* (« temps ») ou *place* (« lieu »). Pour cette raison, environ 25 couples ont été filtrés, donnant une liste finale de 258 couples, qui ont ensuite été utilisés pour créer des réseaux de concepts (2.3.2) et des cartes de corpus navigables (2.4.3). L'annexe A dans la thèse affiche la liste finale après le filtrage manuel (p. 210), ainsi que les éléments filtrés (p. 212).

Extraction de termes

Par « extraction de termes » nous faisons référence à la tâche connue sous le nom de « keyphrase extraction » en anglais (Kim et al., 2010 ; Turney, 2000). La tâche consiste à identifier des séquences de mots qui représentent les termes les plus importants dans un texte. La technologie a été utilisée pour l'indexation bibliographique ou l'amélioration de moteurs de recherche. Dans les applications en Humanités numériques, l'extraction de termes est parfois utilisée pour donner un aperçu d'un corpus (par exemple, G. Moretti et al., 2016 ; Rayson, 2008).

Outil appliqué : L'extraction de termes a été effectuée avec YATEA (Aubin et al., 2006), un extracteur à base de règles.[19] YATEA prend en entrée un texte étiqueté avec des catégories grammaticales. Nous avons effectué cet étiquetage avec TREETAGGER (Schmid, 1994).[20] Sur la base d'un ensemble configurable de séquences de catégories grammaticales, YATEA identifie des syntagmes nominaux. L'outil filtre ensuite les syntagmes nominaux obtenus, afin d'éliminer ceux, qui, tout en correspondant à l'un des séquences de catégories grammaticales acceptées, ne sont pas informatifs.

[19] http://search.cpan.org/~thhamon/Lingua-YaTeA/lib/Lingua/YaTeA.pm
[20] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Sélection des motifs : Les termes avec au moins 10 occurrences dans le corpus sont initialement conservés, ce qui fournit une liste d'environ 2550 candidats termes. Cette liste est ensuite filtrée avec des expressions régulières pour éliminer les termes mal formés, par exemples les séquences contenant des signes de ponctuation du fait du mauvais formatage du corpus. D'autres termes sont éliminés car ils sont jugés non informatifs, par exemple des termes contenant des déterminants comme *such* (« tel ») ou *certain* (« certain »). Après avoir appliqué ce filtrage avec des expressions régulières, la liste est enfin filtrée manuellement pour éliminer les termes non pertinents restants. La liste finale contient approximativement 1950 termes, dont les 250 les plus fréquents sont utilisés pour créer des cartes sémantiques. La liste des termes retenus figure en annexe (p. 213).

La fréquence minimale sélectionnée pour les termes (10) est inférieure à celle fixée pour les mentions obtenues par Entity Linking (définie à 100, cf. p. 274). Les termes contiennent généralement plusieurs mots et sont donc généralement moins fréquents que des mots simples. Une fréquence de 10 pour les termes a donc semblé pertinente, même pour les mots simples dans la mesure où un filtrage manuel avait préalablement été effectué.

2.3.2 Regroupement de termes par proximité sémantique

La plate-forme CorText, qui permet l'analyse lexicale et la création de cartes sémantiques a ensuite été utilisée pour l'analyse. Avec CorText, des clusters de termes sont établis selon leur similarité distributionnelle, c'est-à-dire en prenant en compte les mots apparaissant dans le contexte des termes. Nous avons choisi comme contexte une fenêtre de cinq phrases autour du terme. Le score de similarité est calculé avec la mesure définie dans Rule et al. (2015, "Supporting Information", p. 1). C'est une mesure inspirée de Weeds et al. (2005), qui repose sur l'information mutuelle ponctuelle.

Le **réseau est filtré** lors de sa création, pour obtenir des clusters pertinents, en supprimant des liens faibles qui pourraient empêcher des liens plus importants d'être clairement visibles. Le premier filtrage appliqué supprime les liens dont le poids est inférieur à un seuil donné. Ce seuil est calculé avec l'algorithme par défaut de CorText, qui produit un réseau maximalement connecté, en évitant des composants déconnectés (Rule et al., 2015, "Supporting Information", p. 2). Une autre type de filtrage consiste à ne retenir, pour chaque nœud, que les voisins les plus fortement connectés, selon la mesure de similarité mentionnée dans le paragraphe précédant. Le nombre de voisins les mieux connectés retenus a été fixé à 10 pour tous les réseaux.

Des **communautés sont calculées** sur le réseau, c'est-à-dire des groupes de nœuds fortement interconnectés. L'algorithme *Louvain* (Blondel et al., 2008)

a été utilisé. Dans la visualisation, les nœuds sont colorés selon leur communauté et les **communautés sont étiquetées** en utilisant les noms de leurs deux nœuds les plus centraux (c'est-à-dire les nœuds qui reçoivent le plus de liens à partir d'autres nœuds dans la même communauté). L'algorithme d'étiquetage vise à sélectionner des étiquettes qui capturent les principaux thèmes représentés par les éléments lexicaux de la communauté. Un exemple montrant les communautés et une légende avec leurs étiquettes est dans la figure 5.

Des réseaux qui permettent d'examiner l'**évolution temporelle** du corpus peuvent également être créés avec CorText. Nous avons utilisé la fonction *Heatmap* à cette fin. Les *heatmaps* montrent les zones du réseau sur lesquelles le contenu du corpus se concentre, à partir d'une « périodisation » (c'est-à-dire un découpage en tranches temporelles généralement de taille fixe) prédéfinie. Nous avons divisé le corpus en décennies, et utilisé la statistique χ^2 comme mesure de surreprésentation de certains termes par période. Des exemples de heatmaps se trouvent dans la figure 7.

2.3.3 Visualisation du réseau

La plate-forme CorText utilise un algorithme de « spatialisation » fondé sur les forces (*force-based layout*). Ce type de spatialisation simule un système physique où des forces répulsives éloignent les nœuds les uns des autres (comme s'ils étaient des particules chargées), tandis que des forces d'attraction exercées par les liens resserrent les nœuds ensemble (comme un ressort), jusqu'à ce que les forces se stabilisent (Jacomy et al., 2014).

Étant donné que les liens du réseau encodent une similarité sémantique, les nœuds qui sont les plus proches dans le réseau sont reliés thématiquement, partageant des contextes communs. Les nœuds qui possèdent des liens avec d'autres nœuds situés dans deux clusters différents représentent des concepts reliés aux thèmes des deux clusters à la fois.

La taille du nœud dans le réseau est en fonction du nombre de ses cooccurrents. Les communautés (p. 276) sont représentées par des couleurs.

2.4 Interface de navigation

L'interface de navigation donne accès à notre échantillon du corpus *Transcribe Bentham* (p. 266) via une recherche sur le texte intégral et via des versions navigables des réseaux construits à partir du corpus comme décrit ci-dessus (2.3.3). Les *heatmap* montrent en outre l'évolution thématique du corpus au cours du temps. La première exigence demandée aux réseaux est de refléter le contenu du corpus sans introduire d'erreurs évidentes. On espère aussi que l'aperçu du corpus fourni par le réseau, ou les connexions entre les

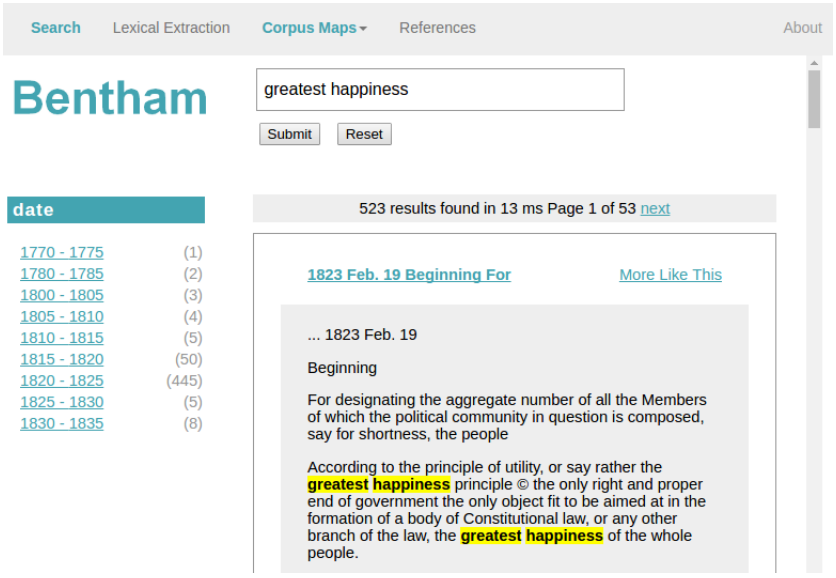


FIGURE 4 – Structure de notre interface utilisateur pour naviguer dans les transcriptions des manuscrits de Bentham en utilisant des réseaux conceptuels et la recherche en texte intégral. La capture d'écran affiche l'index de recherche, en montrant les résultats pour la requête *greatest happiness* (« plus grand bonheur »). Des facettes par date sont disponibles à gauche, pour filtrer les résultats par période de 5 ans ; le nombre entre parenthèses indique les documents renvoyés par période. Les cartes de concepts sont accessibles à partir du menu `Corpus Maps`.

concepts du réseaux, puissent suggérer des nouvelles idées de recherche aux experts (voir l'évaluation de l'interface utilisateur à la p. 285 pour une discussion sur ce point).

Notre interface utilisateur complémente les plates-formes existantes pour naviguer dans le corpus, mentionnées à la p. 269 : la base de données Bentham Papers (figure 2) et la plate-forme des bibliothèques d'UCL (figure 3). La base de données Bentham Papers offre une recherche détaillée fondée sur les métadonnées, et tant l'outil développé par les bibliothèques d'UCL que notre interface permet la recherche dans le texte intégral des transcriptions. Outre le texte transcrit correspondant à la requête, notre interface indique en plus les termes recherchées (en les mettant en surbrillance dans le texte) et avec un regroupement des résultats par date. Notre interface innove surtout par la production de réseaux navigables de concepts, qui n'étaient pas disponibles jusque là avec les autres outils.

2.4.1 Structure de l'interface

La vue d'accueil de l'interface montre l'index de recherche, comme on voit sur la figure 4. Le menu `Corpus Maps` (« Cartes du corpus ») permet d'accéder aux réseaux de concepts navigables et aux heatmaps. Le menu `Lexical Extraction` (« Extraction lexicale ») fournit des informations à l'utilisateur sur la création des listes de termes utilisées pour modéliser le corpus (cf. 2.3.1). Des informations sur les types de cartes créées et leur usage sont

disponibles sur la page `Introduction` du menu `Corpus Maps`. Finalement, le menu `Search` permet de retourner sur la vue d'accueil (c.à.d. afficher l'index de recherche). Les paragraphes suivants décrivent l'interface de recherche et chaque type de carte créée pour le corpus.

2.4.2 Index de recherche

Le corpus a été indexé dans un serveur de recherche SOLR, qui est basé sur le moteur LUCENE.[21] C'est un outil de recherche largement utilisé, qui permet des requêtes HTTP pour l'indexation et la recherche d'information. Les résultats sont triés selon des scores de pertinence calculés par une méthode classique, en utilisant un modèle vectoriel avec des poids *tf-idf* (cf. Manning et al., 2008, chap. 6 pour une description de ce modèle et des poids *tf-idf*). L'outil retourne les documents correspondant à une requête, avec les termes recherchés en surbrillance. Les résultats sont également regroupés par date (tranches de 5 ans dans notre configuration). Les fonctions sont visibles sur la figure 4.

2.4.3 Cartes navigables du corpus

La plate-forme CORTEXT exporte les réseaux créés dans un format standard pour représenter des graphes, appelé *GEXF*.[22] Les exports GEXF des réseaux ont été utilisés pour rendre les réseaux « navigables » avec deux outils différents : le plugin d'export Sigma JS de Gephi et l'explorateur de cartes (*Project Explorer*) de la librairie TinawebJS.

Les réseaux navigables obtenus avec les deux outils permettent de rechercher un nœud dans le réseau. Cependant, certaines fonctionnalités sont exclusives à chaque outil, ce qui rend les outils complémentaires. Dans TinawebJS, tous les nœuds correspondants à une requête sont mis en surbrillance dans le réseau, ce qui donne un meilleur aperçu de l'emplacement des résultats dans le réseau que le résultat du plugin d'export SigmaJS, qui fournit juste une liste de nœuds pertinents. TinawebJS se distingue aussi en fournissant une légende contenant les noms des clusters du réseau. En revanche, le réseau Sigma JS permet d'examiner plus facilement le contexte local de chaque nœud, car lors d'un clic sur un nœud, celui-ci et ses voisins immédiats sont isolés de manière plus claire qu'avec TinawebJS : ainsi, avec le réseau Sigma JS, il est facile de naviguer en cliquant successivement sur les nœuds et leurs voisins, ce qui peut révéler des connexions inédites. Un exemple de réseau visualisé avec TinawebJS se trouve sur la figure 5 et un exemple pour le plugin d'export Sigma JS de Gephi se trouve sur la figure 6.

[21] Solr : https://lucene.apache.org/solr/

[22] GEXF veut dire *Graph Exchange XML Format*. Ce format a été créé par le projet Gephi https://gephi.org/gexf/format/

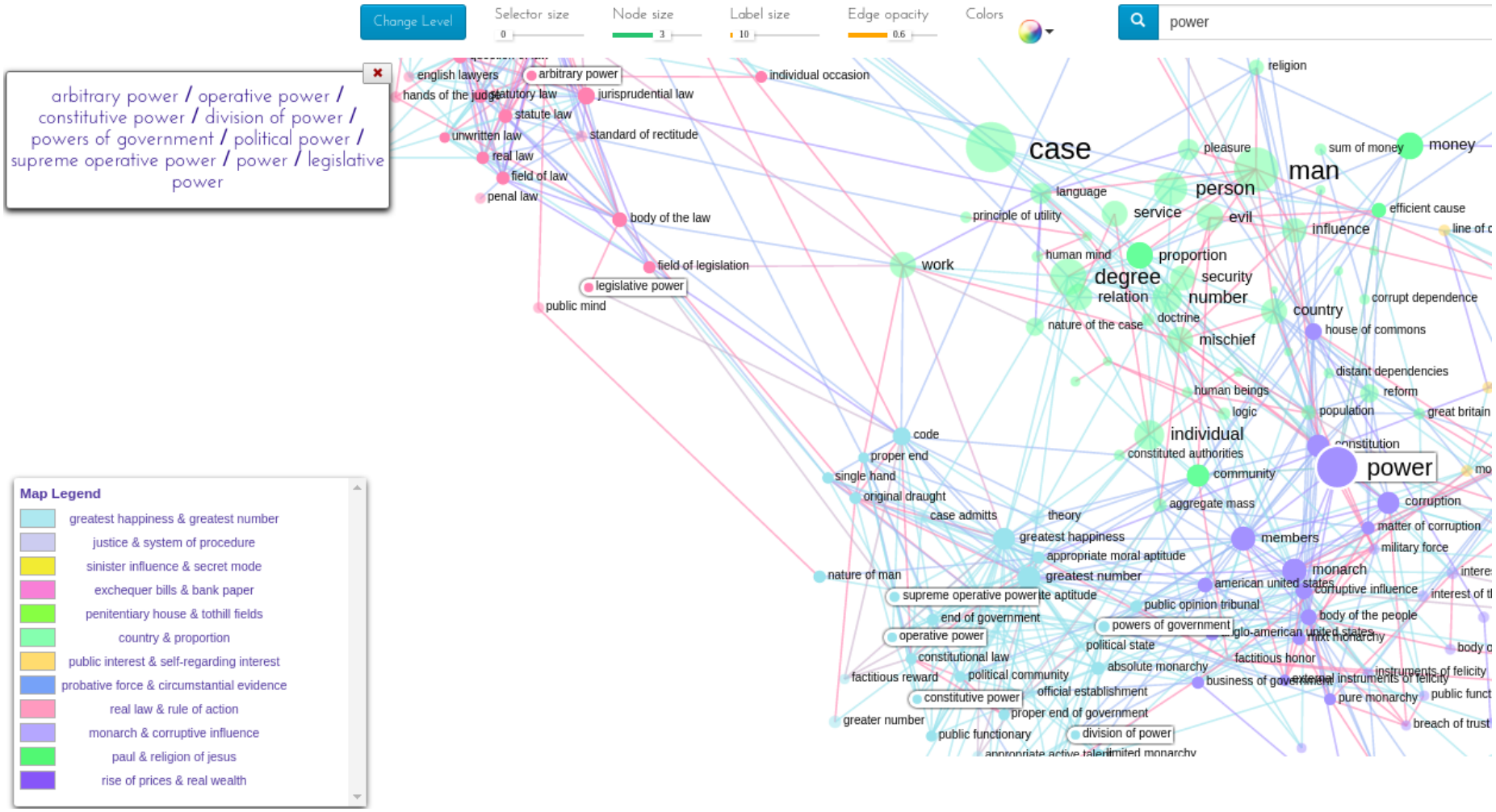


FIGURE 5 – Rendu par TinawebJS du réseau de 250 termes. Les nœuds qui contiennent le mot *power* (« pouvoir ») ont été identifiés, et l'outil les met en surbrillance dans le réseau. Cette analyse met en évidence différentes communautés (identifiées par des couleurs différentes), c'est-à-dire différents contextes dans lesquels la notion de pouvoir est évoquée par Bentham. Le résultat montre que la notion de « pouvoir » est évoquée en même temps que la notion de monarque (communauté *monarch & corruptive influence*) et en même temps que ce que Bentham considérait l'objectif des gouvernants, c'est-à-dire maximiser le plus grand bonheur du plus grand nombre (communauté turquoise, première dans la légende). Le pouvoir dans un contexte législatif est également discuté (communauté rose pâle contenant les nœuds *arbitrary power* et *legislative power*). La bibliothèque TinawebJS a été créée par l'Institut des sciences complexes et le laboratoire CAMS à Paris.

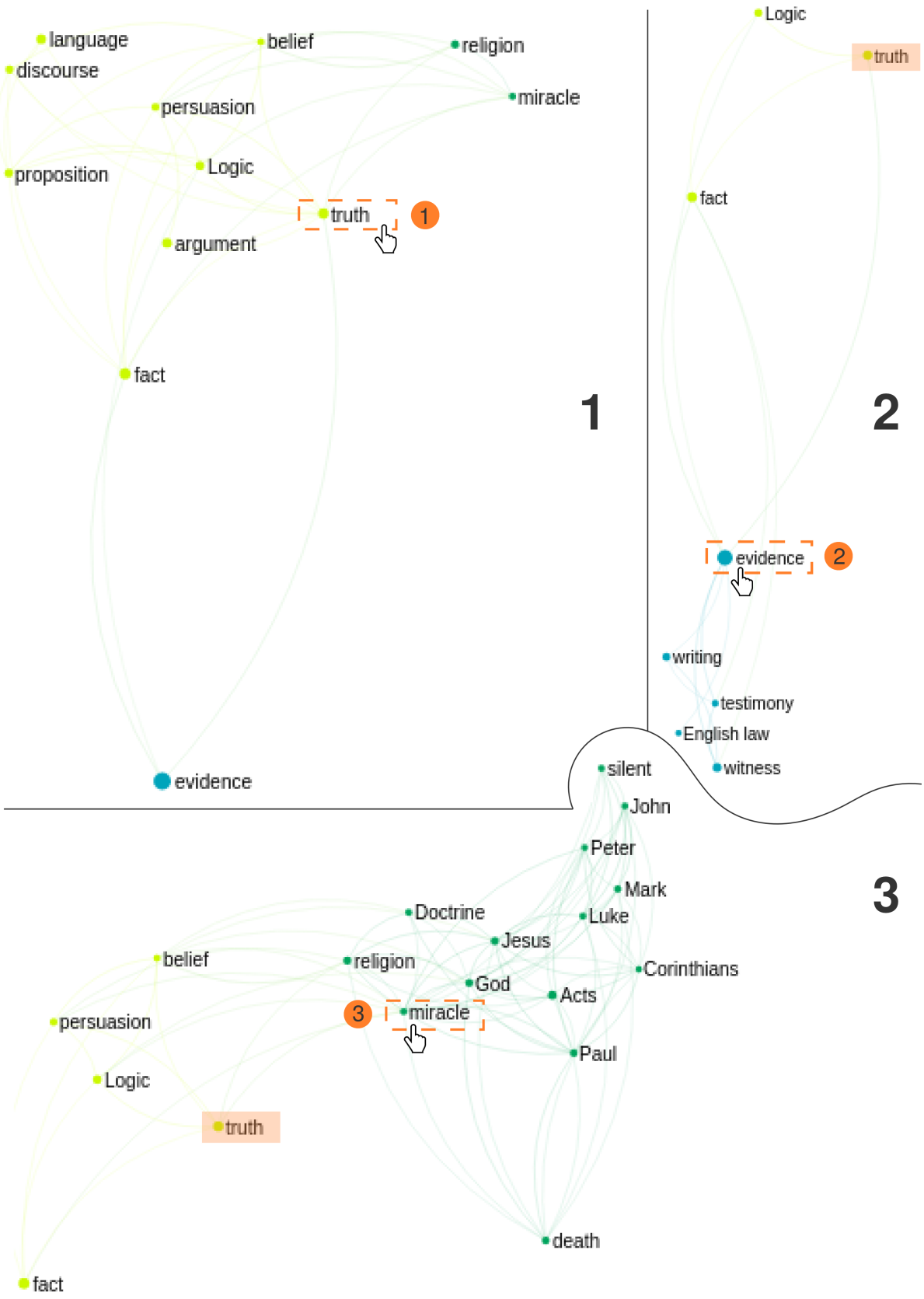


FIGURE 6 – Navigation dans le corpus et exploration du contexte local par sélection successivement de nœuds voisins (l'interface peut être testée sur http://apps.lattice.cnrs.fr/bentham/bentham-js-more.html). **(1)** Le nœud *truth* (« vérité ») est sélectionné. Il a des voisins dans le cluster vert clair *discourse & proposition*, et est lié via *evidence* au cluster bleu *court & procedure*. Il a également des voisins dans le cluster sur la religion (vert foncé). **(2)** En sélectionnant *evidence*, on voit les nœuds les plus proches de *truth* dans le cluster bleu lié à la judicature. **(3)** En sélectionnant *miracle*, on voit les nœuds les plus proches de *truth* dans le cluster sur la religion. Donc, à partir d'un nœud donné comme *truth*, nous pouvons naviguer dans le réseau en cliquant séquentiellement sur ses nœuds voisins et sur l'ensemble de nœuds liés à chaque voisin. Cette exploration peut éventuellement suggérer des connexions inconnues auparavant du chercheur.

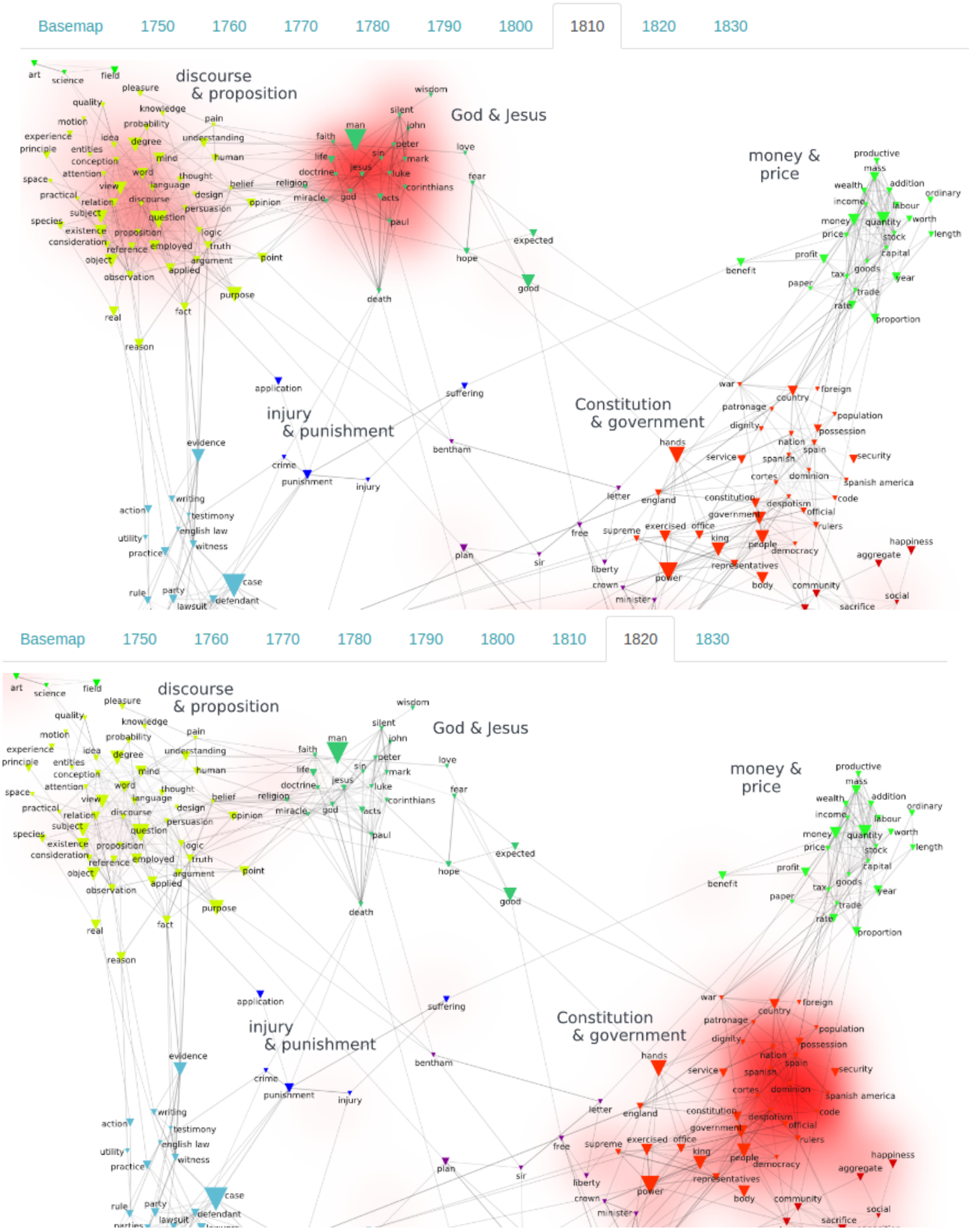


FIGURE 7 – Heatmaps par décennie, fondées sur des mentions de concepts obtenues grâce au module d'Entity Linking. La nuance rouge devient plus foncée lorsque la surreprésentation d'un thème dans une décennie donnée augmente. Par exemple, **en haut**, dans les années **1810**, le corpus se concentre sur le discours humain et la religion (clusters pour *discourse & proposition* et *God & Jesus*). **Bas :** Dans les années **1820**, les manuscrits se concentrent fortement sur les notions exprimées dans le cluster *Constitution & government*. Un expert de Bentham a confirmé que ces heatmaps correspondent à l'évolution temporelle des écrits de Bentham : dans les années 1820, Bentham a ainsi écrit ou commenté des codes constitutionnels pour plusieurs pays. Voir http://apps.lattice.cnrs.fr/bentham/heatmaps-more.html pour une meilleure résolution d'image.

Outre les cartes navigables qui viennent d'être présentées, des **heatmaps** ont été créées, qui montrent les zones saillantes du corpus par décennie, comme décrit à la p. 277. Un exemple de heatmap se trouve sur la figure 7. Les zones en rouge montrent comment, dans les années 1810, les manuscrits se concentrent sur le raisonnement humain et la religion, c'est-à-dire les communautés appelées *discourse & proposition* (« discours & proposition ») et *God & Jesus* (« Dieu & Jésus »), alors que dans les années 1820, les manuscrits se concentrent sur le cluster *Constitution & government* (« Constitution & gouvernement »). Un expert a confirmé le bien fondé de ces représentations (voir p. 285).

2.5 Évaluation par des experts du domaine

Cette sous-section décrit une évaluation par deux experts de l'œuvre de Bentham. Le premier expert est chercheur post-doctoral spécialisé en Bentham à University College London (UCL), et ancien coordinateur du projet Transcribe Bentham. L'autre expert est professeur d'Humanités numériques à UCL et a également participé au projet Transcribe Bentham. Dans ce résumé, nous nous concentrons surtout sur les commentaires fournis par le spécialiste de Bentham.

Comme notre enquête n'a impliqué que deux experts, l'exercice doit être considéré comme une validation préliminaire du travail effectué. L'**objectif de cette évaluation** était surtout d'obtenir des renseignements sur les sujets suivants :

- **Plausibilité des représentations :** Existe-t-il des erreurs évidentes qui nous invalideraient les réseaux de concepts générés ?
- **Utilité de chaque type d'analyse** pour créer les représentations sous forme de réseau. Nous avons extrait deux types d'expression : des mentions de concepts DBpedia, et des termes, comme décrit dans 2.3.1. Notre hypothèse est que les termes soient perçus par les experts comme plus utiles pour les utilisateurs spécialisés que les mentions de concepts DBpedia, car les termes sont plus susceptibles de représenter des concepts spécifiques de la pensée de Bentham que des concepts d'une base de connaissances générique comme DBpedia.
- Possibilité de donner naissance à **nouvelles idées** sur le corpus, à travers la découverte de connexions spécifiques dans le réseau de termes ou les heatmaps produites à partir du corpus.

2.5.1 Procédure d'évaluation

Trois exemples d'utilisation de l'interface ont été montrés aux experts. Nous leur avons signalé que les termes qui ont des liens avec deux clusters correspondent à des concepts liés au contenu des deux clusters à la fois (figure 8).

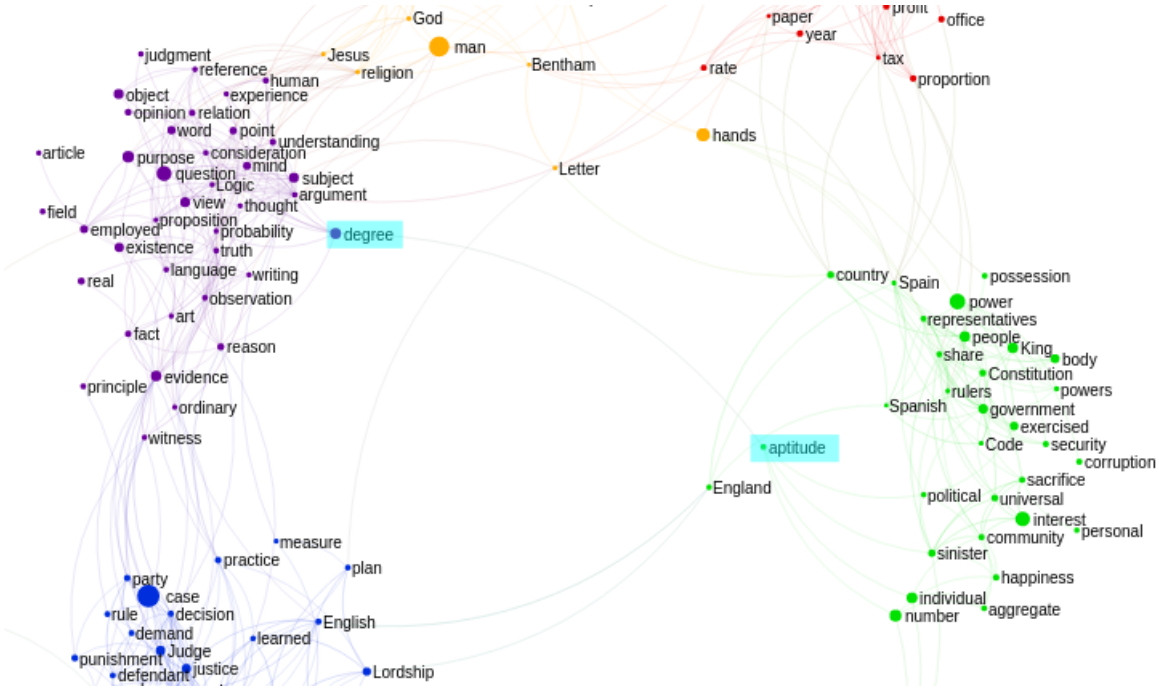


FIGURE 8 – Évaluation de l'interface : Exemple de deux nœuds reliant deux clusters dans la carte de 150 mentions de concepts DBpedia (nœuds *degree* (« dégré ») et *aptitude*). Le réseau est visualisé avec le plugin d'export SIGMAJS de l'outil GEPHI pour la visualisation de graphes

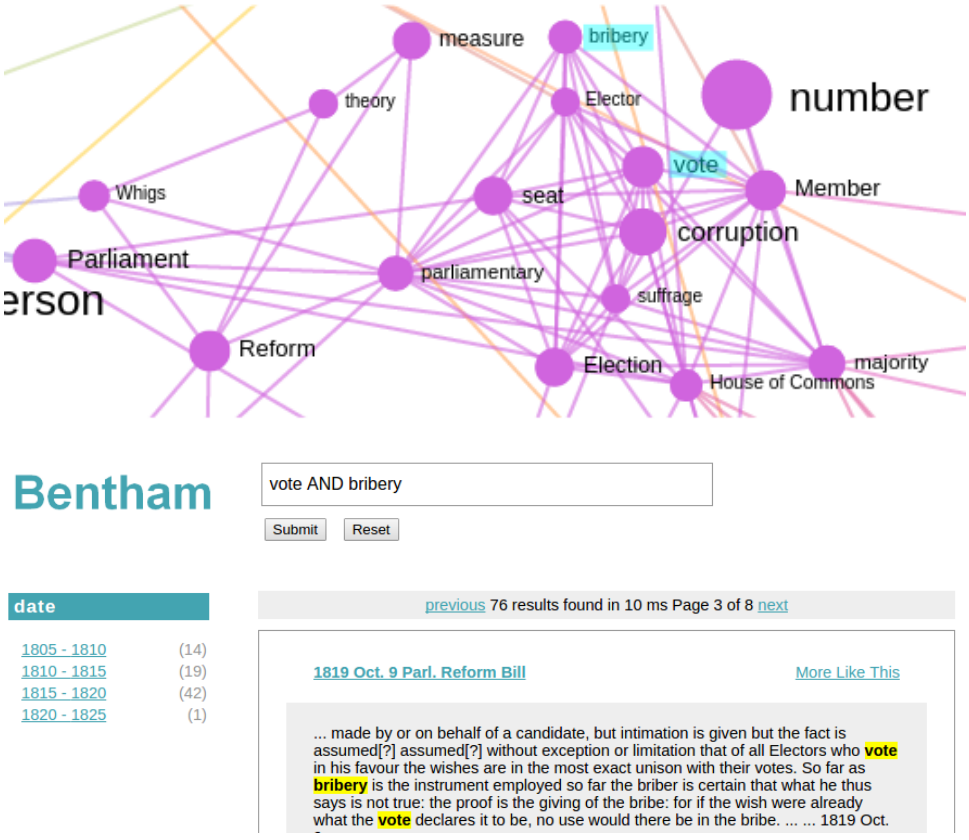


FIGURE 9 – Évaluation de l'interface : Exemple de vérification dans l'index de recherche des contextes de cooccurrence des nœuds du réseau, ici *vote* et *bribery* (« vote » et « subornation »). Le réseau (**haut**) montre les concepts connectés. L'index de recherche (**bas**) permet de chercher leurs contextes de cooccurrence

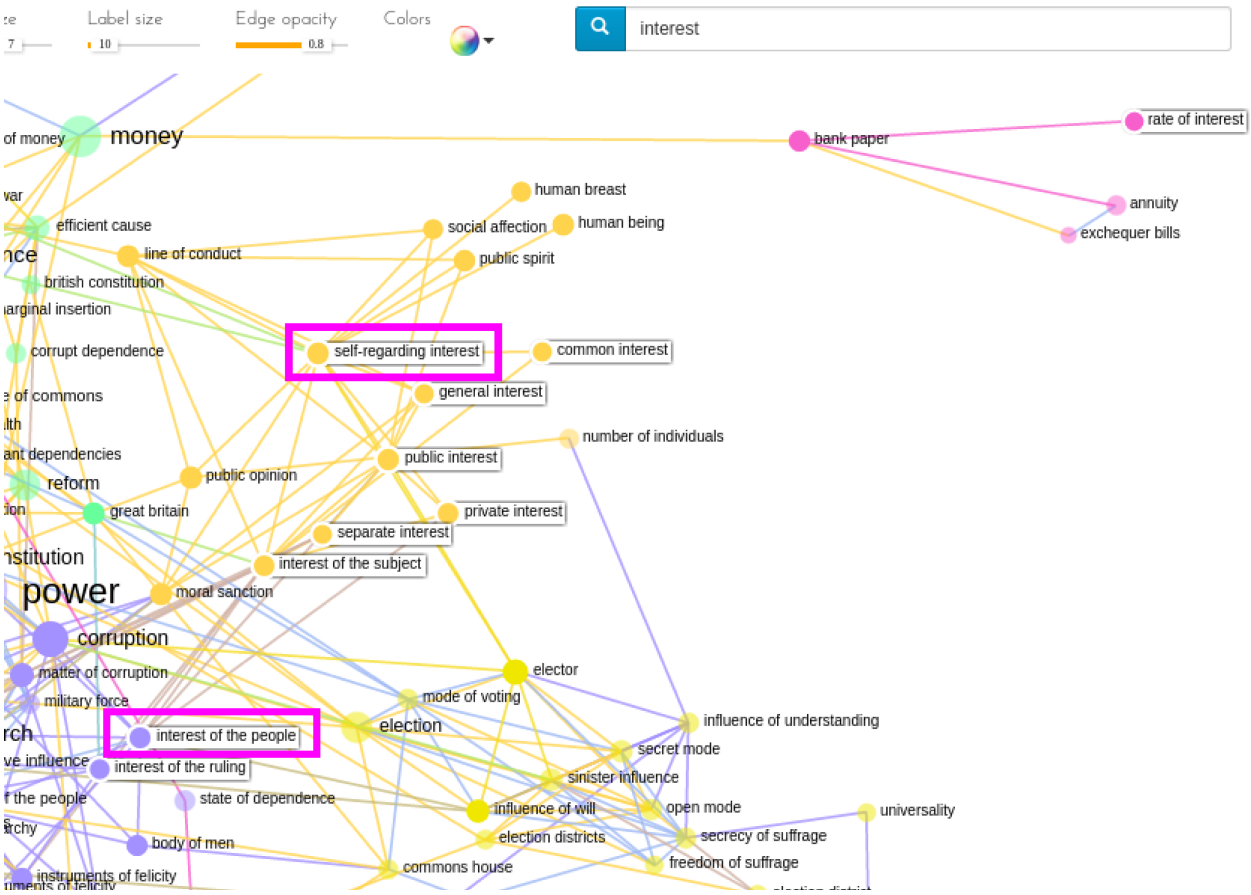


FIGURE 10 – Évaluation de l'interface utilisateur : termes qui correspondent à la requête *interest* (« intérêt ») dans la carte de 250 termes. Les résultats montrent des synonymes et antonymes pour le terme *sinister interest* (« intérêt sinistre »). Le réseau est visualisé avec TINAWEBJS. En partant d'un terme important dans le corpus, comme *sinister interest*, l'expert du domaine a identifié parmi les résultats des quasi-synonymes pour le terme, comme *self-regarding interest* (« intérêt individuel »), et des quasi-antonymes, comme *interest of the people* (« intérêt du peuple »). Cela suggère l'utilité des cartes pour trouver des formulations alternatives d'un concept donné et pour examiner les termes voisins.

Nous leur avons également montré comment l'index de recherche peut être utilisé pour trouver le contexte de chacun des nœuds du réseau, ainsi que des contextes ou les nœuds du réseau sont en cooccurrence (figure 9). Nous leur avons également montré comment utiliser les fonctions de recherche dans les réseaux navigables (voir la figure 10).

En ce qui concerne les réseaux navigables, des cartes de termes extraits par le module d'Entity Linking et par l'extraction de motifs ont été montrées aux experts. Concernant les heatmaps, seulement celles fondées sur les résultats du module d'Entity Linking ont été discutés avec les experts, pour des raisons de temps.

2.5.2 Résumé des résultats de l'évaluation

Au sujet de la plausibilité des représentations générées pour le corpus, l'expert de l'œuvre de Bentham a confirmé que l'aperçu des thèmes identifiés

dans les réseaux navigables, ainsi que les connexions entre les différents nœuds dans les graphes sont en accord avec ses connaissances sur le corpus et surtout que les cartes ne présentent pas d'erreurs évidentes. Concernant les heatmaps par décennie, il a constaté que les zones du corpus identifiées comme étant saillantes au cours de chaque décennie correspondent aux intérêts principaux de Bentham pour chacune de ces périodes.

L'expert n'a pas fait de « découverte » fondamentale à travers l'utilisation de l'application mais ce n'était pas obligatoirement l'objet de cette première expérience. Les réseaux fondés sur les termes ont été jugés par l'expert comme plus informatifs que les réseaux basés sur les mentions de concepts DBpedia, extraites par le module d'Entity Linking. Les termes arrivent à exprimer des concepts précis, caractéristiques de la pensée de Bentham, comme *sinister interest* ou *operative power*.[23] L'expert a trouvé que les mentions de concepts DBpedia représentent des éléments de base, comme *interest* (« intérêt ») ou *power* (« pouvoir »), de la signification de certains termes définis par Bentham, mais sans correspondre de façon plus exacte aux expressions précises que Bentham a utilisées.

Même si l'expert n'a pas obtenu de nouvelles idées de recherche en utilisant l'application, il considère que l'application a un potentiel pour aider à acquérir de nouvelles connaissances sur le corpus de la façon suivante : dans les réseaux de motifs, il a trouvé des formulations alternatives pour des termes caractéristiques de Bentham. Par exemple, un concept de base dans les écrits de Bentham est l'idée de *sinister interet*.[24] Les réseaux montrent des termes proches à ce dernier, comme *self-regarding interest*, *interest of the subjet* et *private interest*.[25] Selon l'expert, de telles formulations alternatives sont utiles pour le travail éditorial sur le corpus. Elles sont utiles pour déterminer la façon dont Bentham aborde certaines notions, et identifier des passages où il utilise des formulations alternatives. Pour l'expert, les réseaux sémantiques sont importants sur le plan terminologique.

Sur la base de ces derniers commentaires de la part de l'expert, les travaux futurs les plus pertinents seraient la création de modèles de sémantique distributionnelle pour l'analyse des termes en corpus. Ces modèles permettent d'obtenir des ensembles d'expressions liées (synonymes, hyponymes, etc.)

[23] Des traductions littérales seraient « intérêt sinistre » et « pouvoir opérationnel ». La signification exacte des termes n'est pas essentielle pour la discussion ici. Pour *sinister interest*, cf. la note 24.

[24] Les nuances du terme *sinister interest* ne sont pas essentielles pour la discussion ici. Sommairement, Bentham considère que l'intérêt général doit primer comme motivation des actions d'un gouvernant sur l'intérêt individuel, qu'il appelle *sinister interest* (« intérêt sinistre ») dans ce contexte.

[25] On pourrait traduire ces expressions par « intérêt individuel », « intérêt du sujet » et « intérêt privé » respectivement.

par l'analyse de similarités contextuelles. Un modèle distributionnel qui pourrait être facilement utilisé est word2vec (Mikolov et al., 2013).

2.5.3 Interprétabilité des résultats de l'évaluation

La nature visuelle du produit que nous évaluons ici (des cartes-réseaux de concepts) pose une difficulté particulière. Comme Rieder et al. (2012) le suggère, après de nombreux autres comme Heintz (2007, p. 78), il est facile de surestimer la valeur et la fiabilité d'une représentation visuelle, et de recréer une réalité externe à partir d'une représentation, c'est-à-dire en quelque sorte inverser le processus d'interprétation, en cherchant inconsciemment à faire correspondre la réalité à la représentation, plutôt que l'inverse. Il est donc primordial que l'expert puisse se renseigner sur les méthodes et les biais éventuels impliqués dans la production des cartes sémantiques à sa disposition et puisse éventuellement les remettre en question.

Sans aller jusqu'à proposer de fausses cartes, ce qui poserait des problèmes éthiques évidents, il semble important d'essayer de produire des cartes alternatives (à partir de différents algorithmes par exemple). Nous l'avons fait ici en proposant deux séries de cartes différentes fondées sur deux techniques de repérage de termes différentes mais il serait souhaitable d'aller plus loin dans cette voie.

2.6 Conclusions et perspectives

Une analyse de corpus aboutissant à des cartes sémantiques a été présentée et appliquée aux manuscrits de Jeremy Bentham. Des éléments lexicaux permettant de modéliser le corpus ont été identifiés de deux manières différentes : avec un module d'Entity Linking d'une part et avec un extracteur de termes d'autre part. Des réseaux de cooccurrence ont été créés sur la base de ces éléments lexicaux. Les réseaux ont été rendus navigables avec deux bibliothèques de visualisation de graphes, et des « heatmaps » ont été créés pour représenter l'évolution temporelle du contenu du corpus. Outre l'indexation en texte intégral, l'analyse a permis de produire des réseaux de concepts représentant les thèmes abordés dans le corpus. Toutes les cartes représentant le corpus, ainsi que l'index de recherche, sont accessibles depuis une interface utilisateur.[26]

Une évaluation de l'application par deux experts du domaine a donné les résultats suivants : les réseaux créés avec les termes extraits du corpus ont été jugés plus informatifs pour les spécialistes que ceux créés à partir des mentions de concepts DBpedia. Cela n'est pas surprenant car DBpedia est une base de connaissances trop générique pour un corpus spécialisé comme les manuscrits de Bentham. Les experts ont considéré que les réseaux sont

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

intéressants pour identifier des formulations alternatives de concepts importants chez Bentham. Les outils dont ils disposent à l'heure actuelle sont moins performants sur ce point qui est important pour le travail éditorial sur l'œuvre du philosophe. Pour améliorer encore l'analyse sémantique au niveau terminologique, l'intégration d'un modèle de similarité distributionnelle serait à envisager.

3 Étude de cas : Le *Bulletin des Négociations de la Terre (ENB)*

La deuxième étude de cas vise à analyser les déclarations des différents acteurs impliqués dans le *Bulletin des Négociations de la Terre* (*ENB* selon son acronyme anglais, pour *Earth Negotiations Bulletin*, que nous utilisons ici). Il s'agit d'un corpus de rapports issus des sommets climatiques où des traités internationaux comme le Protocole de Kyoto de 1997 ou les Accords de Paris de 2015 ont été signés.[27] Dans la mesure où s'agit d'analyser des négociations, il est important de savoir quels acteurs soutiennent telle ou telle proposition, et quels acteurs s'y opposent. Nous décrivons ici une chaîne de traitement linguistique et un modèle du domaine permettant d'extraire automatiquement ce type d'informations. Nous avons également créé une interface utilisateur qui permet d'explorer le corpus en fonction de l'information extraite : les acteurs qui soutiennent ou s'opposent un point de la négociation peuvent être recherchés, ainsi que les acteurs qui sont d'accord ou en désaccord sur un sujet donné.[28]

La présentation de cette étude de cas est structurée comme suit. Nous décrivons tout d'abord le corpus (3.1) et nous donnons un aperçu des travaux antérieurs sur la question (3.2). Nous détaillons ensuite notre système d'extraction de relations, ainsi qu'une évaluation quantitative de ses résultats (3.3). Nous poursuivons avec la présentation de l'interface utilisateur permettant l'exploration dynamique du corpus (3.4), et une évaluation qualitative par des experts du domaine (3.5). Nous concluons enfin avec un bilan et des perspectives (3.6).

3.1 Description du corpus

Le *Bulletin des Négociations de la Terre* est une publication qui couvre les négociations internationales sur les politiques climatiques.[29] Son 12^{e} volume couvre les Conférences des Parties (COP), c.à.d. des sommets annuels qui

[27] http://unfccc.int/portal_francophone/accord_de_paris/items/10081.php

[28] L'interface se trouve sur http://apps.lattice.cnrs.fr/ie/uidev/

[29] http://enb.iisd.org/enb/vol12/

ont servi à négocier des traités internationaux dans le domaine, comme la COP 21 qui s'est déroulée à Paris fin 2015.[27]

Le corpus ENB fournit des rapports quotidiens sur les déclarations des participants lors des négociations. Il vise un ton objectif et impose l'utilisation d'un vocabulaire semi-contrôlé, notamment un ensemble de verbes de parole et de prise de position censé être neutre pour rendre compte des débats. Par exemple, *objected* (« a fait l'objection ») ou *stated* (« a affirmé ») plutôt que *attacked* (« a attaqué ») ou *accused* (« a accusé »). Le corpus a aussi tendance à utiliser une variété limitée de structures syntaxiques, afin d'éviter des moyens stylistiques qui pourraient mettre l'accent sur les interventions de certains participants en détriment des autres.

Le corpus est publié par l'*International Institute for Sustainable Development (Institut international pour le développement durable)*.[30] Les éditeurs du corpus sont des experts en politique climatique. Il est donc un corpus créé par des experts à l'intention des experts.[31]

Notre échantillon du corpus couvre 23 conférences sur le climat (COP) qui se sont tenues entre 1995 et 2015. Les rapports COP du corpus ENB sont de deux types : des rapports quotidiens (publiés chaque jour pendant la durée d'une COP) et des résumés correspondant à une COP au complet. Nous avons seulement inclus les rapports quotidiens dans notre échantillon, car les résumés correspondant à une COP complète ont en fait tendance à simplement reproduire le contenu des rapports quotidiens, ce qui entraînerait une duplication du contenu pour nos analyses. L'échantillon contient donc 258 rapports quotidiens, correspondant à environ 500 000 mots (24 000 phrases).

Le corpus publié en ligne est en HTML. Nous avons créé des scripts permettant de télécharger les fichiers HTML puis les « nettoyer » afin de créer deux versions, en texte pur et en XML structuré, de chaque rapport.

3.2 Travaux précédents sur le corpus

Venturini et al. (2014) et Baya-Laffite et al. (2016) détaillent des recherches visant à produire une **cartographie** du corpus ENB, de manière similaire à ce que nous avons fait avec le corpus Bentham (section 2) avec la plate-forme CorText (cf. 2.3.2). L'objectif de ces études était d'observer l'évolution des thèmes abordés lors des conférences sur le climat depuis les années 1990. Les modélisations produites ne permettent toutefois pas d'avoir accès aux relations de soutien et d'opposition entre acteurs sur les différents sujets des négociations, faute d'avoir recours à des technologies permettant d'identifier

[30] http://www.iisd.org/
[31] http://enb.iisd.org/about/team/

de telles relations. L'objectif principal de notre travail ici consiste précisément à essayer de répondre à cette limite.

Salway et al. (2014) ont appliqué une méthode inspirée de techniques d'« **induction de grammaire** » au corpus pour en extraire, de façon peu supervisée, des motifs autour des acteurs et des concepts en jeu. Certains des patrons induits par leur méthode sont pertinents et permettent d'analyser des relations de soutien et d'opposition entre les acteurs. Cependant, le résultat de leur analyse n'est pas exhaustif par rapport au repérage du contenu des déclarations des acteurs.

Une **interface de navigation** pour le corpus ENB a été créée par le médialab de Sciences Po (Venturini et al., 2015). Pour développer cette interface l'équipe du médialab a demandé à des experts du domaine d'identifier une liste de sujets importants dans le corpus. Cette analyse est faite manuellement, sur la base d'une extraction lexicale automatique effectuée avec CorText. Chaque paragraphe du corpus est ensuite étiqueté avec un ou plusieurs termes de cette liste. L'interface obtenue permet de rechercher des paragraphes où un acteur ou un sujet est mentionné ; une copie de l'interface se trouve dans la thèse complète en anglais (p. 168).

L'interface du médialab offre une recherche fondée sur les métadonnées et non sur le texte intégral, ce qui peut parfois être gênant : l'interface que nous visons doit au contraire fournir une recherche en texte intégral. Une caractéristique intéressante de l'interface du médialab est qu'elle donne un aperçu de la distribution temporelle des résultats pour une requête. Elle montre aussi la distribution des acteurs et sujets dans les résultats. L'interface du médialab fournit des résultats au niveau des paragraphes et des documents, alors qu'il nous a semblé qu'il serait souvent plus pertinent de procéder à une analyse au niveau de la phrase. Enfin, les thèmes annotés dans l'interface du médialab ont été créées manuellement par des experts du domaine. Nous visons au contraire à produire des annotations automatiquement à partir d'une chaîne de traitement linguistique.

3.3 Extraction de relations

Les relations que nous extrayons pour cette étude de cas correspondent à l'attitude des acteurs vis-à-vis des thèmes abordés lors des conférences, sur la base des prédicats utilisés (verbes de parole, marqueurs de prise de position). Nous annotons ensuite sur cette base les relations d'appui et d'opposition entre acteurs, ainsi que les points de la négociation sur lesquelles ils sont en accord ou en désaccord.

Dans la thèse complète, un état de l'art sur les différentes technologies qui peuvent être utilisées pour extraire des relations a été fourni. Ici, nous

limitons la discussion au système que nous avons créé à cette fin, et qui repose sur une technologie appelée *étiquetage en rôles sémantiques* (*Semantic Role Labeling* ou *SRL*). L'analyse exige en outre de disposer d'un analyseur en dépendances, c'est-à-dire d'un outil d'analyse automatique des fonctions syntaxiques.

La relation entre un acteur et une prise de position est formalisée sous la forme d'une *proposition*, qui est définie comme un triplet du type ⟨*acteur*, *prédicat*, *message*⟩. Un prédicat peut être un verbe de parole, par exemple *state* (« affirmer »). Il peut aussi être un nom de parole comme *objection* (« objection »). Des exemples de propositions extraites par notre système sont disponibles sur la figure 11 (p. 296), qui montre notre interface de navigation pour le corpus. Le panneau de gauche affiche les triplets extraits : la colonne *Actor* pour les acteurs, *Action* pour les prédicats et *Point* pour les messages. Le panneau de gauche montre les phrases à partir desquelles les propositions ont été extraites.

Nous décrivons ici le système développé pour extraire les propositions : tout d'abord, la chaîne de traitements TAL exploitée par le système (3.3.1), puis le modèle du domaine, contenant des acteurs et des prédicats, qui a été utilisé pour identifier des propositions pertinentes (3.3.2). Nous décrivons ensuite les règles qui s'appliquent à la sortie de la chaîne de traitements TAL pour identifier les propositions, compte tenu du modèle de domaine (3.3.3). L'efficacité du système, qui a été évaluée par rapport à un jeu de référence, sera enfin discutée.

3.3.1 Chaîne de traitements TAL

Les modules TAL utilisés pour l'extraction de propositions font partie de la librairie IXA PIPELINE (Agerri et al., 2014). Cette librairie a été choisie car ses résultats sont à l'état de l'art d'après des évaluations récentes et elle utilise un format d'annotation XML facile à exploiter.

La **tokénisation** et l'**étiquetage grammatical** (part-of-speech tagging) ont été effectués avec les modèles par défaut de IXA PIPELINE pour l'anglais. L'**analyse en constituants syntaxiques** (requise par le module de coréférence) a également été réalisée avec le modèle par défaut de la librairie.

Des **chaînes de coréférence** ont été fournies par l'outil COREFGRAPH.[32] Il s'agit d'une implémentation en Python du module DCOREF de Stanford (Lee et al., 2013), et son format d'annotation est compatible avec la librairie IXA. Compte tenu de certains caractéristiques non-standards dans l'usage des pronoms dans notre corpus (par ex. *he* pour désigner un pays), nous avons créé des règles spécifiques pour la résolution des anaphores pronominales,

[32] https://bitbucket.org/Josu/corefgraph

en fonction de la sortie de COREFGRAPH et de l'analyse en dépendances ; des détails seront fournis à la p. 294. La coréférénce et la résolution d'anaphores sont évidemment des problèmes complexes en TAL. Nous avons conscience de ne pas avoir traité tous les cas possibles de ces phénomènes. Nous nous sommes plutôt concentrés sur un sous-ensemble des cas possibles qui était le plus pertinent pour notre corpus (et relativement bien pris en compte par l'analyseur utilisé), mais la coréférence dans un sens plus large n'a pas été abordée.

L'**analyse en dépendances** et l'**étiquetage en rôles sémantiques** (SRL) ont été effectués avec le module IXA-PIPE-SRL.[33] Il s'agit d'un outil compatible avec la pipeline IXA. Il donne accès aux outils d'analyse en dépendances et SRL dans la librairie MATE TOOLS (Björkelund et al., 2010), en introduisant quelques améliorations.[34] Les schémas d'annotation en dépendances syntaxiques et rôles sémantiques sont ceux de la conférence CoNLL (Buchholz et al., 2006 pour la syntaxe, Carreras et al., 2005 pour le SRL).[35] L'étiquetage SRL correspond aux bases de données PROPBANK et NOMBANK (Palmer et al., 2005 et Meyers et al., 2004 respectivement).

Format d'annotation : Une caractéristique assez pratique des outils exploités est qu'ils utilisent tous un format d'entrée et sortie commun : le *NLP Annotation Format* ou *NAF* (Fokkens et al., 2014) ; on peut traduire le nom par « Format d'annotation pour le TAL ». Il s'agit d'un format XML composé de couches, chacune représentant une étape de l'analyse linguistique automatique (catégories grammaticales, analyse syntaxique, SRL, etc.). La librairie KAFNAFPARSERPY a été utilisée pour gérer les annotations NAF.[36]

Comme il a été mentionné, l'objectif du système est d'extraire des propositions, c.à.d. des triplets de forme $\langle actor, prédicat, message \rangle$.[37] Après l'extraction des propositions, les messages de celles-ci ont été annotés avec des motifs, avec des concepts DBpedia et avec des concepts d'un thesaurus sur la politique climatique. Ces annotations permettent de comparer les différents messages des acteurs entre eux sur l'interface utilisateur.

L'extraction de termes a été réalisée avec YATEA (Aubin et al., 2006). Cet outil extrait tant des termes composé de un ou plusieurs mots, de façon non

[33] https://github.com/newsreader/ixa-pipe-srl

[34] Les améliorations consistent à une annotation simultanée vers plusieurs bases de données lexicales, grâce à l'outil PREDICATE MATRIX (matrice de prédicats) par López de Lacalle et al., 2016. Cf. le site mentionné sur la note 33.

[35] CoNLL : *Conference on Natural Language Learning* (« Conférence sur l'apprentissage en langage naturel »), une conférence spécialisée dans les méthodes d'apprentissage statistique pour la création d'outils et ressources d'analyse linguistique.

[36] https://github.com/cltl/KafNafParserPy

[37] Des exemples de propositions extraites par notre système se trouvent sur la figure 11 (p. 296). Le panneau de gauche affiche les triplets extraits, et le panneau de droite montre les phrases où ils ont été extraits.

supervisée, en utilisant des critères syntaxiques et statistiques. Nous avons également appliqué cet outil au corpus de Bentham (cf. p. 275).

La tâche d'Entity Linking vers DBpedia a été effectuée avec DBPEDIA SPOTLIGHT, qui a déjà été décrit à la p. 273.

Concernant le thesaurus du domaine, il a été exploité pour identifier des sujets plus spécifiques au climat qu'il serait impossible de repérer en utilisant seulement la base de connaissances DBpedia. Le thesaurus spécialisé que nous avons appliqué est le *Climate Thesaurus* (Bauer et al., 2011).[38] Il couvre l'énergie renouvelable et un ensemble de pratiques de gestion du changement climatique connues sous le nom de « développement compatible avec le climat ». Des termes relatifs à la politique climatique pertinents pour analyser les négociations sur le climat font partie du thesaurus.

3.3.2 Modèle du domaine

Le modèle du domaine contient des acteurs et des prédicats. Les acteurs représentent les participants aux négociations internationales sur le climat et sont formalisés par un lien (un « mapping ») entre les variantes mentionnées dans le texte et le nom canonique utilisé dans DBpedia.[39] Le modèle contient également des entrées (sous forme de lemmes) pour des prédicats de parole et de prise de position, qu'ils soient verbaux et nominaux. Certains verbes dans le modèle sont des verbes de récit neutre, comme *announce* (« annoncer »). D'autres verbes expriment des notions comme le soutien ou l'opposition, et l'accord ou le désaccord. Un exemple serait *criticize* (« critiquer »). Les verbes sont contenus dans la base de données PROPBANK (Palmer et al., 2005). Il faut souligner l'utilisation de prédicats nominaux, par exemple *announcement* (« annonce ») ou *objection* (« objection »). Ces prédicats nominaux (souvent des noms déverbaux) expriment des notions similaires aux prédicats verbaux du modèle, et ils sont bien décrits dans la base de données NOMBANK (Meyers et al., 2004). Le type de chaque prédicat (c.à.d. support, opposition ou récit neutre) est également spécifié dans le modèle.

Afin de garantir une bonne robustesse du système, le système d'extraction de propositions peut également être appliqué sans avoir recours à une liste prédéfinie d'acteurs. Dans ce cas, seront considérés comme des acteurs les éléments jouant un rôle argumental dans la structure prédicative identifiée à partir des prédicats décrits dans les ressources utilisées.

[38] http://www.climatetagger.net/climate-thesaurus/. Le thesaurus a également été connu sous le nom de *Reegle Thesaurus*.

[39] Par exemple l'acteur *United_States* (les États-Unis), peut être mentionné avec des variantes comme *the US*, *the USA*, *the United States* etc.

La liste complète des acteurs et prédicats du modèle (en anglais) se trouve dans l'annexe B.

3.3.3 Règles d'extraction

Plusieurs règles d'analyse ont été mises en place pour identifier des acteurs des propositions, sur la base des données décrites dans PROPBANK. Dans PROPBANK, le rôle *A0* correspond à l'agent d'un prédicat, le rôle *A1* est le patient ou l'objet, et les rôles *AM* représentent des compléments de temps, lieu etc.[40] La plupart des prédicats dans notre modèle impliquent un émetteur et un thème, ainsi que l'attitude de l'émetteur par rapport au thème (soutien, opposition ou récit neutre). Dans ce sens, l'argument *A0* d'un prédicat correspond généralement à l'émetteur du message, et l'argument *A1* contient généralement le message, c'est-à-dire le ou les points de la négociation abordés par l'acteur.

Dans certaines phrases, le rôle *A1* contient des acteurs plutôt qu'un point de négociation. C'est notamment le cas dans des constructions comme *China, opposed by the EU, accepted ...*.[41] L'agent de *opposed by* (« à qui s'est opposé... ») est l'agent d'une proposition qui contredit la proposition du verbe principal *accepted* (« a accepté »). Une règle d'extraction a été créée pour ces cas où un acteur s'oppose à la proposition exprimée par le verbe principal.

D'autres règles d'extraction ont été créées, pour traiter des cas plus spécifiques que les cas généraux qui viennent d'être évoqués. Par exemple, une règle ajoute les rôles sémantiques de type *A2* au message quand cela est pertinent. Les acteurs pertinents sont aussi parfois à rechercher dans des rôles de compléments étiquetés *AM-MNR* (ce qui correspond au complément de manière).

La **négation** est traitée en cherchant des rôles sémantiques *AM-NEG* liés aux prédicats, ainsi que des éléments lexicaux négatifs comme *not*, *cannot*, *lack of* (« pas », « peut pas », « manque de ») dans une fenêtre de deux mots précédant le prédicat. La négation est évidemment une question complexe en TAL et cette façon simple d'aborder la négation est relativement *ad hoc*, nous en avons conscience. Par exemple, les doubles négations ne sont pas abordées, et une phrase comme *There was no lack of disagreement*,[42] où les éléments soulignés sont négatifs tous les deux, ne serait pas traitée correctement.

Résolution d'anaphores pronominales : Dans le corpus, les pronoms personnels *he* et *she* (« il », « elle ») peuvent servir d'anaphore pour désigner un

[40] Palmer et al. (2005) décrit les rôles PROPBANK et Dowty (1991, p. 572) décrit les notions d'agent et patient qui sont à la base de ces rôles.

[41] *La Chine, à laquelle s'est opposée l'UE, a accepté ...*

[42] Littéralement traduit comme *Il n'y a pas eu de manque de désaccord*, pour exprimer qu'il y a eu beaucoup de désaccord, où la double négation est exprimée par *pas ... de manque de* (no lack of).

pays, en plus du pronom inanimé *it*.[43] Deux règles ont été créées pour gérer cette utilisation non-standard, et la résolution d'anaphores est limitée aux cas couverts par ces règles :

1. Le sujet du verbe principal d'une phrase (selon l'analyse de dépendances) peut être considéré comme l'antécédent d'un pronom *he* ou *she* (« il », « elle ») en position initiale dans la phrase suivante.
2. Parmi les antécédents possibles selon les chaînes de coréférence fournies par CorefGraph, seulement ceux qui sont dans la même phrase que le pronom ou dans la phrase précédente sont considérés.

Comme indiqué, ces règles ne couvrent pas tous les cas possibles. L'analyse de la résolution des anaphores est un phénomène complexe que nous n'avons pas cherché à traiter de manière exhaustive.

Enfin, pour faciliter les recherches par date, on assigne aux propositions la date du rapport dont ils ont été extraites.

Les propositions reçoivent un score de confiance en fonction de leur valeur informationnelle attendue. Par exemple, les facteurs suivants diminuent le score de confiance : un message très court (un ou deux mots), l'application de la résolution des anaphores ou la présence de caractères inhabituels dans le nom des acteurs. Les propositions qui montrent ces caractéristiques sont susceptibles d'être moins informatives, ou ont plus de chances de contenir des erreurs, du fait des traitements additionnels qu'on leur a appliqué.

Toutes les règles d'extraction de propositions et la procédure pour assigner des scores de confiance sont implémentées dans le langage de programmation Python.

Évaluation du système

L'efficacité du système a été évaluée par rapport à un jeu de test annoté manuellement, contenant 100 phrases et environ 300 propositions. Le jeu de test comprend des phrases qui représentent les défis posés par le corpus, c'est-à-dire des phrases avec plusieurs acteurs et prédicats, ou contenant une négation. Pour compter une proposition comme correcte, ses trois éléments (acteur, prédicat et message) doivent correspondre exactement à la référence manuelle. Sur la base de cette définition d'un résultat correct, le système a atteint un score F1 de 0,69, avec une précision (P) de 0,687 et rappel (R) de 0,693.[44] La qualité d'extraction fournie par le système a été jugée suffisante lors de l'évaluation par des experts du domaine.

[43] Le genre du pronom dépend alors du genre du ou de la représentant(e) pour ce pays dans les négociations.

[44] Les définitions habituelles ont été utilisées pour ces métriques : $P = \frac{vrais\ positifs}{vrais\ positifs + faux\ positifs}$; $R = \frac{vrais\ positifs}{vrais\ positifs + faux\ négatifs}$; $F1 = 2\frac{P \cdot R}{P+R}$

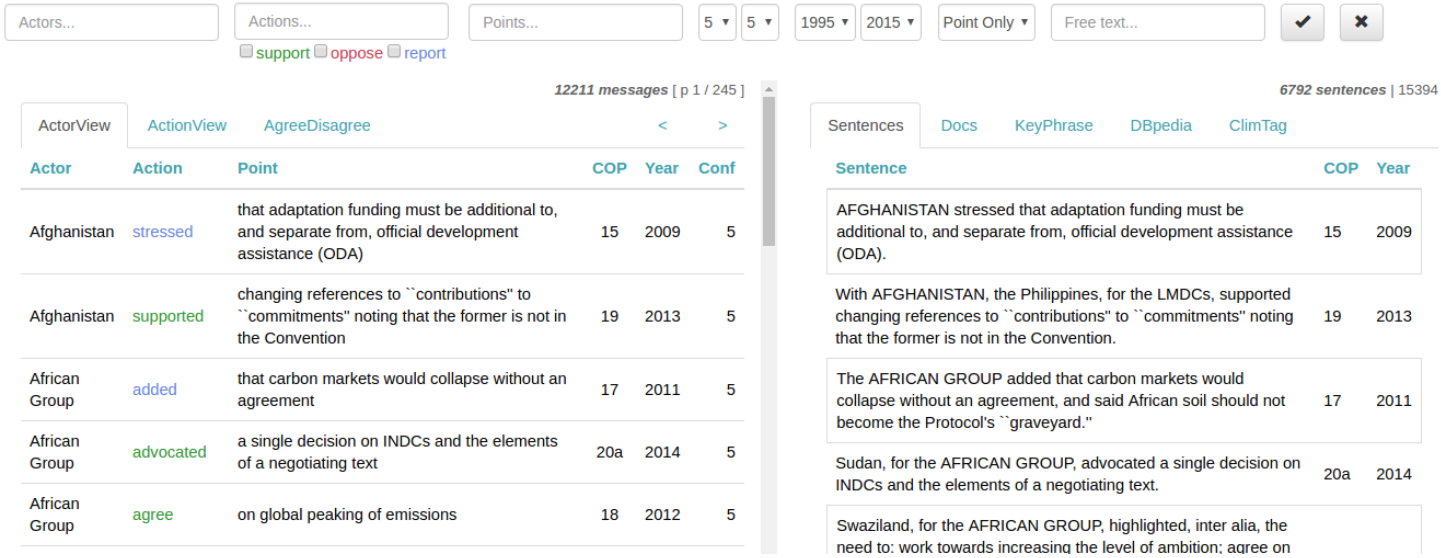


FIGURE 11 – Vue de l'interface utilisateur de l'application de navigation dans le corpus ENB. Le volet de gauche affiche les propositions et chaque onglet dans le volet de droite affiche différents types d'informations. Les phrases et les documents contenant des propositions pertinentes sont d'abord extraits, ainsi que les termes, les concepts DBpedia et les concepts du Climate Thesaurus mentionnés dans ces documents. Sur la gauche, les champs *Actors*, *Actions*, *Points* permettent de rechercher des éléments précis à l'intérieur des propositions (acteurs, prédicats et messages respectivement). Les types de prédicats peuvent être sélectionnés à partir des cases à cocher sous le champ *Actions* (*support*, *oppose*, *report*, pour les prédicats de soutien, opposition et récit neutre respectivement). La zone *Texte libre* (à droite) sert à la recherche dans le texte intégral. Les résultats peuvent être filtrés par « confiance » et par date avec des menus déroulants.

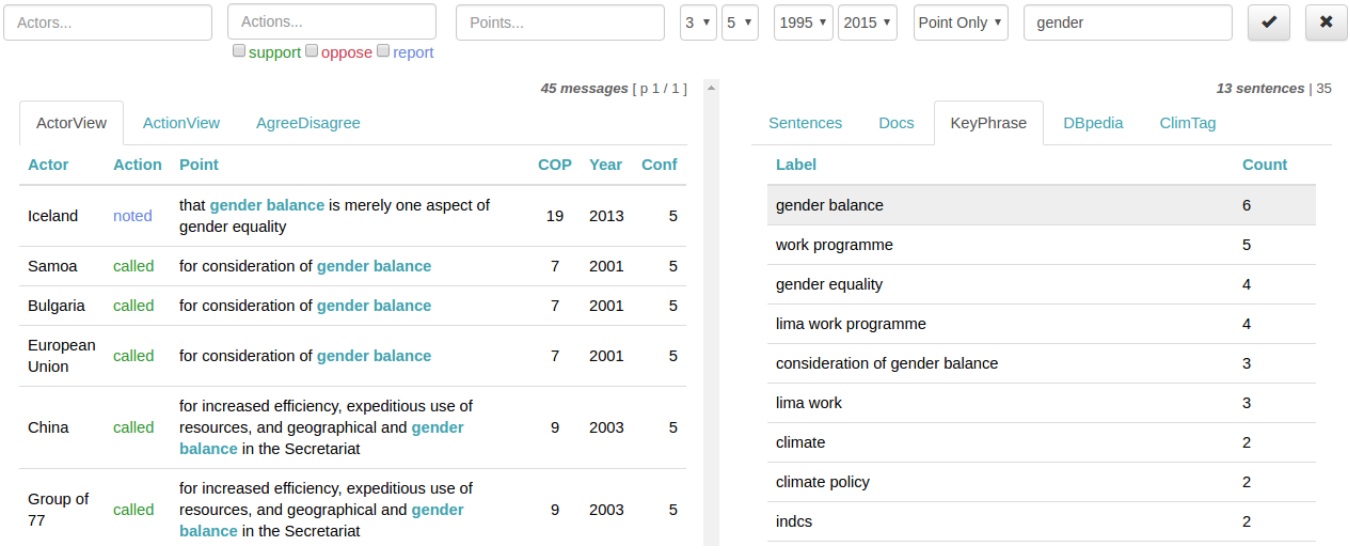
3.4 Interface de navigation

L'interface permet aux chercheurs d'explorer les positions des pays dans la négociation et de les comparer en fonction des termes extraits dans leurs messages, ainsi que des concepts de DBpedia et des concepts du thesaurus du domaine (le *Climate Thesaurus*). La vue principale de l'interface est affichée dans la figure 11.[45]

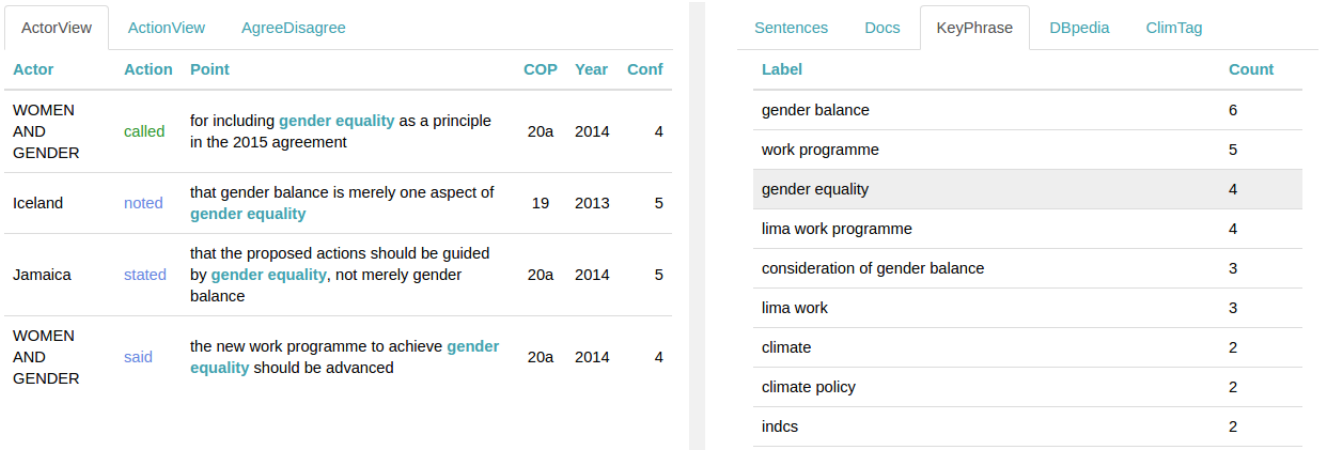
Les propositions contenant un certain acteur ou un prédicat donné sont recherchées respectivement avec les champs de recherche *Actors* (« acteurs ») et *Actions* (« actions »). Les propositions correspondant à la requête sont affichées sur le panneau de gauche, et sur le panneau de droite on peut voir les phrases et les documents dont elles ont été extraites. Une recherche sur le texte intégral peut être effectuée avec le champ de recherche *Free text* (« texte libre »). Pour les requêtes en texte libre, les phrases correspondant à la requête sont affichées sur le panneau de droite, et les propositions qui ont été annotées par le système dans ces phrases sont affichées sur le panneau de gauche.

Les onglets correspondant aux termes (*KeyPhrase*), aux concepts *DBpedia* et aux concepts du Climate Thesaurus (onglet *ClimTag*) sur le panneau de droite donnent une vue d'ensemble du contenu des propositions correspondant à une requête.

[45] L'interface se trouve sur `http://apps.lattice.cnrs.fr/ie/uidev/`

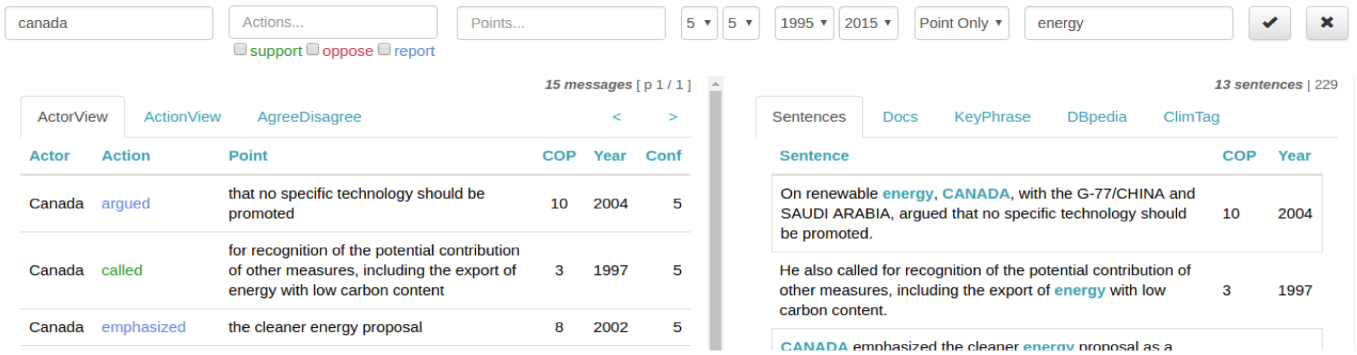


A. Propositions contenant *gender balance* (« équilibre des genres »)

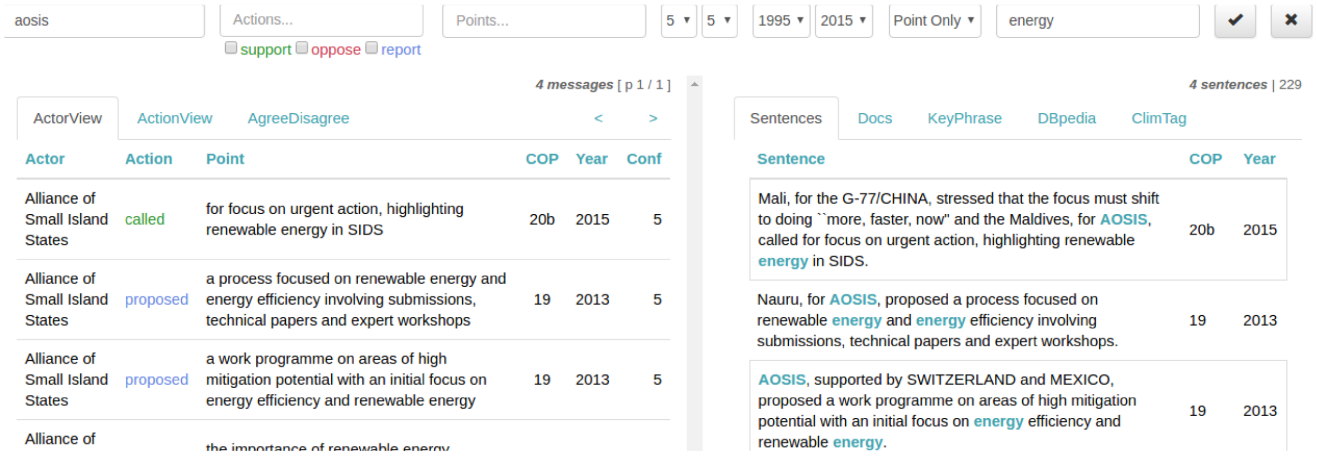


B. Propositions contenant *gender equality* (« égalité de genre »)

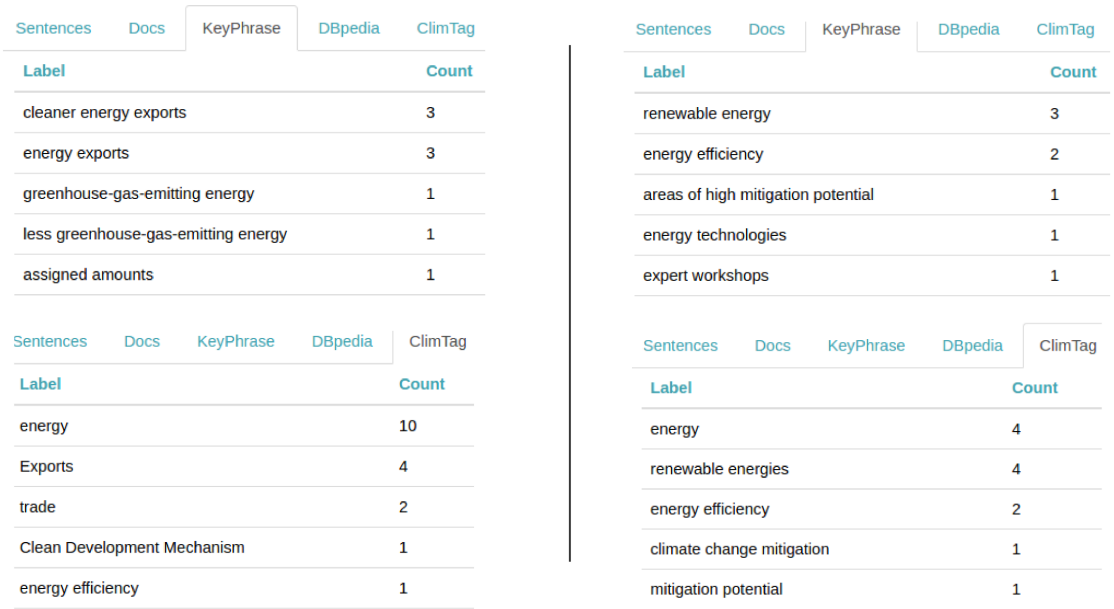
FIGURE 12 – Interface ENB : Le mot *gender* (« genre ») a été recherché dans le corpus, avec le champ *Free text* (« Texte libre »). Le volet de gauche montre les propositions extraites où *gender* a été trouvé. Dans le volet de droite, les termes de ces propositions s'affichent. Dans l'image supérieure, les propositions correspondant à *gender balance* (« équilibre des genres ») sont sélectionnées. Dans l'image inférieure, les propositions correspondant à *gender equality* (« égalité de genre ») ont été choisies. Avec l'aperçu des messages contenant le terme visé, nous voyons que certains pays, outre des acteurs non gouvernementaux comme *Women and Gender* font une déclaration plus forte que d'autres autour du genre, parlant d'égalité plutôt que d'équilibre.



A1. Requête avec *Canada* comme acteur et *energy* (« *energie* ») dans le texte des phrases. Affichage du panneau des phrases.



B1. Requête avec *AOSIS* comme acteur et *energy* (« *energie* ») dans le texte des phrases. Affichage du panneau des phrases.



A2. *Canada* & *energy*:
Motifs (haut) et Climate Thesaurus (bas)

B2. *AOSIS* & *energy*:
Motifs (haut) et Climate Thesaurus (bas)

FIGURE 13 – Interface pour le corpus ENB : Comparaison des déclarations de deux acteurs sur l'énergie (*energy*) à travers les termes extraits et les concepts du thesaurus. **A1** et **B1** montrent des requêtes qui renvoient les propositions dont l'acteur est le Canada (A1) ou l'AOSIS (B1), extraites des phrases contenant le mot *energy*. **A2** et **B2** montrent les 5 concepts du Climate Thesaurus les plus fréquents dans le messages des propositions (A2 pour le Canada, B2 pour l'AOSIS). Nous voyons que les deux pays parlent d'*energy efficiency* (« efficacité énergétique »), mais seul le Canada parle de *energy exports* (« exportation d'énergie »), peut-être parce qu'il possède de nombreuses ressources énergétiques renouvelables.

En résumé, les utilisateurs peuvent rechercher des propositions émises par un acteur donné, via un prédicat donné ou pour un type de prédicat donné, et contenant les termes de la requête. Des métadonnées (termes du thesaurus et concepts DBpedia) sont affichées à côté des propositions correspondant à une requête, pour fournir un aperçu du contenu des messages des propositions. Cela peut aider à répondre à des questions comme les suivantes :

- Quels termes ou concepts sont mentionnés dans les propositions où un acteur (par exemple *la Chine*) utilise un prédicat d'opposition ?
- Quels acteurs mentionnent un concept donné (par exemple, *droits humains*) et avec quels verbes ?
- Quels autres termes ou concepts DBpedia sont trouvés dans le contexte des termes de la requête ?

Des captures d'écran montrant les types de requêtes possibles avec l'interface sont fournies ci-dessous. La figure 12 montre les différents acteurs qui ont fait des déclarations sur le genre, ainsi que les sujets abordés dans ces déclarations. La figure 13 compare les déclarations du Canada et de la Chine concernant l'énergie.

En plus de la vue principale de l'interface qui vient d'être décrite, l'application a également une vue appelée *AgreeDisagree* (qu'on pourrait traduire par « AccordDésaccord »), et que l'on peut voir sur la figure 14. Cette vue permet de sélectionner deux acteurs et affiche ensuite les termes, concepts DBpedia et concepts Climate Thesaurus dans les propositions où ces acteurs sont d'accord ou en désaccord. L'accord et le désaccord ont été déterminés en fonction d'indices explicites, par exemple la séquence *opposed by* (« opposé par ») dans une phrase comme *Actor A, auquel s'est opposé Actor B ...*.[46]

L'export de résultats n'est actuellement pas possible avec l'interface ; il serait utile comme travail futur d'implémenter cette fonction.

3.5 Évaluation par les experts du domaine

Les objectifs de l'application étaient tout d'abord de donner un aperçu du corpus, aidant à répondre à des questions comme celles énumérées dans la section précédente. Idéalement, l'outil devrait aussi aider un expert du domaine à acquérir de nouvelles connaissances sur le corpus, par exemple en l'aidant à développer des idées de recherche nouvelles.

Une expérience avec des experts du domaine a été menée pour évaluer dans quelle mesure l'application atteint ces objectifs. L'interface a été évaluée par trois experts du domaine qui avaient déjà publié des recherches sur ce corpus.

[46] *L'acteur A, opposé par l'acteur B ...*

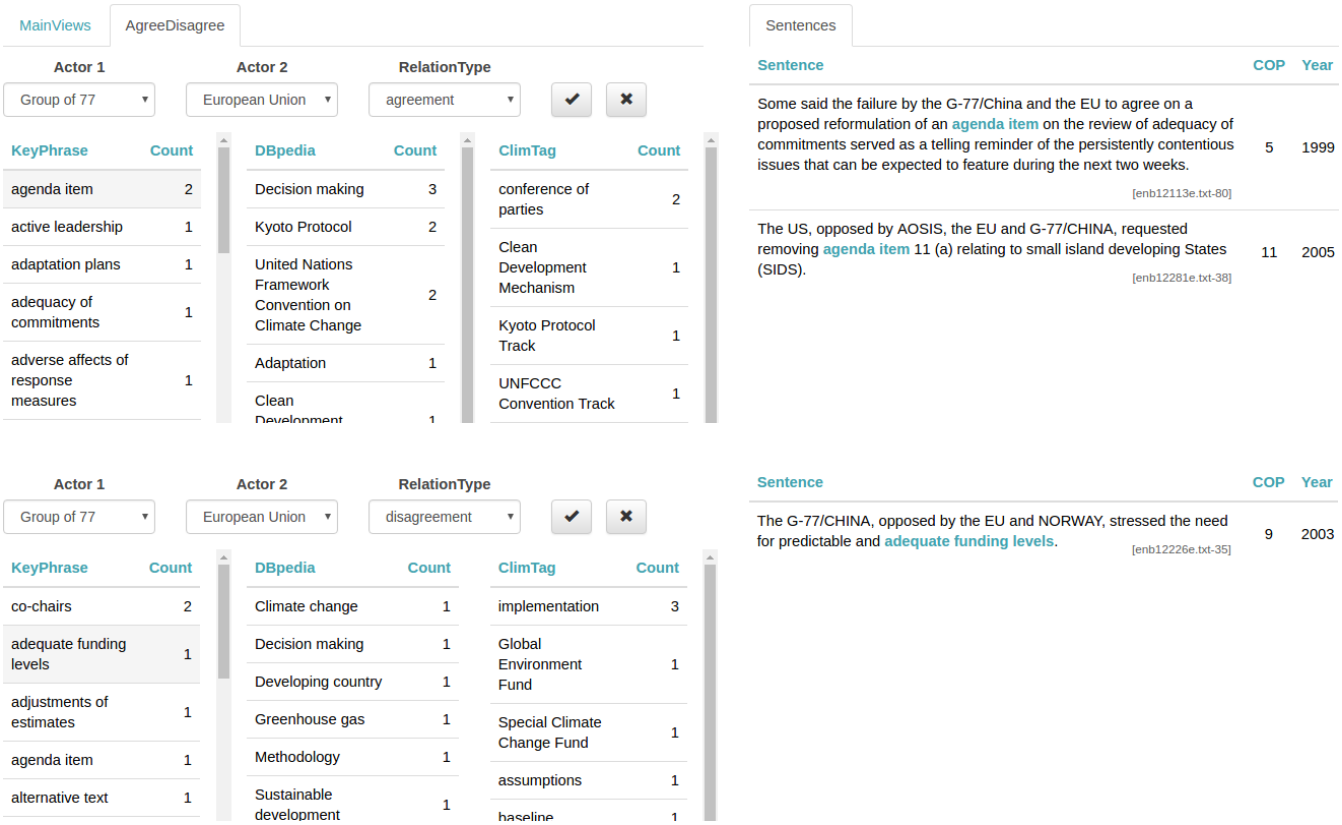


FIGURE 14 – Accord et désaccord entre les acteurs *European Union* (« Union européenne », un groupe de pays développés) et le *Group of 77* (« Groupe des 77 », un groupe de pays en développement). Les colonnes du volet de gauche montrent les métadonnées extraites des messages des propositions pour lesquels ces acteurs étaient en accord (*agreement*) **(haut)** ou désaccord (*disagreement*) **(bas)**. En cliquant sur les métadonnées, on affiche les phrases d'où ces propositions sont extraites. Concernant le désaccord, nous voyons que le financement des politiques climatiques est une des questions où les deux acteurs ont des opinions opposées : le motif *adequate funding levels* (« niveaux de financement adéquats ») et certains concepts du thesaurus climatique dans la colonne *ClimTag* font référence au financement.

Concernant l'aperçu du corpus montré par l'interface, les experts ont apprécié les possibilités de navigation fournies par une recherche différenciée pour chaque élément des propositions, par exemple l'aide offerte par l'interface pour examiner le comportement d'un acteur dans le corpus. Un des experts a souligné que des possibilités d'agrégation globales de résultats seraient aussi utiles : par exemple, en plus d'afficher les propositions où *le Canada* est l'acteur, afficher également combien de fois l'acteur utilise chaque prédicat. Ou pour un certain prédicat, comme *rejected* (« rejeté »), afficher combien de fois chaque acteur l'utilise—ceci pourrait par exemple donner une indication des pays qui ont une tendance plutôt défensive. Les données pour créer ces agrégations peuvent être obtenues manuellement à partir de l'interface actuelle, mais cela demande un effort à l'utilisateur qui pourrait être automatisé. Ceci serait assez simple à mettre en place à partir de l'application actuelle.

L'évaluation a aussi montré que l'application peut aider les experts dans leurs recherches. Ainsi, les experts ont apprécié de voir des déclarations

d'acteurs qui ne sont pas fréquemment étudiés et qui ont peu couverts par des travaux antérieurs sur le corpus, comme par exemple les organisations de peuples autochtones ou les groupes comme Women and Gender (« Femmes et Genre »). Les recherches antérieures avaient essentiellement couvert une liste prédéfinie de pays et groupes de pays (p. 289). A l'inverse, dans notre application, outre les pays et leurs groupes, tout acteur agent d'un verbe de parole ou de prise de position est susceptible d'être annoté en tant qu'acteur de la négociation. Une experte a mentionné que l'analyse détaillée des phrases du corpus, articulées comme des triplets $\langle$*acteur, prédicat, message*$\rangle$ est utile pour la comparaison des pays selon différentes facettes. Elle a suggéré qu'une dimension possible, à laquelle elle n'avait pas pensée avant, est la dimension formelle et juridique de l'argumentation de certains acteurs, au détriment du thème de la gestion du changement climatique lui-même. Enfin, l'un des experts a suggéré que l'information d'accord/désaccord entre les acteurs pourrait être utilisée pour créer des réseaux sémantiques. C'est effectivement une perspective simple à mettre en œuvre, sur le modèle de ce que nous avons fait pour le corpus Bentham par exemple.

3.6 Conclusions et perspectives

L'application permettant de naviguer dans les données du *Bulletin des Négociations de la Terre* montre comment les analyses TAL (impliquant dans ce cas des dépendances syntaxiques et des rôles sémantiques) permettent d'obtenir des résultats plus structurés qu'une simple analyse de cooccurrences. Les commentaires des experts sur l'application ont suggéré qu'elle est pertinente pour eux, et nous avons pu mettre au jour des exemples où de connaissances nouvelles ont été identifiées grâce à l'utilisation de notre application. Les experts ont aussi mentionné de possibles améliorations de l'interface de navigation, comme par exemple l'ajout d'une fonction pour exporter les résultats ou des possibilités d'agrégations additionnelles sur les données extraites.

4 Conclusions

La thèse a examiné la façon dont plusieurs technologies de Traitement automatique des langues (TAL) peuvent aider à accéder aux informations pertinentes dans de grands corpus textuels. Deux technologies, le liage d'entités (Entity Linking) et l'extraction de termes, ont été utilisées afin d'annoter les acteurs et les concepts en question dans les corpus. Des méthodes d'extraction de relations ont été employées pour déterminer comment ces acteurs et ces concepts sont reliés les uns aux autres. Les annotations TAL ont été intégrées dans des applications de navigation en corpus, qui combinent la

recherche en texte intégral, les réseaux sémantiques, et la recherche structurée basée sur les annotations. Comme la qualité des résultats du TAL varie selon le corpus, il a été nécessaire d'effectuer certains développements afin d'adapter les outils aux corpus d'application.

Les limites de l'annotation de concepts avec des bases de connaissances de domaine général ont été discutées. Du fait de ces limites il est utile de compléter les analyses génériques par des bases de connaissances spécifiques au domaine ou par des méthodes fondées sur des données (*data-driven*), telles que l'extraction de termes et les modèles de similarité distributionnelle.

La thèse complète porte sur trois études de cas illustrant notre utilisation de l'extraction de termes, du liage d'entités, de l'extraction de relations, ainsi que les développements que nous avons effectués pour adapter les outils aux corpus étudiés. Dans ce résumé en français, nous nous sommes concentrés sur deux études de cas simplement : les manuscrits de Jeremy Bentham et le *Bulletin des Négociations de la Terre*. Dans le premier cas, nos analyses du corpus sont fondées sur la cooccurrence de termes et sur des réseaux de concepts créés à partir de ceux-ci. Pour la deuxième application, les méthodes d'extraction des relations ont été utilisées afin de fournir des informations précises sur les liens entre acteurs, concepts et prises de position. S'agissant d'un corpus de négociations diplomatiques, l'extraction de relations a été appliquée pour établir comment certains acteurs se situent par rapport aux autres, suivant des relations de soutien et d'opposition.

Les interfaces utilisateur que nous avons développées ont été évaluées positivement par des entretiens avec des experts. L'interface pour le corpus Bentham fournit un aperçu adéquat des sujets abordés dans le corpus. En utilisant l'interface pour le *Bulletin des Négociations de la Terre*, les experts ont été convaincus de l'intérêt de l'application pour leurs recherches, ce qui constitue une bonne validation de notre travail.

Bibliography

Agerri, Rodrigo, Josu Bermudez, and German Rigau (2014). "IXA pipeline: Efficient and Ready to Use Multilingual NLP tools." *LREC: The 9th Language Resources and Evaluation Conference*, pp. 3823–3828. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/775_Paper.pdf.

Agichtein, Eugene and Luis Gravano (2000). "Snowball: Extracting relations from large plain-text collections". *Proceedings of the fifth ACM conference on Digital libraries*. ACM, pp. 85–94. URL: http://dl.acm.org/citation.cfm?id=336644.

Agirre, Eneko and Philip Edmonds (2007). *Word sense disambiguation: Algorithms and applications*. Vol. 33. Springer Science & Business Media.

Aguilar, Jacqueline, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis (2014). "A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards". *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pp. 45–53. URL: http://anthology.aclweb.org/W/W14/W14-29.pdf#page=55.

Ahn, David (2006). "The stages of event extraction". *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*. Association for Computational Linguistics, pp. 1–8. URL: http://dl.acm.org/citation.cfm?id=1629236.

Alex, Beatrice, Kate Byrne, Claire Grover, and Richard Tobin (2015). "Adapting the Edinburgh Geoparser for Historical Georeferencing". *International Journal of Humanities and Arts Computing* 9.1, pp. 15–35. ISSN: 1753-8548, 1755-1706. DOI: 10.3366/ijhac.2015.0136. URL: http://www.euppublishing.com/doi/10.3366/ijhac.2015.0136.

Artstein, Ron and Massimo Poesio (2008). "Inter-coder agreement for computational linguistics". *Computational Linguistics* 34.4, pp. 555–596. URL: http://dl.acm.org/citation.cfm?id=1479206.

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

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

Baerg, Nicole Rae, Will Lowe, Simone Paolo Ponzetto, Heiner Stuckenschmidt, and Cäcilia Zirn (2014). "Estimating Central Bank Preferences".

NLP Unshared Task in PoliInformatics. URL: http://www.cs.cmu.edu/~nasmith/unshared2014/Baerg.pdf.

Baker, Collin F, Charles J Fillmore, and John B Lowe (1998). "The Berkeley Framenet Project". *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, pp. 86–90.

Baker, Collin, Michael Ellsworth, and Katrin Erk (2007). "SemEval'07 task 19: frame semantic structure extraction". *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics, pp. 99–104.

Bamman, David and Gregory Crane (2007). "The Latin Dependency Treebank in a cultural heritage digital library". *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pp. 33–40.

Bamman, David and Noah A. Smith. "Open Extraction of Fine-Grained Political Statements". *EMNLP*. URL: http://anthology.aclweb.org/D/D15/D15-1008.pdf.

Bamman, David, Ted Underwood, and Noah A. Smith (2014). "A Bayesian Mixed Effects Model of Literary Character." *ACL (1)*, pp. 370–379. URL: http://www.aclweb.org/anthology/P14-1035.

Banko, Michele, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni (2007). "Open information extraction for the web". *IJCAI*. Vol. 7, pp. 2670–2676. URL: http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-429.pdf.

Banko, Michele, Oren Etzioni, and Turing Center (2008). "The Tradeoffs Between Open and Traditional Relation Extraction." *ACL*. Vol. 8, pp. 28–36. URL: http://anthology.aclweb.org/P/P08/P08-1.pdf#page=72.

Bastian, Mathieu, Sebastien Heymann, Mathieu Jacomy, et al. (2009). "Gephi: an open source software for exploring and manipulating networks." *ICWSM* 8, pp. 361–362. URL: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/viewFile/154/1009/.

Bauer, Florian, Denise Recheis, and Martin Kaltenböck (2011). "Data. reegle. info–A new key portal for Open Energy Data". *Environmental Software Systems. Frameworks of eEnvironment*. Springer, pp. 189–194. URL: http://link.springer.com/chapter/10.1007/978-3-642-22285-6_21.

Baya-Laffite, Nicolas and Jean-Philippe Cointet (2016). "Mapping Topics in International Climate Negotiations: A Computer-Assisted Semantic Network Approach". *Innovative Methods in Media and Communication Research*. Ed. by Sebastian Kubitschko and Anne Kaun. DOI: 10.1007/978-3-319-40700-5_14. Cham: Springer International Publishing, pp. 273–291. URL:

http://link.springer.com/10.1007/978-3-319-40700-5_14.

Bentham, Jeremy (1968 – ongoing). *The Collected Works of Jeremy Bentham*. Ed. by P. Schofield, J.H. Burns, J.R. Dinwiddy, and F. Rosen (General editors).

Beretta, Francesco (2015). "Publishing and sharing historical data on the semantic web : the SyMoGIH project – symogih.org". *Workshop: Semantic Web Applications in the Humanities*. Göttingen Centre for Digital Humanities. Göttingen, Germany. URL: https://halshs.archives-ouvertes.fr/halshs-01136533.

Berry, David, ed. (2012). *Understanding Digital Humanities*. Palgrave.

Biemann, Chris, Gregory R. Crane, Christiane D. Fellbaum, and Alexander Mehler (2014). "Computational Humanities-bridging the gap between Computer Science and Digital Humanities (Dagstuhl Seminar 14301)". *Dagstuhl Reports* 4.7. URL: http://drops.dagstuhl.de/opus/volltexte/2014/4792/.

Björkelund, Anders, Bernd Bohnet, Love Hafdell, and Pierre Nugues (2010). "A high-performance syntactic and semantic dependency parser". *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*. Association for Computational Linguistics, pp. 33–36. URL: http://dl.acm.org/citation.cfm?id=1944293.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre (2008). "Fast unfolding of communities in large networks". *Journal of statistical mechanics: theory and experiment* 2008.10, P10008.

Bohnet, Bernd (2010). "Very high accuracy and fast dependency parsing is not a contradiction". *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, pp. 89–97. URL: http://dl.acm.org/citation.cfm?id=1873792.

Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor (2008). "Freebase: a collaboratively created graph database for structuring human knowledge". *Proceedings of the ACM SIGMOD international conference on Management of data*, pp. 1247–1250.

Bonial, Claire and Martha Palmer (2016). "Comprehensive and Consistent PropBank Light Verb Annotation". *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia.

Bordea, Georgeta, Kartik Asooja, Paul Buitelaar, and Leona O'Brien (2014). "Gaining insights into the Global Financial Crisis using Saffron". URL: http://www.academia.edu/download/33977225/saffronFinancialCrisis.pdf.

Borin, Lars, Dimitrios Kokkinakis, and Leif-Jöran Olsson (2007). "Naming the past: Named entity and animacy recognition in 19th century Swedish literature". *Proceedings of the Workshop on Language Technology for Cultural*

Heritage Data (LaT-eCH 2007), pp. 1–8. URL: http://www.anthology.aclweb.org/W/W07/W07-09.pdf#page=11.

Bosque, Ignacio (2001). "On the weight of light predicates". *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, pp. 23–38.

Bourreau, Pierre and Thierry Poibeau (2014). "Mapping the Economic Crisis: Some Preliminary Investigations". *NLP Unshared Task in PoliInformatics*. URL: http://arxiv.org/abs/1406.4211.

Brando, Carmen, Francesca Frontini, and Jean-Gabriel Ganascia (2015). "Disambiguation of named entities in cultural heritage texts using linked data sets". *East European Conference on Advances in Databases and Information Systems*. Springer, pp. 505–514. URL: http://link.springer.com/chapter/10.1007/978-3-319-23201-0_51.

Brooke, Julian, Timothy Baldwin, and Adam Hammond. "Bootstrapped Text-level Named Entity Recognition for Literature". URL: http://people.eng.unimelb.edu.au/tbaldwin/pubs/acl2016-ner.pdf.

Buchholz, Sabine and Erwin Marsi (2006). "CoNLL-X shared task on multilingual dependency parsing". *Proceedings of the Tenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 149–164. URL: http://dl.acm.org/citation.cfm?id=1596305.

Bunescu, Razvan C. and Marius Pasca (2006). "Using Encyclopedic Knowledge for Named entity Disambiguation." *EACL*. Vol. 6. 00538, pp. 9–16. URL: http://www.cs.utexas.edu/~ml/papers/encyc-eacl-06.pdf.

Butt, Miriam (2010). "The light verb jungle: Still hacking away." *Complex predicates in cross-linguistic perspective*, pp. 48–78.

Carreras, Xavier and Lluís Màrquez (2005). "Introduction to the CoNLL-2005 shared task: Semantic role labeling". *Proceedings of the Ninth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 152–164. URL: http://dl.acm.org/citation.cfm?id=1706571.

Causer, T., J. Tonra, and V. Wallace (2012). "Transcription maximized; expense minimized? Crowdsourcing and editing The Collected Works of Jeremy Bentham". *Literary and Linguistic Computing* 27.2, pp. 119–137. ISSN: 0268-1145, 1477-4615. DOI: 10.1093/llc/fqs004. URL: http://llc.oxfordjournals.org/cgi/doi/10.1093/llc/fqs004.

Causer, Tim and Melissa Terras (2014a). "Crowdsourcing Bentham: beyond the traditional boundaries of academic history". *International Journal of Humanities and Arts Computing* 8.1, pp. 46–64.

– (2014b). "'Many hands make light work. Many hands together make merry work': Transcribe Bentham and crowdsourcing manuscript collections". *Crowdsourcing Our Cultural Heritage*, pp. 57–88.

Chang, Angel X, Valentin I Spitkovsky, Christopher D Manning, and Eneko Agirre (2016). "Evaluating the word-expert approach for Named-Entity Disambiguation". *arXiv preprint arXiv:1603.04767*.

Chavalarias, David and Jean-Philippe Cointet (2013). "Phylomemetic Patterns in Science Evolution—The Rise and Fall of Scientific Fields". *PLoS ONE* 8.2, e54847. DOI: 10.1371/journal.pone.0054847. URL: http://dx.doi.org/10.1371/journal.pone.0054847.

Chen, Zheng and Heng Ji (2009). "Language specific issue and feature exploration in Chinese event extraction". *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Association for Computational Linguistics, pp. 209–212. URL: http://dl.acm.org/citation.cfm?id=1620910.

Chinchor, Nancy A (1998). *Overview of muc-7/met-2*. Tech. rep. SCIENCE APPLICATIONS INTERNATIONAL CORP SAN DIEGO CA.

Chinchor, Nancy and Elaine Marsh (1998). "Muc-7 information extraction task definition". *Proceeding of the seventh message understanding conference (MUC-7), Appendices*, pp. 359–367. URL: http://www.aclweb.org/anthology/M/M98/M98-1027.pdf.

Choi, Jinho D, Joel R Tetreault, and Amanda Stent (2015). "It Depends: Dependency Parser Comparison Using A Web-based Evaluation Tool." *ACL (1)*, pp. 387–396.

Christensen, Janara, Stephen Soderland, Oren Etzioni, et al. (2011). "An analysis of open information extraction based on semantic role labeling". *Proceedings of the sixth international conference on Knowledge capture*. ACM, pp. 113–120. URL: http://dl.acm.org/citation.cfm?id=1999697.

Clark, Micah, Adam Dalton, Tomas By, Yorick Wilks, Samira Shaikh, Ching-Sheng Lin, and Tomek Strzalkowski (2014). "Influence and Belief in Congressional Hearings". *NLP Unshared Task in PoliInformatics*. URL: http://www.academia.edu/download/41731274/Influence_and_Belief_in_Congressional_He20160129-8815-rq3tdd.pdf.

Clement, Tanya, Sara Steger, John Unsworth, and Kirsten Uszkalo (2008). *How Not To Read A Million Books*. URL: http://people.brandeis.edu/~unsworth/hownot2read.html.

Coll Ardanuy, Mariona, Maarten van den Bos, and Caroline Sporleder (2016a). "You shall know people by the company they keep: person name disambiguation for social network construction". *LaTeCH 2016*, p. 63. URL: https://www.aclweb.org/anthology/W/W16/W16-21.pdf#page=75.

Coll Ardanuy, Mariona, Jürgen Knauth, Andrei Beliankou, Maarten van den Bos, and Caroline Sporleder (2016b). "Person-Centric Mining of Historical Newspaper Collections". *Research and Advanced Technology for Digital Libraries*. Springer, Cham, pp. 320–331. DOI: 10.1007/978-3-319-43997-6_25. URL: https://link.springer.com/chapter/10.1007/978-3-319-43997-6_25.

Coll Ardanuy, Mariona and Caroline Sporleder (2015). "Clustering of Novels Represented as Social Networks". *LiLT (Linguistic Issues in Language Technology)* 12. URL: http://csli-lilt.stanford.edu/ojs/index.php/LiLT/article/view/60.

Conde, Angel, Mikel Larrañaga, Ana Arruarte, Jon A. Elorriaga, and Dan Roth (2016). "litewi: A combined term extraction and entity linking method for eliciting educational ontologies from textbooks". *Journal of the Association for Information Science and Technology* 67.2, pp. 380–399. ISSN: 2330-1643. DOI: 10.1002/asi.23398. URL: http://dx.doi.org/10.1002/asi.23398.

Cornolti, Marco (2012). "A Framework to compare text annotators and its applications". MSc Thesis. University of Pisa.

Cornolti, Marco, Paolo Ferragina, and Massimiliano Ciaramita (2013). "A framework for benchmarking entity-annotation systems". *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 249–260. URL: http://dl.acm.org/citation.cfm?id=2488411.

Covington, Michael A (1990). *A Dependency Parser for Variable-Word-Order Languages. Research Report AI–1990–01*. Artificial Intelligence Programs. The University of Georgia.

Cripps, Paul, Anne Greenhalgh, Dave Fellows, Keith May, and David Robinson (2004). "Ontological Modelling of the work of the Centre for Archaeology". *CIDOC CRM Technical Paper*.

Crofts, Nick, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff (2011). *Definition of the CIDOC conceptual reference model*. URL: http://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_5.0.4.pdf.

Cucerzan, Silviu (2007). "Large-Scale Named Entity Disambiguation Based on Wikipedia Data." *EMNLP-CoNLL*. Vol. 7. Citeseer, pp. 708–716.

Cunningham, Hamish (2005). "Information extraction, automatic". *Encyclopedia of language and linguistics*, pp. 665–677. URL: http://www.academia.edu/download/27742540/preprint.pdf.

Cunningham, Hamish, Diana Maynard, Kalina Bontcheva, and Valentin Tablan (2002). "A framework and graphical development environment for robust NLP tools and applications." *ACL*, pp. 168–175.

Cybulska, Agata and Piek Vossen (2011). "Historical event extraction from text". *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Association for Computational Linguistics, pp. 39–43. URL: http://dl.acm.org/citation.cfm?id=2107642.

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

Das, Dipanjan, Nathan Schneider, Desai Chen, and Noah A Smith (2010). "Probabilistic frame-semantic parsing". *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*. Association for Computational Linguistics, pp. 948–956.

De La Clergerie, Éric Villemonte, Olivier Hamon, Djamel Mostefa, Christelle Ayache, Patrick Paroubek, and Anne Vilnat (2008). "Passage: from French parser evaluation to large sized treebank". *LREC: The 8th Language Resources and Evaluation Conference*, pp. 3570–3576. URL: http://repository.dlsi.ua.es/251/1/pdf/908_paper.pdf.

De Marneffe, Marie-Catherine and Christopher D. Manning (2008). "The Stanford typed dependencies representation". *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*. Association for Computational Linguistics, pp. 1–8. URL: http://dl.acm.org/citation.cfm?id=1608859.

Diesner, Jana (2012). "From Texts to Networks: Detecting and Managing the Impact of Methodological Choices for Extracting Network Data from Text Data". PhD thesis. Carnegie Mellon University. DOI: 10.1007/s13218-012-0225-0. URL: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA600419.

Dowty, David (1991). "Thematic Proto-Roles and Argument Selection". *Language* 67.3, p. 547. ISSN: 00978507. DOI: 10.2307/415037. URL: http://www.jstor.org/stable/415037?origin=crossref.

Ehrmann, Maud, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan (2016). "Diachronic Evaluation of NER Systems on Old Newspapers". *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*. Bochumer Linguistische Arbeitsberichte, pp. 97–107. URL: https://infoscience.epfl.ch/record/221391.

Ellis, Joe, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie Strassel (2015). "Overview of linguistic resources for the TAC KBP 2015 evaluations: Methodologies and results". *Proceedings of TAC KBP 2015 Workshop, National Institute of Standards and Technology*, pp. 16–17.

URL: http://tac.nist.gov/publications/2015/additional.papers/TAC2015.KBP_resources_overview.proceedings.pdf.

Elson, David K., Nicholas Dames, and Kathleen R. McKeown (2010). "Extracting social networks from literary fiction". *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, pp. 138–147. URL: http://dl.acm.org/citation.cfm?id=1858696.

Erdmann, Alexander, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe (2016). "Challenges and Solutions for Latin Named Entity Recognition". *LT4DH 2016*, p. 85. URL: http://www.aclweb.org/anthology/W/W16/W16-40.pdf#page=97.

Erk, Katrin and Sebastian Pado (2006). "Shalmaneser–a toolchain for shallow semantic parsing". *Proceedings of LREC*. Vol. 6.

Etzioni, Oren, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam (2011). "Open Information Extraction: The Second Generation." *IJCAI*. Vol. 11, pp. 3–10. URL: http://www.cs.washington.edu/research/projects/aiweb/media/papers/etzioni-ijcai2011.pdf.

Ezzat, Mani (2014). "Acquisition de relations entre entités nommées à partir de corpus". PhD thesis. Institut national des langues et civilisations orientales. URL: http://www.theses.fr/2014INAL0008/document.

Fellbaum, Christiane, ed. (1998). *Wordnet: An electronic lexical database*.

Ferragina, Paolo and Ugo Scaiella (2010). "Tagme: On-the-fly annotation of short text fragments (by Wikipedia entities)". *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, pp. 1625–1628. URL: http://dl.acm.org/citation.cfm?id=1871689.

Fillmore, Charles J, Christopher R Johnson, and Miriam RL Petruck (2003). "Background to framenet". *International journal of lexicography* 16.3, pp. 235–250.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning (2005). "Incorporating non-local Information into Information Extraction Systems by Gibbs Sampling". *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 363–370.

Fiscus, Jonathan G. (1997). "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)". *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, pp. 347–354. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=659110.

Fitzpatrick, Kathleen (2010). *Reporting from the Digital Humanities 2010 Conference – ProfHacker - Blogs - The Chronicle of Higher Education*. URL: http://www.chronicle.com/blogs/profhacker/reporting-from-the-digital-humanities-2010-conference/25473.

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

Fokkens, Antske, Aitor Soroa, Zuhaitz Beloki, Niels Ockeloen, German Rigau, Willem Robert van Hage, and Piek Vossen (2014). "NAF and GAF: Linking linguistic annotations". *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pp. 9–16.

Frontini, Francesca, Carmen Brando, and Jean-Gabriel Ganascia (2015). "Domain-adapted named-entity linker using Linked Data". *Workshop on NLP Applications: Completing the Puzzle co-located with the 20th International Conference on Applications of Natural Language to Information Systems (NLDB 2015)*. Passau, Germany. URL: https://hal.archives-ouvertes.fr/hal-01203356.

Frontini, Francesca, Carmen Brando, Marine Riguet, Clémence Jacquot, and Vincent Jolivet (2016). "Annotation of Toponyms in TEI Digital Literary Editions and Linking to the Web of Data". *MATLIT: Materialidades da Literatura* 4.2, pp. 49–75. ISSN: 2182-8830. URL: http://iduc.uc.pt/index.php/matlit/article/view/2375.

Généreux, Michel (2007). "Cultural heritage digital resources: from extraction to querying". URL: http://eprints.brighton.ac.uk/3211/.

Gregory, Ian, Alistair Baron, David Cooper, Andrew Hardie, Patricia Murrieta-Flores, and Paul Rayson (2014). "Crossing Boundaries: Using GIS in Literary Studies, History and Beyond". *Collections électroniques de l'INHA. Actes de colloques et livres en ligne de l'Institut national d'histoire de l'art*. INHA. URL: https://inha.revues.org/4931.

Grimmer, J. and B. M. Stewart (2013). "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts". *Political Analysis* 21.3, pp. 267–297. ISSN: 1047-1987, 1476-4989. DOI: 10.1093/pan/mps028. URL: http://pan.oxfordjournals.org/cgi/doi/10.1093/pan/mps028.

Grishman, Ralph (2015). "Information Extraction". *Intelligent Systems, IEEE* 30.5, pp. 8–15. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7243219.

Grishman, Ralph and Beth Sundheim (1996). "Message Understanding Conference-6: A Brief History." *COLING*. Vol. 96, pp. 466–471. URL: http://www.alta.asn.au/events/altss_w2003_proc/altss/courses/molla/C96-1079.pdf.

Grover, Claire, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball (2010). "Use of the Edinburgh geoparser for georeferencing digitized historical collections". *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 368.1925, pp. 3875–3889. ISSN: 1364-503X, 1471-2962. DOI: 10.1098/rsta.2010.0149. URL: http://rsta.royalsocietypublishing.org/content/368/1925/3875.

Guibon, Gaël, Isabelle Tellier, Mathieu Constant, Sophie Prévost, and Kim Gerdes (2014). "Parsing poorly standardized language dependency on Old French". *Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, pp. 51–61.

– (2015a). "Analyse syntaxique de l'ancien français: quelles propriétés de la langue influent le plus sur la qualité de l'apprentissage?" *TALN 22*.

Guibon, Gaël, Isabelle Tellier, Sophie Prévost, Mathieu Constant, and Kim Gerdes (2015b). "Searching for Discriminative Metadata of Heterogenous Corpora". *Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, pp. 72–82.

Hachey, Ben, Joel Nothman, and Will Radford. "Cheap and easy entity evaluation". URL: http://anthology.aclweb.org/P/P14/P14-2076.pdf.

Hajič, Jan, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. (2009). "The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages". *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, pp. 1–18. URL: http://dl.acm.org/citation.cfm?id=1596411.

Hasibi, Faegheh, Krisztian Balog, and Svein Erik Bratsberg. "On the Reproducibility of the TAGME Entity Linking System". URL: http://hasibi.com/files/ecir2016-ort.pdf.

Haug, Dag TT and Marius Jøhndal (2008). "Creating a parallel treebank of the Old Indo-European Bible translations". *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pp. 27–34.

Hearst, Marti A. (1997). "TextTiling: Segmenting text into multi-paragraph subtopic passages". *Computational linguistics* 23.1, pp. 33–64. URL: http://dl.acm.org/citation.cfm?id=972687.

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

Heiden, Serge, Jean-Philippe Magué, and Bénédicte Pincemin (2010). "TXM: Une plateforme logicielle open-source pour la textométrie-conception et développement". *10th International Conference on the Statistical Analysis of Textual Data-JADT 2010*. Vol. 2. 3. Edizioni Universitarie di Lettere Economia Diritto, pp. 1021–1032.

Heintz, Bettina (2007). "Zahlen, Wissen, Objektivität: Wissenschaftssoziologische Perspektiven". *Zahlenwerk*. Ed. by Andrea Mennicken and Hendrik Vollmer. Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 65–85. URL: http://link.springer.com/10.1007/978-3-531-90449-8_4.

Hobbs, Jerry R., Douglas Appelt, John Bear, David Israel, Megumi Kameyama, and Mabry Tyson (1993). "Fastus: A system for extracting information from text". *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, pp. 133–137. URL: http://dl.acm.org/citation.cfm?id=1075701.

Hoffart, Johannes, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum (2011). "Robust disambiguation of named entities in text". *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 782–792. URL: http://dl.acm.org/citation.cfm?id=2145521.

Hulden, Vilja (2016). "Whodunit... and to Whom? Subjects, Objects, and Actions in Research Articles on American Labor Unions". *LaTeCH 2016*, pp. 140–145. URL: http://www.aclweb.org/anthology/W/W16/W16-21.pdf#page=152.

Inglese, Guglielmo (2015). "Towards a Hittite Treebank. Basic Challenges and Methodological Remarks". *Corpus-Based Research in the Humanities (CRH)*.

Isaksen, Leif, Rainer Simon, Elton T.E. Barker, and Pau de Soto Cañamares (2014). "Pelagios and the Emerging Graph of Ancient World Data". *Proceedings of the 2014 ACM Conference on Web Science*. WebSci '14. Bloomington, Indiana, USA: ACM, pp. 197–201. DOI: 10.1145/2615569.2615693. URL: http://doi.acm.org/10.1145/2615569.2615693.

Jacomy, Mathieu, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian (2014). "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software". *PloS one* 9.6, e98679.

Ji, Heng, H. T. Dang, J. Nothman, and B. Hachey (2014). "Overview of TAC-KBP 2014 Entity Discovery and Linking Tasks". *Proc. Text Analysis Conference (TAC2014)*. URL: http://nlp.cs.rpi.edu/paper/edl2014overview.pdf.

Ji, Heng, Joel Nothman, Ben Hachey, and Radu Florian (2015). "Overview of TAC-KBP2015 Tri-lingual Entity Discovery and Linking". *Text Analysis*

Conference. URL: https://pdfs.semanticscholar.org/955a/78a8a5e4e31d10ffc827f365bd4c4f30d563.pdf.

Jing, Yufeng and W Bruce Croft (1994). "An association thesaurus for information retrieval". *Intelligent Multimedia Information Retrieval Systems and Management-Volume 1*. Le Centre de Hautes Etudes Internationales d'Informatique Documentaire, pp. 146–160.

Johansson, Richard and Pierre Nugues (2007). "LTH: semantic structure extraction using nonprojective dependency trees". *Proceedings of the 4th international workshop on semantic evaluations*. Association for Computational Linguistics, pp. 227–230.

– (2008). "Dependency-based syntactic-semantic analysis with PropBank and NomBank". *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 183–187. URL: http://dl.acm.org/citation.cfm?id=1596355.

Jurafsky, Dan and James H Martin (2009). *Speech and language processing*. Pearson.

Kelly, Diane (2007). "Methods for Evaluating Interactive Information Retrieval Systems with Users". *Foundations and Trends in Information Retrieval* 3.1—2, pp. 1–224. ISSN: 1554-0669, 1554-0677. DOI: 10.1561/1500000012. URL: http://www.nowpublishers.com/article/Details/INR-012.

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

Kim, Su Nam, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin (2010). "Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles". *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pp. 21–26. URL: http://dl.acm.org/citation.cfm?id=1859668.

Kipper-Schuler, Karin (2005). "VerbNet: A broad-coverage, comprehensive verb lexicon". PhD thesis. University of Pennsylvania. URL: http://repository.upenn.edu/dissertations/AAI3179808/.

Klein, Ewan, Beatrice Alex, and Jim Clifford (2014). "Bootstrapping a historical commodities lexicon with SKOS and DBpedia". *Proceedings of the EACL LaTeCH Workshop*, pp. 13–21. URL: http://www.aclweb.org/anthology/W/W14/W14-06.pdf#page=23.

Kleinnijenhuis, Jan, Wouter van Atteveldt, and Antske Fokkens (2014). "Chicken or Egg? The reciprocal influence of press and politics". *NLP Unshared Task in PoliInformatics*. URL: http://www.cs.cmu.edu/~nasmith/unshared2014/Kleinnijenhuis.pdf.

Kulkarni, Sayali, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti (2009). "Collective annotation of Wikipedia entities in web text". *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 00202. ACM, pp. 457–466. URL: http://dl.acm.org/citation.cfm?id=1557073.

Latour, Bruno (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press.

Lauscher, Anne, Federico Nanni, Pablo Ruiz Fabo, and Simone Ponzetto (2016). "Entities as Topic Labels: Combining Entity Linking and Labeled LDA to Improve Topic Interpretability and Evaluability". *Italian Journal of Computational Linguistics, Special Issue on Digital Humanities and Computational Linguistics*. URL: http://www.ai-lc.it/IJCoL/v2n2/4-lauscher_et_al.pdf.

Law, John and John Hassard (1999). "Actor network theory and after".

LDC (2005). *ACE (Automatic Content Extraction) English Annotation Guidelines for Events. Version 5.4.3*. Linguistic Data Consortium. URL: https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf.

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky (2013). "Deterministic coreference resolution based on entity-centric, precision-ranked rules". *Computational Linguistics* 39.4, pp. 885–916.

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

Levy, Roger and Galen Andrew (2006). "Tregex and Tsurgeon: tools for querying and manipulating tree data structures". *Proceedings of the fifth international conference on Language Resources and Evaluation*. Citeseer, pp. 2231–2234.

Li, Qi, Heng Ji, Yu Hong, and Sujian Li (2014a). "Constructing Information Networks Using One Single Model." *EMNLP*, pp. 1846–1851. URL: http://www.aclweb.org/anthology/D/D14/D14-1198.pdf.

Li, Qi, Heng Ji, and Liang Huang (2013). "Joint Event Extraction via Structured Prediction with Global Features." *ACL (1)*, pp. 73–82. URL: http://anthology.aclweb.org/P/P13/P13-1008.pdf.

Li, William P., David Larochelle, and Andrew W. Lo (2014b). "Estimating Policy Trajectories During the Financial Crisis". *NLP Unshared Task in PoliInformatics*. URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2447293.

Lieberman, Henry. *The Tyranny of Evaluation*. URL: http://web.media.mit.edu/~lieber/Misc/Tyranny-Evaluation.html.

López de Lacalle, Maddalen, Egoitz Laparra, Itziar Aldabe, and German Rigau (2016). "A Multilingual Predicate Matrix". *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA).

López de Lacalle, Maddalen, Egoitz Laparra, and German Rigau (2014). "Predicate Matrix: extending SemLink throughWordNet mappings". *The 9th edition of the Language Resources and Evaluation Conference. Reykjavik, Iceland*. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/589_Paper.pdf.

Makhoul, John, Francis Kubala, Richard Schwartz, Ralph Weischedel, et al. (1999). "Performance measures for information extraction". *Proceedings of DARPA broadcast news workshop*, pp. 249–252.

Manning, Christopher D, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky (2014). "The Stanford CoreNLP Natural Language Processing toolkit." *ACL (System Demonstrations)*, pp. 55–60.

Manning, Christopher D, Prabhakar Raghavan, Hinrich Schütze, et al. (2008). *Introduction to information retrieval*. Vol. 1. 1. Cambridge university press Cambridge.

Marciniak, Daniel (2016). "Computational text analysis: Thoughts on the contingencies of an evolving method". *Big Data & Society* 3.2. ISSN: 2053-9517. DOI: 10.1177/2053951716670190. URL: http://bds.sagepub.com/content/3/2/2053951716670190.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini (1993). "Building a Large Annotated Corpus of English: The Penn Treebank". *Comput. Linguist.* 19.2, pp. 313–330. ISSN: 0891-2017. URL: http://dl.acm.org/citation.cfm?id=972470.972475.

Al-Maskari, Azzah and Mark Sanderson (2010). "A review of factors influencing user satisfaction in information retrieval". *Journal of the American Society for Information Science and Technology* 61.5, pp. 859–868. URL: http://onlinelibrary.wiley.com/doi/10.1002/asi.21300/full.

Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni (2012). "Open language learning for information extraction". *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pp. 523–534. URL: http://dl.acm.org/citation.cfm?id=2391009.

McGillivray, Barbara, Marco Passarotti, and Paolo Ruffolo (2009). "The Index Thomisticus Treebank Project: Annotation, Parsing and Valency Lexicon." *TAL* 50.2, pp. 103–127.

Meeks, Elijah and Scott B. Weingart (2012). "The Digital Humanities Contribution to Topic Modeling". *Journal of Digital Humanities* 2.1. URL: http:

//journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/.

Mélanie, Frédérique, Johan Ferguth, Katherine Gruel, and Thierry Poibeau (2015). "Archaeology in the Digital Age: From Paper to Databases". *Digital Humanities 2015*.

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

Mesquita, Filipe (2015). "Extracting Information Networks from Text". PhD thesis. University of Alberta. URL: https://era.library.ualberta.ca/public/view/item/uuid:bde10153-7348-4d37-8747-a3314b936afc/DS2/de_Sa_Mesquita_Filipe_201503_PhD.pdf.

Mesquita, Filipe, Jordan Schmidek, and Denilson Barbosa (2013). "Effectiveness and Efficiency of Open Relation Extraction". *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 447–457. URL: http://www.aclweb.org/anthology/D13-1043.

Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman (2004). "The NomBank project: An interim report". *HLT-NAACL 2004 workshop: Frontiers in corpus annotation*, pp. 24–31. URL: http://www.aclweb.org/website/old_anthology/W/W04/W04-2705.pdf.

Mihalcea, Rada and Andras Csomai (2007). "Wikify!: linking documents to encyclopedic knowledge". *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 00508. ACM, pp. 233–242. URL: http://dl.acm.org/citation.cfm?id=1321475.

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

Miller, John E. and Kathleen F. McCoy (2014). "Changing Focus of the FOMC Through the Financial Crisis". *NLP Unshared Task in PoliInformatics*. URL: http://www.academia.edu/download/34123959/fomc.pdf.

Milne, David and I. Witten (2008a). "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links". *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pp. 25–30. URL: http://www.aaai.org/Papers/Workshops/2008/WS-08-15/WS08-15-005.pdf.

Milne, David and Ian H. Witten (2008b). "Learning to link with wikipedia". *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, pp. 509–518. URL: http://dl.acm.org/citation.cfm?id=1458150.

Mitamura, Teruko, Zhengzhong Liu, and Eduard Hovy (2015). "Overview of TAC KBP 2015 Event Nugget Track". *Text Analysis Conference*. URL: http://cairo.lti.cs.cmu.edu/kbp/2015/event/Mitamura,%20Liu,%20Hovy%20-%202016%20-%20Overview%20of%20TAC%20KBP%202015%20Event%20Nugget%20Track.pdf.

Morales, Michelle, David Brizan, Hussein Ghaly, Thomas Hauner, Min Ma, and Andrew Rosenberg (2014). "Social Network Analysis in the EStimation of Bank Financial Strength During the Financial Crisis". *NLP Unshared Task in PoliInformatics*.

Moretti, Franco (2005). *Graphs, maps, trees: abstract models for a literary history*. Verso.

Moretti, Giovanni, Rachele Sprugnoli, Stefano Menini, and Sara Tonelli (2016). "ALCIDE: Extracting and visualising content from large document collections to support humanities studies". *Knowledge-Based Systems* 111, pp. 100–112. ISSN: 0950-7051. DOI: http://dx.doi.org/10.1016/j.knosys.2016.08.003. URL: http://www.sciencedirect.com/science/article/pii/S0950705116302635.

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

Moro, Andrea and Roberto Navigli (2015). "SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking". *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics, pp. 288–297. URL: http://www.aclweb.org/anthology/S15-2049.

Moro, Andrea, Alessandro Raganato, and Roberto Navigli (2014). "Entity Linking meets Word Sense Disambiguation: A Unified Approach". *Transactions of the Association for Computational Linguistics* 2. 00014. URL: http://www.transacl.org/wp-content/uploads/2014/05/54.pdf.

Nadeau, David and Satoshi Sekine (2007). "A survey of named entity recognition and classification". *Lingvisticae Investigationes* 30.1, pp. 3–26.

Nanni, Federico and Pablo Ruiz Fabo (2016). "Entities as topic labels: Improving topic interpretability and evaluability combining Entity Linking and Labeled LDA". *Digital Humanities Conference (DH 2016)*. Jagiellonian University & Pedagogical University, Kraków, Poland: Alliance of Digital Humanities Organizations (ADHO), pp. 632–635. URL: https://arxiv.org/abs/1604.07809.

Navigli, Roberto (2009). "Word sense disambiguation: A survey". *ACM Computing Surveys* 41.2, pp. 1–69. ISSN: 03600300. DOI: 10.1145/1459352.1459355. URL: http://portal.acm.org/citation.cfm?doid=1459352.1459355.

– (2012). "A quick tour of word sense disambiguation, induction and related approaches". *International Conference on Current Trends in Theory and Practice of Computer Science*. Springer, pp. 115–129. URL: http://link.springer.com/10.1007/978-3-642-27660-6_10.

Navigli, Roberto and Simone Paolo Ponzetto (2012). "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network". *Artificial Intelligence* 193, pp. 217–250. ISSN: 00043702. DOI: 10.1016/j.artint.2012.07.001. URL: http://linkinghub.elsevier.com/retrieve/pii/S0004370212000793.

NIST-ACE (2005). *The ACE 2005 (ACE05) Evaluation Plan*. URL: http://www.itl.nist.gov/iad/mig/tests/ace/2005/doc/ace05-evalplan.v3.pdf.

Nivre, Joakim, Johan Hall, and Jens Nilsson (2004). "Memory-Based Dependency Parsing". *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*. Ed. by Hwee Tou Ng and Ellen Riloff. Boston, Massachusetts, USA: Association for Computational Linguistics, pp. 49–56.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. (2016). "Universal dependencies v1: A multilingual treebank collection". *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 1659–1666.

Palmer, Martha (2009). "Semlink: Linking propbank, verbnet and framenet". *Proceedings of the Generative Lexicon Conference*. GenLex-09, 2009 Pisa, Italy, pp. 9–15.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury (2005). "The proposition bank: An annotated corpus of semantic roles". *Computational linguistics* 31.1, pp. 71–106. URL: http://dl.acm.org/citation.cfm?id=1122628.

Piatti, Barbara, Hans Rudolf Bär, Anne-Kathrin Reuschel, Lorenz Hurni, and William Cartwright (2009). "Mapping Literature: Towards a Geography of Fiction". *Cartography and Art*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–16. URL: http://link.springer.com/10.1007/978-3-540-68569-2_15.

Plank, Barbara (2016). "What to do about non-standard (or non-canonical) language in NLP". *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pp. 13–20. URL: http://arxiv.org/abs/1608.07836.

Poibeau, Thierry (2002). "Extraction d'information à base de connaissances hybrides". Thèse de doctorat dirigée par Kayser, Daniel Informatique Paris 13 2002. PhD thesis. Université Paris-Nord. URL: http://www.theses.fr/2002PA132001.

Poibeau, Thierry and Pablo Ruiz Fabo (2015). "Generating Navigable Semantic Maps from Social Sciences Corpora". *Digital Humanities Conference (DH 2015)*. Sydney, Australia: Alliance of Digital Humanities Organizations (ADHO). URL: https://arxiv.org/abs/1507.02020.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong (2013). "Towards Robust Linguistic Analysis using OntoNotes." *CoNLL*, pp. 143–152.

Pradhan, Sameer, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue (2011). "CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes". *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, pp. 1–27. URL: http://dl.acm.org/citation.cfm?id=2132937.

Prévost, Sophie and Achim Stein, eds. (2013). *Syntactic Reference Corpus of Medieval French (SRCMF)*. Lyon/Stuttgart: ENS de Lyon; Lattice, Paris; ILR University of Stuttgart. URL: http://srcmf.org.

Raganato, Alessandro, Jose Camacho-Collados, Antonio Raganato, and Yunseo Joung (2016). "Semantic Indexing of Multilingual Corpora and its Application on the History Domain". *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 140–147. URL: http://www.aclweb.org/anthology/W/W16/W16-40.pdf#page=152.

Rao, Delip, Paul McNamee, and Mark Dredze (2010). "Streaming cross document entity coreference resolution". *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, pp. 1050–1058.

– (2013). "Entity linking: Finding extracted entities in a knowledge base". *Multi-source, Multilingual Information Extraction and Summarization*. Springer, pp. 93–115. URL: http://link.springer.com/chapter/10.1007/978-3-642-28569-1_5.

Ratinov, Lev, Dan Roth, Doug Downey, and Mike Anderson (2011). "Local and global algorithms for disambiguation to wikipedia". *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pp. 1375–1384. URL: http://dl.acm.org/citation.cfm?id=2002642.

Rayson, Paul (2008). "From key words to key semantic domains". *International Journal of Corpus Linguistics* 13.4, pp. 519–549. 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.

Řehůřek, Radim and Petr Sojka (2010). "Software Framework for Topic Modelling with Large Corpora". *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, pp. 45–50.

Reuschel, Anne-Kathrin and Lorenz Hurni (2011). "Mapping literature: Visualisation of spatial uncertainty in fiction". *The Cartographic Journal* 48.4, pp. 293–308.

Rieder, Bernhard and Theo Röhle (2012). "Digital methods: Five challenges". *Understanding digital humanities*. Ed. by David Berry. Palgrave, pp. 67–84.

Riloff, Ellen, Rosie Jones, et al. (1999). "Learning dictionaries for information extraction by multi-level bootstrapping". *AAAI/IAAI*, pp. 474–479. URL: http://www.aaai.org/Papers/AAAI/1999/AAAI99-068.pdf.

Rizzo, Giuseppe, Marieke van Erp, and Raphaël Troncy (2014). "Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web." *LREC, The 9th Language Resources and Evaluation Conference*, pp. 4593–4600.

Rizzo, Giuseppe and Raphaël Troncy (2012). "NERD: a framework for unifying named entity recognition and disambiguation extraction tools". *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 73–76. URL: http://dl.acm.org/citation.cfm?id=2380936.

Rose, Stuart, Dave Engel, Nick Cramer, and Wendy Cowley (2010). "Automatic keyword extraction from individual documents". *Text Mining: Applications and Theory*. Ed. by Michael W Berry and Jacob Kogan. Wiley, pp. 1–20.

Ruiz Fabo, Pablo, Clément Plancq, and Thierry Poibeau (2016a). "Climate Negotiation Analysis". *Digital Humanities Conference (DH 2016)*. Jagiellonian University & Pedagogical University, Kraków, Poland: Alliance of Digital Humanities Organizations (ADHO), pp. 663–666. URL: http://dh2016.adho.org/abstracts/81.

– (2016b). "More than Word Cooccurrence: Exploring Support and Opposition in International Climate Negotiations with Semantic Parsing". *LREC: The 10th Language Resources and Evaluation Conference*, pp. 1902–1907.

Ruiz Fabo, Pablo and Thierry Poibeau (2015a). "Combining Open Source Annotators for Entity Linking through Weighted Voting". *Joint Conference*

*on Lexical and Computational Semantics (*SEM 2015)*, pp. 211–215. URL: http://aclweb.org/anthology/S/S15/S15-1025.pdf.

Ruiz Fabo, Pablo and Thierry Poibeau (2015b). "EL92: Entity Linking Combining Open Source Annotators via Weighted Voting". *9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 355–359. URL: https://halshs.archives-ouvertes.fr/hal-01173968/document.

Ruiz Fabo, Pablo, Thierry Poibeau, and Frédérique Mélanie (2015c). "ELCO3: Entity Linking with Corpus Coherence Combining Open Source Annotators". *2015 Conference of the North American Chapter of the Association for Computational Linguistics–Human Language Technologies (NAACL HLT 2015)*. URL: https://aclweb.org/anthology/N/N15/N15-3010.pdf.

Rule, Alix, Jean-Philippe Cointet, and Peter S. Bearman (2015). "Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014". *Proceedings of the National Academy of Sciences* 112.35, pp. 10837–10844. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1512221112. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1512221112.

Salton, Gerard, Edward A. Fox, and Harry Wu (1983). "Extended Boolean Information Retrieval". *Commun. ACM* 26.11, pp. 1022–1036. ISSN: 0001-0782. DOI: 10.1145/182.358466. URL: http://doi.acm.org/10.1145/182.358466.

Salton, Gerard, Anita Wong, and Chung-Shu Yang (1975). "A vector space model for automatic indexing". *Communications of the ACM* 18.11, pp. 613–620. URL: http://dl.acm.org/citation.cfm?id=361220.

Salway, Andrew, Samia Touileb, and Endre Tvinnereim (2014). "Inducing Information Structures for Data-driven Text Analysis". *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science (LaTeCH)*, pp. 28–32. URL: http://acl2014.org/acl2014/W14-25/W14-25-2014.pdf#page=40.

Sanabila, Hadaiq Rolis and Ruli Manurung (2014). "Towards automatic wayang ontology construction using relation extraction from free text". *EACL 2014*, p. 128. URL: http://anthology.aclweb.org/W/W14/W14-06.pdf#page=138.

Schmid, Helmut (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees". *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Schreibman, Susan, Ray Siemens, and John Unsworth, eds. (2004). *A Companion to Digital Humanities*. Blackwell.

Schrodt, P. A. (2014). *TABARI: Textual Analysis by Augmented Replacement Instructions*. URL: http://eventdata.parusanalytics.com/software.dir/tabari.html.

Schrodt, Philip A., John Beieler, and Muhammed Idris (2014). "Three's a Charm?: Open Event Data Coding with EL:DIABLO, PETRARCH, and the Open Event Data Alliance". *ISA Annual Convention*. URL: http://parusanalytics.com/eventdata/papers.dir/Schrodt-Beieler-Idris-ISA14.pdf.

Schrodt, Philip A and David Van Brackle (2013). "Automated coding of political event data". *Handbook of computational approaches to counterterrorism*. Springer, pp. 23–49.

Scrivner, Olga and Sandra Kübler (2012). "Building an Old Occitan corpus via cross-Language transfer." *KONVENS*, pp. 392–400.

Sekine, Satoshi and Chikashi Nobata (2004). "Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy." *LREC, the 4th Language Resources and Evaluation Company*, pp. 1977–1980.

Smith, Noah A., Claire Cardie, Anne Washington, and John Wilkerson (2014). "Overview of the 2014 NLP Unshared Task in PoliInformatics". *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Baltimore, MD, USA: Association for Computational Linguistics, pp. 5–7. URL: http://www.aclweb.org/anthology/W14-2505.

Solan, Zach, David Horn, Eytan Ruppin, and Shimon Edelman (2005). "Unsupervised learning of natural languages". *Proceedings of the National Academy of Sciences of the United States of America* 102.33, pp. 11629–11634. URL: http://www.pnas.org/content/102/33/11629.short.

Song, Zhiyi, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma (2015). "From light to rich ERE: annotation of entities, relations, and events". *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*, pp. 89–98.

Sporleder, Caroline (2010). "Natural language processing for cultural heritage domains". *Language and Linguistics Compass* 4.9, pp. 750–768.

Stein, Achim (2014). "Parsing Heterogeneous Corpora with a Rich Dependency Grammar." *LREC*, pp. 2879–2886.

– (2016). "Old French Dependency Parsing: Results of Two Parsers Analysed from a Linguistic Point of View". *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA).

Straka, Milan, Jan Hajič, and Jana Straková (2016). "UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing". *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA).

Strube, Michael and Simone Ponzetto. "WikiRelate! Computing Semantic Relatedness Using Wikipedia". *Proceedings of the 21st National Conference*

on Artificial Intelligence, AAAI, pp. 1419–1424. URL: http://www.aaai.org/Papers/AAAI/2006/AAAI06-223.pdf.

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum (2007). "Yago: a core of semantic knowledge". *Proceedings of the 16th international conference on World Wide Web*. ACM, pp. 697–706. URL: http://dl.acm.org/citation.cfm?id=1242667.

Surdeanu, Mihai and Heng Ji (2014). "Overview of the english slot filling track at the tac2014 knowledge base population evaluation". *Proc. Text Analysis Conference (TAC2014)*. URL: http://clulab.cs.arizona.edu/papers/kbp2014_draft.pdf.

Surdeanu, Mihai, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre (2008). "The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies". *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 159–177. URL: http://dl.acm.org/citation.cfm?id=1596352.

Szpektor, Idan, Ido Dagan, Alon Lavie, Danny Shacham, and Shuly Wintner (2007). "Cross lingual and semantic retrieval for cultural heritage appreciation". *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pp. 65–72. URL: http://www.academia.edu/download/30790379/W07-09.pdf#page=75.

Taylor, Ann (2007). "The York—Toronto—Helsinki parsed corpus of Old English prose". *Creating and Digitizing Language Corpora*. Springer, pp. 196–227.

Taylor, Ann and Anthony S Kroch (1994). "The Penn-Helsinki Parsed Corpus of Middle English". *MS. University of Pennsylvania*.

Terras, Melissa, Julianne Nyhan, Edward Vanhoutte, et al., eds. (2013). *Defining Digital Humanities: A Reader*. Ashgate.

Tesnière, Lucien (1959). *Eléments de syntaxe structurale*. Librairie C. Klincksieck.

Tieberghien, Estelle, Frédérique Mélanie, Pablo Ruiz Fabo, Thierry Poibeau, Tim Causer, and Melissa Terras (2016). "Mapping the Bentham Corpus". *Digital Humanities Conference (DH 2016)*. Jagiellonian University & Pedagogical University, Kraków, Poland: Alliance of Digital Humanities Organizations (ADHO), pp. 279–282. URL: http://dh2016.adho.org/abstracts/372.

Ting, Kai Ming and Ian H. Witten (1997). "Stacked generalization: when does it work?" URL: http://www.cms.waikato.ac.nz/~ml/publications/1997/Ting-Witten-General97.pdf.

Tjong Kim Sang, Erik F. and Fien De Meulder (2003). "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition". *Proceedings of the seventh conference on Natural language learning*

at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, pp. 142–147. URL: http://dl.acm.org/citation.cfm?id=1119195.

Tong, Hanghang, Christos Faloutsos, and Jia-yu Pan (2006). "Fast Random Walk with Restart and Its Applications". *Sixth International Conference on Data Mining (ICDM'06)*. IEEE, pp. 613–622.

Toselli, A.H. and E. Vidal (2015). "Handwritten Text Recognition Results on the Bentham Collection with Improved Classical N-Gram-HMM methods". *International Workshop on Historical Document Imaging and Processing (HIP)*. ACM.

Traub, Myriam C. and Jacco van Ossenbruggen, eds. (2015). *Workshop on Tool Criticism in the Digital Humanities*. Centrum Wiskunde & Informatica, KNAW eHumanities, and Amsterdam Data Science Center. URL: http://persistent-identifier.org/?identifier=urn:nbn:nl:ui:18-23500.

Turney, Peter D (2000). "Learning algorithms for keyphrase extraction". *Information retrieval* 2.4, pp. 303–336.

Usbeck, Ricardo, Michael Röder, A-C Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, et al. (2015). "GERBIL-general entity annotation benchmark framework". *24th WWW conference*.

Vala, Hardik, David Jurgens, Andrew Piper, and Derek Ruths (2015). "Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On The Difficulty of Detecting Characters in Literary Texts". *Proceedings of Empirical Methods in Natural Language Processing*. URL: http://www.aclweb.org/anthology/D15-1088.

Van Atteveldt, Wouter (2008). *Semantic network analysis: techniques for extracting, representing and querying media content (PhD Dissertation)*. Charleston, SC: BookSurge.

Van Atteveldt, Wouter, Tamir Sheafer, Shaul R. Shenhav, and Yair Fogel-Dror (2017). "Clause Analysis: Using Syntactic Information to Automatically Extract Source, Subject, and Predicate from Texts with an Application to the 2008–2009 Gaza War". *Political Analysis*, pp. 1–16.

Van Erp, Marieke, Pablo Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Jörg Waitelonis (2016). "Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job". *10th International Conference on Language Resources and Evaluation (LREC)*. URL: http://jplu.github.io/publications/van_Erp_Plu-LREC2016.pdf.

Van Erp, MGJ, A Van den Bosch, S Wubben, S Hunt, P Lendvai, and L Borin (2009). "Instance-driven discovery of ontological relation labels". Association for Computational Linguistics.

Venturini, Tommaso, N. Baya Laffite, J.-P. Cointet, I. Gray, V. Zabban, and K. De Pryck (2014). "Three maps and three misunderstandings: A digital mapping of climate diplomacy". *Big Data & Society* 1.2. ISSN: 2053-9517. DOI: 10.1177/2053951714543804. URL: http://bds.sagepub.com/lookup/doi/10.1177/2053951714543804.

Venturini, Tommaso and Daniele Guido (2012). "Once Upon a Text: an ANT Tale in Text Analysis". *Sociologica* 6.3. [Note: "ANT" in the article title refers to a social science approach called "Actor–Network Theory (ANT)"]. URL: http://www.rivisteweb.it/doi/10.2383/72700.

Venturini, Tommaso, Benjamin Ooghe-Tabanou, Mathieu Jacomy, Paul Girard, Kari de Pryck, Gabriel Varela, Alex Constantin, Oleksii Boiarskyi, Karl Aberer, Alexis Jacomy, Thomas Dupeyrat, Thomas Busson, Léo Bonnargent, and Jérémy Lesceau. *Climate Negotiations Browser*. Ed. by Frédéric Mion. Sciences Po. URL: http://www.climatenegotiations.org.

Vieira, Rodrigo (2015). "Adapting State-of-the-Art Named Entity Recognition and Disambiguation Frameworks for Handling Clinical Text". MSc Thesis. Instituto Superior Técnico.

Vlachidis, Andreas (2012). "Semantic Indexing via Knowledge Organization Systems: Applying the CIDOC-CRM to Archaeological Grey Literature". PhD thesis. University of Glamorgan. URL: http://hypermedia.research.southwales.ac.uk/media/files/documents/2013-07-11/Andreas-Vlachidis_Thesis_print_ready.pdf.

Volk, Martin, Noah Bubenhofer, Adrian Althaus, Maya Bangerter, Lenz Furrer, and Beni Ruef (2010). "Challenges in Building a Multilingual Alpine Heritage Corpus." *LREC, Language Resources and Evaluation Conference*.

Vossen, Piek, German Rigau, Luciano Serafini, Pim Stouten, Francis Irving, and Willem Robert Van Hage (2014). "NewsReader: recording history from daily news streams". *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*. Reykjavik, Iceland. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/436_Paper.pdf.

Waitelonis, Jörg, Henrik Jürges, and Harald Sack. "Don't compare Apples to Oranges–Extending GERBIL for a fine grained NEL evaluation". URL: http://hpi.de/fileadmin/user_upload/fachgebiete/meinel/papers/Web_3.0/2016Waitelonis_SEMANTICS2016.pdf.

Walker, Christopher, Stephanie Strassel, Julie Medero, and Kazuaki Maeda (2005). *ACE 2005 Multilingual Training Corpus - Linguistic Data Consortium*. URL: https://catalog.ldc.upenn.edu/LDC2006T06.

Wang, Haochang, Tiejun Zhao, Hongye Tan, and Shu Zhang (2008). "Biomedical Named Entity Recognition Based on Classifiers Ensemble." *International Journal of Computer Science and Applications* 5.2, pp. 1–11. URL: http://www.tmrfindia.org/ijcsa/v5i21.pdf.

Wang, Lu, Parvaz Mahdabi, Joonsuk Park, Dinesh Puranam, Bishan Yang, and Claire Cardie (2014). "Cornell Expert Aided Query-focused Summarization (CEAQS): A Summarization Framework to PoliInformatics". *NLP Unshared Task in PoliInformatics*. URL: http://www.cs.cornell.edu/~luwang/papers/PoliInformatics.pdf.

Wang, Ting, Yaoyong Li, Kalina Bontcheva, Hamish Cunningham, and Ji Wang (2006). "Automatic extraction of hierarchical relations from text". *European Semantic Web Conference*. Springer, pp. 215–229.

Weeds, Julie and David Weir (2005). "Co-occurrence retrieval: A flexible framework for lexical distributional similarity". *Computational Linguistics* 31.4, pp. 439–475. URL: http://www.mitpressjournals.org/doi/abs/10.1162/089120105775299122.

Weischedel, Ralph, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue (2011). "OntoNotes: A large training corpus for enhanced processing". *Handbook of Natural Language Processing and Machine Translation. Springer*.

Wolpert, DH (1992). "Stacked generalization". *Neural Networks* 5, pp. 241–259. URL: http://www.cs.utsa.edu/~bylander/cs6243/wolpert92stacked.pdf.

Zirn, Cäcilia, Michael Schäfer, Michael Strube, Simone Paolo Ponzetto, and Heiner Stuckenschmidt (2014). "Exploring structural features for position analysis in political discussions". *NLP Unshared Task in PoliInformatics*. URL: http://www.computerlanguste.de/acl2014/UnsharedTaskTaskACL2014Zirn.pdf.

Résumé

La recherche en Sciences humaines et sociales repose souvent sur de grands corpus textuels, impossibles de lire en détail. Le Traitement automatique des langues (TAL) identifie des concepts et des acteurs importants dans un corpus et les relations entre eux, ce qui peut fournir une vue d'ensemble utile pour les experts d'un domaine, les aidant à identifier les zones du corpus pertinentes pour leurs recherches.

Pour annoter de grands corpus, nous avons appliqué le liage d'entités (Entity Linking), pour identifier des acteurs et concepts. Les relations entre ceux-ci ont été déterminées sur la base d'une chaîne de traitements TAL, qui étiquette des fonctions sémantiques et syntaxiques.

Des outils de TAL génériques ont été utilisés. L'efficacité des méthodes de TAL dépend du corpus, et des développements ont été effectués pour mieux s'adapter à nos corpus.

Trois corpus ont été analysés. D'abord, les manuscrits de Jeremy Bentham, un corpus de philosophie politique des 18^{e} et 19^{e} siècles. Ensuite, le corpus *PoliInformatics*, sur la crise financière américaine de 2007. Enfin, le *Bulletin des Négociations de la Terre* (*ENB*), qui couvre les sommets internationaux sur la politique climatique, où des traités comme le Protocole de Kyoto ont été négociés.

Des interfaces de navigation de corpus ont été développées, qui combinent les réseaux et la recherche structurée fondée sur des annotations TAL. Par exemple, l'interface *ENB* permet de voir les acteurs qui ont exprimé de l'opposition sur un sujet. Les relations entre acteurs et concepts sont exploitées, au-delà de la co-occurrence entre termes.

Les interfaces ont été évaluées par des experts de domaine. Nous avons tenté de déterminer si les experts peuvent avoir une meilleure compréhension du corpus grâce aux applications, en trouvant des faits nouveaux. Ceci a été attesté avec l'interface *ENB*, ce qui est une bonne validation du travail effectué.

Mots Clés

Liage d'entité, wikification, extraction de relations, extraction de propositions, visualisation de corpus, Traitement automatique des langues, Humanités numériques

Abstract

Social sciences and Humanities research is often based on large textual corpora, unfeasible to read in detail. Natural Language Processing (NLP) identifies important concepts and actors in a corpus, and the relations between them, which can provide a useful overview for domain-experts, helping identify corpus areas relevant for their research.

To annotate large corpora, we first applied Entity Linking, to identify corpus actors and concepts. The relations between these were determined based on an NLP pipeline, which provides semantic role labeling and syntactic dependencies among other information.

Generic NLP tools were used. As the efficacy of NLP methods depends on the corpus, some technological development was undertaken to better adapt to our corpora.

Three corpora were analyzed. First, the manuscripts of Jeremy Bentham (a 18th-19th century corpus in political philosophy). Second, the *PoliInformatics* corpus, about the American financial crisis of 2007. Third, the *Earth Negotiations Bulletin* (*ENB*), which covers international climate policy summits, where treaties like the Kyoto Protocol or the Paris Agreements get negotiated.

Corpus navigation interfaces were developed. They combine networks, full-text search and structured search based on NLP annotations. As an example, in the *ENB* corpus UI, negotiation actors having expressed support or opposition about a given issue can be searched. Relation information between actors and concepts is employed, beyond simple term co-occurrence.

The UIs were evaluated by domain-experts. We tried to determine whether experts could gain new insight on the corpus by using the applications, e.g if they found new evidence or research ideas. This was attested with the *ENB* interface, which is a good validation of the work carried out.

Keywords

Entity Linking, Wikification, Relation Extraction, Proposition Extraction, Corpus Visualization, Natural Language Processing, Digital Humanities