LREC 2018 Workshop

ParlaCLARIN:
Creating and Using Parliamentary Corpora

PROCEEDINGS

Edited by

Darja Fišer, Maria Eskevich, Franciska de Jong

ISBN: 978-0-306-40615-7
EAN: 4 003994 155486

7 May 2018

Proceedings of the LREC 2018 Workshop
"ParlaCLARIN: LREC2018 workshop on creating and using parliamentary corpora"

7 May 2018 – Miyazaki, Japan

Edited by Darja Fišer, Maria Eskevich, Franciska de Jong

https://www.clarin.eu/ParlaCLARIN

Acknowledgments: Organisation of the workshop is supported by CLARIN ERIC.

Organising Committee

- Maria Eskevich, CLARIN ERIC, The Netherlands **
- Darja Fišer, University of Ljubljana and Jožef Stefan Institute, Slovenia*
- Franciska de Jong, CLARIN ERIC, The Netherlands

*: Chair of the Programme Committee
**: Chair of the Organising Committee

Programme Committee

Chairs

- Darja Fišer, University of Ljubljana and Jožef Stefan Institute, Slovenia, Chair
- Franciska de Jong, CLARIN ERIC, The Netherlands, Co-chair

Members

- Darius Amilevičius, Vytautas Magnus University, Lithuania
- Ilze Auziņa, University of Latvia, Latvia
- Kaspar Beelen, University of Amsterdam, The Netherlands
- Andreas Blätte, University of Duisburg-Essen, Germany
- Anastasia Deligiaouri, Western Macedonia University of Applied Sciences, Greece
- Griet Depoorter, Dutch Language Institute, Belgium
- Francesca Frontini, Université Paul Valéry - Montpellier, France
- Katerina T. Frantzi, University of the Aegean, Greece
- Maria Gavriilidou, ILSP/Athena RC, Greece
- Goran Glavaš, University of Mannheim, Germany
- Barbora Hladka, Charles University, Czech Republic
- Laura Hollink, Centrum Wiskunde & Informatica, The Netherlands
- Caspar Jordan, Swedish National Data Service, Sweden
- Martijn Kleppe, National Library of the Netherlands, The Netherlands
- Krister Lindén, University of Helsinki, Finland
- Bente Maegaard, University of Copenhagen, Denmark
- Maarten Marx, University of Amsterdam, The Netherlands
- Karlheinz Moerth, Austrian Academy of Sciences, Austria
- Monica Monachini, National Research Council of Italy, Italy
- Federico Nanni, University of Mannheim, Germany
- Jan Odijk, Utrecht University, The Netherlands

- Petya Osenova, IICT-BAS and Sofia University “St. Kl. Ohridski", Bulgaria
- Simone Paolo Ponzetto, University of Mannheim, Germany
- Wim Peters, University of Strathclyde, UK
- Stelios Piperidis, Athena RC/ILSP, Greece
- Valeria Quochi, National Research Council of Italy, Italy
- Ineke Schuurman, KU Leuven, Belgium
- Inguna Skadiņa, University of Latvia, Latvia
- Sara Tonelli, Fondazione Bruno Kessler, Italy
- Jurgita Vaičenonienė, Vytautas Magnus University, Lithuania
- Tamás Váradi, Hungarian Academy of Sciences, Hungary
- Tanja Wissik, Austrian Academy of Sciences, Austria
- Martin Wynne, Bodleian Libraries, University of Oxford, UK

Preface

Parliamentary data is a major source of socially relevant content. It is available in ever larger quantities, is multilingual, accompanied by rich metadata, and has the distinguishing characteristic that it spoken language produced in controlled circumstances that has been traditionally transcribed but now increasingly released also in audio and video formats. All those factors in combination require solutions related to its archiving, structuring, synchronization, visualization, querying and analysis. Furthermore, adequate approaches to its exploitation also have to take into account the need of researchers from vastly different Humanities and Social Sciences fields, such as political sciences, sociology, history, and psychology.

Given the maturity, variety, and potential of this type of language data as well as the rich metadata it is complemented with, it is urgent to gather researchers both from the side of those producing parliamentary corpora and making them available, as well as those making use of them for linguistic, historical, political, sociological etc. research in order to share methods and approaches of compiling, annotating and exploring them in order to achieve harmonization of the compiled resources, and to ensure current and future comparability of research on national datasets as well as promote transnational analyses.

An inspiring CLARIN-PLUS cross-disciplinary workshop "Working with parliamentary records" (https://www.clarin.eu/event/2017/clarin-plus-workshop-working-parliamentary-records) that was held in Sofia, Bulgaria, in Spring 2017, and a comprehensive overview of a multitude of the existing parliamentary resources within the CLARIN infrastructure (https://office.clarin.eu/v/CE-2017-1019-Parliamentary-data-report.pdf) clearly indicated a need for better harmonization, interoperability and comparability of the resources and tools relevant for the study of parliamentary discussions and decisions, not only in Europe but worldwide.

As a follow-up, this ParlaCLARIN workshop brings together researchers who are interested in compiling, annotating, structuring, linking and visualising parliamentary records that are suitable for research in a wide range of disciplines in the Humanities and Social Sciences. The accepted papers address the following topics:

- Creation and annotation of parliamentary data in textual and/or spoken format
- Annotation standards and best practices for parliamentary corpora
- Accessibility, querying and visualisation of parliamentary data
- Text analytics, semantic processing and linking of parliamentary data
- Parliamentary corpora and multilinguality
- Studies based on parliamentary corpora

The workshop programme is composed of:

- Keynote talk by Cornelia Ilie, Malmö University, Sweden, and Hellenic American University, Athens, Greece, titled "Applying Multi-Perspective Approaches to the Analysis of Parliamentary Data";
- Panel discussion "Infrastructural Support for Research on Parliamentary Data" with panelists: Andreas Blätte, University of Duisburg-Essen, Germany; Cornelia Ilie, Malmö University, Sweden, and Hellenic American University, Athens, Greece; Federico Nanni, University of Mannheim, Germany; Jan Odijk, Utrecht University, The Netherlands; and the keynote speaker;
- 15 presentations of long (3) and short (12) papers by 64 authors from 13 countries.

We have received a lot of interesting submissions and would therefore like to thank the reviewers for their careful and constructive reviews which have contributed to the quality of the event.

D. Fišer, F. de Jong, M. Eskevich May 2018

Programme

Welcome and Introduction

09.00 – 09.15 Darja Fišer, Jakob Lenardič. CLARIN Corpora for Parliamentary Discourse Research

Session 1: Creating parliamentary corpora

09.15 – 09.45 Andrej Pančur, Mojca Šorn, Tomaž Erjavec
SlovParl 2.0: The Collection of Slovene Parliamentary Debates from the Period of Secession

09.45 – 10.00 Maciej Ogrodniczuk. Polish Parliamentary Corpus

10.00 – 10.15 Tanja Wissik, Hannes Pirker. ParlAT beta Corpus of Austrian Parliamentary Records

10.15 – 10.30 Onur Güngör, Mert Tiftikci, Çağil Sönmez
A Corpus of Grand National Assembly of Turkish Parliament's Transcripts

Invited talk

11.00 – 12.00 Cornelia Ilie. Applying Multi-Perspective Approaches to the Analysis of Parliamentary Data

Session 2: Enriching parliamentary corpora

12.00 – 12.15 Federico Nanni, Mahmoud Osman, Yi-Ru Cheng, Simone Paolo Ponzetto and Laura Dietz
UKParl: A Data Set for Topic Detection with Semantically Annotated Text

12.15 – 12.30 Alina Karakanta, Mihaela Vela, Elke Teich
EuroParl-UdS: Preserving and Extending Metadata in Parliamentary Debates

12.30 – 12.45 Roberts Darģis, Ilze Auziņa, Uldis Bojārs, Pēteris Paikens, Artūrs Znotiņš
Annotation of the Corpus of the Saeima with Multilingual Standards

12.45 – 13.00 Gavin Abercrombie, Riza Batista-Navarro
A Sentiment-labelled Corpus of Hansard Parliamentary Debate Speeches

Session 3: Parliamentary data in computational social sciences 1

14.00 – 14.30 Nona Naderi, Graeme Hirst
Automatically Labeled Data Generation for Classification of Reputation Defence Strategies

14.30 – 14.45 Dimitris Gkoumas, Maria Pontiki, Konstantina Papanikolaou, Haris Papageorgiou
Exploring the Political Agenda of the Greek Parliament Plenary Sessions

14.45 – 15.00 Federico Nanni, Goran Glavaš, Simone Paolo Ponzetto, Sara Tonelli, Nicolò Conti, Ahmet Aker, Alessio Palmero Aprosio, Arnim Bleier, Benedetta Carlotti, Theresa Gessler, Tim Henrichsen, Dirk Hovy, Christian Kahmann, Mladen Karan, Akitaka Matsuo, Stefano Menini, Dong Nguyen, Andreas Niekler, Lisa Posch, Federico Vegetti, Zeerak Waseem, Tanya Whyte, Nikoleta Yordanova
Findings from the Hackathon on Understanding Euroscepticism Through the Lens of Textual Data

15.00 – 16.00 **Panel: Infrastructural Support for Research on Parliamentary Data**
Andreas Blätte, Cornelia Ilie, Federico Nanni, Jan Odijk

Session 4: Parliamentary data in computational social sciences 2

16.30 – 17.00 Dorte Haltrup Hansen, Costanza Navarretta, Lene Offersgaard
A Pilot Gender Study of the Danish Parliament Corpus

17.00 – 17.15 Sascha Diwersy, Francesca Frontini, Giancarlo Luxardo
The Parliamentary Debates as a Resource for the Textometric Study of the French Political Discourse

17.15 – 17.30 Andreas Blätte. Using Data Packages to Ship Annotated Corpora of Parliamentary Protocols: The GermaParl R Package

Closing remarks

Table of Contents

Applying Multi-Perspective Approaches to the Analysis of Parliamentary Data

Cornelia Ilie
Malmö University, Sweden; Hellenic American University, Athens, Greece

The growing availability of digital corpora of parliamentary proceedings has provided invaluable resources for large-scale machine-based analysis. However, most corpus-driven methodologies have often been primarily concerned with statistical results and quantitative analyses, rather than with text-based and discourse-driven analyses of the content of parliamentary proceedings (Baker, 2006; Partington, 2012). In general, quantitative corpus linguistic analyses point to general patterns or trends of language usage, but these need to be interpreted - through qualitative analysis - in terms of (institutional) context, party affiliation, interpersonal power balance, and debate topic, to name but a few. A helpful solution would be to set up a mixed toolbox that integrates quantitative techniques with qualitative discourse-analytical approaches.

While transcripts of parliamentary proceedings have been made available in several countries, researchers are still confronted with the controversial question of (in)accuracy. For example, UK Hansard reports, which are theoretically supposed to be verbatim, are actually edited in order to remove the more serious shortcomings of MPs' oral delivery (Slembrouck, 1992). First, intrinsic elements of spontaneous speech, such as false starts, involuntary repetitions, or incomplete sentences, are left out. Second, the written version does not reflect certain features of spoken language, such as intonation, stress and regional accents. Moreover, certain reformulations are produced by Hansard editors to avoid clumsy or unclear messages. (Mollin, 2007) compared a sample of the official transcript to a transcript made from a recording of a House of Commons session and found that characteristics of spoken language, such as incomplete utterances, hesitations and contextual talk had been omitted. Since the transcripts are not entirely accurate, analysts of parliamentary discourse corpora need to watch the video recordings (Ilie, 2010; Ilie, 2013; Ilie, 2018), which can provide important clues about 'missing links', inconsistencies and the like.

Since they are based on a more sophisticated and fine-grained analysis, qualitative approaches to corpus data (Sealey and Bates, 2016) can provide deeper insights into the wide-ranging correlations between the purely linguistic, the contextual and the performative levels of the parliamentary proceedings under consideration. This presentation discusses and illustrates the meaning negotiation that emerges at the interface of the micro- and macro-levels of analysis regarding the parliamentary dialogic confrontations recorded in the Hansard reports. Three interrelated processes of meaning construction and contextualization can be identified in parliamentary debates: lexical selection (key words, labels, forms of address), collocational patterning (clichés, quotations, ritualistic formulas), and interpersonal co-performance (question-answer sequences, statements and counter-statements, follow-ups). These processes can be analysed in relation to metadiscourse and interdiscursivity (using transcripts of parliamentary corpora), whereas the behavioural and interpersonal dynamics need to be analysed in relation to visual prompts (which presupposes access to videorecordings).

Bibliographical References

Baker, P. (2006). *Using Corpora in Discourse Analysis (Continuum Discourse)*. Continuum International Publishing Group.

Ilie, C. (2010). Strategic uses of parliamentary forms of address: The case of the u.k. parliament and the swedish riksdag. *Journal of Pragmatics*, 42(4):885 – 911. Pragmatic Perspectives on Parliamentary Discourse.

Ilie, C. (2013). Gendering confrontational rhetoric: discursive disorder in the british and swedish parliaments. *Democratization*, 20(3):501–521.

Ilie, C. (2018). "Behave yourself, woman!" – Patterns of gender discrimination and sexist stereotyping in parliamentary interaction. *Journal of Language and Politics*, forthcoming.

Mollin, S. (2007). The hansard hazard: gauging the accuracy of british parliamentary transcripts. *Corpora*, 2(2):187–210.

Partington, A., (2012). *Corpus Analysis of Political Language*. *The Encyclopedia of Applied Linguistics*. Blackwell Publishing Ltd.

Sealey, A. and Bates, S. (2016). Prime ministerial self-reported actions in prime minister's questions 1979–2010: A corpus-assisted analysis. *Journal of Pragmatics*, 104(Complete):18–31.

Slembrouck, S. (1992). The parliamentary hansard 'verbatim' report: the written construction of spoken discourse. *Language and Literature*, 1(2):101–119.

CLARIN Corpora for Parliamentary Discourse Research

Darja Fišer[1,2], Jakob Lenardič[1]
[1]Faculty of Arts, University of Ljubljana, Aškerčeva 2, 1000 Ljubljana, Slovenia
[2]Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
{darja.fiser, jakob.lenardic}@ff.uni-lj.si

Parliamentary debates are an important resource for many disciplines in digital humanities and social sciences because they contain impactful information and special, formalized and often persuasive and emotional language. This paper presents the parliamentary corpora in the CLARIN infrastructure and suggests how they could be made more readily available to digital humanities and social sciences researchers in order to promote interdisciplinary, trans-national and cross-cultural studies.

1. Introduction

Parliamentary discourse displays specific institutional discursive features, complies with a set of rules and conventions, is motivated by a wide range of communicative goals such as persuasion, negotiation and agenda-setting along ideological or party lines, and is characterized by institutional role-based commitments, dialogically shaped institutional confrontation and the awareness of a multi-layered audience (Ilie, 2017). Due to their unique content, structure and language, records of parliamentary sessions have been a quintessential resource for a wide range of research questions from a number of disciplines in digital humanities and social sciences for the past 50 years (Chester and Browning, 1962; Franklin and Norton, 1993), such as political science (van Dijk, 2010), sociology (Cheng, 2015), history (Pančur and Šorn 2016), discourse analysis (Hirst et al., 2014), sociolinguistics (Rheault et al., 2015) and multilinguality (Bayley et al., 2004) but has only recently started to acquire a truly interdisciplinary scope (Bayley, 2004; Ihalainen et al., 2016). With an increasingly decisive role of parliaments and their rapidly changing relations with the public, media, government and international organizations, further empirical research and development of richly annotated and integrative analytical tools is necessary to achieve a better understanding of the specificities of parliamentary discourse and its wider societal impact, in particular with studies that take into account diverse parts of society (women, minorities, marginalized groups) and cross-cultural dimensions.

In most countries, access to parliamentary records is becoming increasingly simple due to Freedom of Information Acts, which has sparked a number of national and international initiatives that are compiling parliamentary data into valuable, often richly annotated parliamentary corpora. Several of the developed parliamentary corpora in the CLARIN infrastructure have already been successfully used in scientific research in various disciplines. In computational linguistics, the Lithuanian corpus was the basis for the development of machine learning approaches for classifying political text in accordance with its ideological position (Kapočiūtė-Dzikienė and Krupavičius, 2014), as well as for a stylometric analysis to distinguish the styles of left-wing, centre-wing and right-wing parties (Mandravickaitė and Krilavičius, 2015). Recently, Meurer (2017) has used, among other corpora, *Talk of Norway* to develop dependency relations from LFG structures. In corpus linguistics, Sverredal (2014) has used the *Korp* version of the *Riksdag's Open Data* to conduct a corpus-based analysis of the development of plural forms in Swedish finite verbs. Pančur and Šorn (2016) have argued for the necessity of using corpora in historical studies to aid with exploring large amounts of historical sources with a showcase on the Slovene parliamentary corpus *SlovParl*.

Unfortunately, corpus development efforts are seldom co-ordinated, and as a consequence the resources are not uniformly sampled, annotated, formatted or documented, and in many cases not even made easily accessible. In order to promote comparability and reproducibility of research results as well as foster interdisciplinary, trans-national and cross-cultural studies, this paper gives an overview of the parliamentary corpora available through CLARIN, the European research infrastructure for language resources and technology (Hinrichs and Krauwer, 2016). We also discuss how they could be made more readily available to the heterogeneous research community, especially colleagues without an engineering background.

2. Overview of CLARIN parliamentary corpora

Table 1: Overview of the parliamentary corpora in CLARIN, sorted by country code.

Country	Size (mil tokens)	Period	Linguistic annotation
cz	0.5	/	Speech-text alignment
de	0.4	1998-2015	/
dk	7.3	2008-2010	T, PoS, L
ee	13	1995-2001	/
el	28.7	2011-2015	/
fi	2.2	2008-2016	/
fr	0.17	2002-2012	/
lt	23.9	1990-2013	T, PoS, L
no_1	63.8	1998-2016	T, PoS, L
no_2	29	2008-2015	/
pt	1	1970-2008	T, PoS, L
se	1,250	1971-2016	T, PoS, L, Semantic
si	10.8	1990-1992	T, PoS, L
uk_1	1,600	1803-2005	T, PoS, L
uk_2	0.19	1998-2015	/
eu	588	1996-2011	Sentence alignment

In total, there are 16 parliamentary corpora accessible through the CLARIN infrastructure. Apart from the multilingual *Europarl* corpus (Koehn, 2005), which contains debates from the European parliament in 21 languages, there are 2 corpora of British parliamentary debates, 2 corpora of Norwegian debates and 1 corpus per country, for the following 11 countries: Czech Republic, Denmark, Estonia, Finland, France, Germany, Lithuania, Portugal, Slovenia, and Sweden. Table 1 gives an overview of the identified corpora in terms of size, period, and linguistic annotation.[1] The handles to the corpora are given in the Language resources section at the end of the paper.

2.1. Large monolingual corpora

Czech: ***Czech Parliament Meetings*** (Pražák and Šmidl, 2012) consists of audio recordings and related transcriptions that correspond to approximately 500,000 tokens. It is available for download on the website of the Czech repository *LINDAT* under the public CC-BY licence and for online querying through *KonText*.[2] The transcriptions of parliamentary discussions were semi-automatically aligned to the recordings and annotated with speaker-related information.

Danish: ***DK-CLARIN Almensprogligt korpus - offentlig del: tekster fra Folketinget*** (CLARIN-DK, 2011) includes Danish parliamentary proceedings from 2008-2010 and consists of 7.3 million tokens. The corpus is tokenised, PoS-tagged and lemmatised and is available for download from the Danish repository *CLARIN-DK* under a non-specific public licence.

Estonian: ***Transcripts of Riigikogu*** (University of Tartu, 2014) consists of approximately 13 million tokens and covers the time span 1995-2001. Aside from TEI-annotation, it is unclear how the corpus is annotated. The corpus can be downloaded under a non-specific academic licence on the corpus webpage or accessed online through the *Keeleveeb Query* concordancer provided by CLARIN-Estonia.

Finnish: The ***Eduskunta corpus*** **(**Bartis, 2017a; 2017b; Lennes, 2017) covers Finnish parliamentary data for the period between 2008 and 2016. The corpus consists of 2.2 million tokens. The corpus can be downloaded from the Finnish repository *Language Bank of Finland*, which also provides the associated videos of the sessions, as well as queried through the concordancer *Korp* (Finnish distribution).[3]

Greek: ***Hellenic Parliament Sittings*** (clarin:el, 2015) includes Greek parliamentary proceedings for 2011-2015 and consists of 28.7 million tokens. It is unclear how the corpus is annotated. This corpus is available for download under the academically-restricted CC BY-NC licence from the Greek repository *clarin:el*.

Lithuanian: ***Lithuanian Parliament Corpus for Authorship Attribution*** **(**Kapočiūtė-Dzikienė et al., 2017) includes Lithuanian parliamentary data for 1990-2013 and consists of 23.9 million tokens. The corpus is tokenised, PoS-tagged and lemmatised. This corpus can be downloaded from the CLARIN-LT repository under a CLARIN-LT public licence.

Norwegian: There are two Norwegian parliamentary corpora – ***Talk of Norway*** (Lapponi and Søyland 2016) and ***Proceedings of Norwegian Parliamentary debates*** (Common Language Resources and Technology Infrastructure Norway, 2015). *Talk of Norway* covers Norwegian parliamentary speech for 1998-2016, consists of 63.8 million tokens, and is available for download through the *CLARINO* repository, while *Proceedings of Norwegian Parliamentary Debates* covers a slightly shorter period, 2008-2015, consists of 29 million tokens and is only available for online querying through the concordancer *Corpuscle*.[4] Both corpora are available under the NLOD public licence.

Portuguese: ***PTPARL Corpus*** (ELRA, 2008) covers Portuguese parliamentary proceedings from 1970-2008 and consists of approximately 1 million tokens. The corpus is tokenised, PoS-tagged and lemmatised. It is listed for download in the ELRA catalogue[5] under the non-commercial ELRA END USER and commercial ELRA VAR licences.

Slovene: The ***SlovParl*** (Pančur et al., 2017) corpus covers Slovene parliamentary proceedings for 1990-1992 and in its latest version consists of 10.8 million tokens. The corpus is tokenised, PoS-tagged, and lemmatised. The corpus is available for download through the CLARIN.SI repository under CC BY and available for online querying through the CLARIN.SI concordancers.[6]

Swedish: ***Riksdag's Open Data*** consists of 1.25 billion tokens for 1971-2016 and is thus the second largest of the parliamentary corpora in the CLARIN infrastructure. The corpus is tokenised, PoS-tagged, lemmatised, and contains annotations of lemgrams, compounds and named entities. It is available through the *Språkbanken* repository and can either be downloaded through or queried online through *Korp* (Swedish distribution).[7] The corpus is available under CC BY.

UK: ***The Hansard Corpus*** **(**The SAMUELS Project, 2016) consists of 1.6 billion tokens from 1803-2005 and is the largest parliamentary corpus in the CLARIN infrastructure both in word size and temporal span. The corpus is tokenised, PoS-tagged, lemmatised and also displays seep semantic annotation. It is listed on the website of CLARIN-UK and is available for querying through the *BYU* concordancer.

[1] T = Tokenisation; PoS = Part-of-Speech tagging; L = lemmatisation

[2] http://lindat.mff.cuni.cz/services/kontext/first_form?corpname=czechparl_2012_03_28_cs_w.

[3] https://korp.csc.fi/.

[4] http://clarino.uib.no/korpuskel/page.

[5] http://catalog.elra.info/product_info.php?products_id=1179.

[6] http://www.clarin.si/info/concordances/.

[7] https://spraakbanken.gu.se/korp/.

2.2. Small monolingual corpora

In addition to the large corpora, the three smaller thematic corpora for English, French and German parliamentary speech are available for download under CC BY through the French ORTOLANG repository. These three corpora are the English ***Parliamentary Debates on Europe at the House of Commons*** (Truan, 2016a), the French ***Parliamentary Debates on Europe at the Assemblée nationale*** (Truan, 2016b), and the German ***Parliamentary Debates on Europe at the Bundestag*** (Truan, 2016c). Unlike the previously discussed large corpora, these contain only those parliamentary debates that correspond to the annual European Council meetings at the respective parliaments. The French corpus is for 2002-2012 while the English and German corpora cover a longer period, 1998-2015. In terms of token size, the English and French corpora are the smallest (approximately 190,000 and 173,000 tokens respectively), while the German corpus is slightly larger (approximately 417,000 tokens).

2.3. Multilingual corpus

The ***Europarl*** corpus (Koehn, 2005) is a multilingual parallel corpus of the sessions of the European Parliament. It covers the period 1996-2011, consists of 588 million tokens, is tokenised, sentence aligned and marked for speakers, and is freely available for download on a dedicated page[8] under no specified licence.

2.4. The state of the infrastructure

In general, the identified parliamentary corpora are well integrated within the CLARIN infrastructure. Almost all of the 16 corpora from Table 1 are listed in the Virtual Language Observatory (VLO),[9] which is the main metadata-based portal for language resources of the CLARIN infrastructure and provides the access point to finding resources across the national CLARIN centres (Uytvanck et al., 2012). The only exceptions are (i) *The Hansard Corpus*, (ii) *Hellenic Parliament Sittings* and (iii) the *Riksdag's Open Data* corpus, which are listed only in the respective national repositories (e.g. *CLARIN-UK* for *The Hansard Corpus*).

In terms of availability, 5 corpora can be both downloaded and accessed through an online concordancer (the Czech *Czech Parliament Meetings*, the Estonian *Transcripts of Riigikogu*, the Finnish *Eduskunta corpus*, the Swedish *Riksdag's Open Data*, and the Slovene *SlovParl* corpus), 3 can only be queried through an online environment (the British *Hansard Corpus*, the *Hungarian National Corpus* and *Proceedings of Norwegian Parliamentary Debates*), and the rest of the 9 corpora can only be downloaded.

Parliamentary corpora in the CLARIN infrastructure are described with high-quality metadata. Information on size and time period of the corpora is readily available (except for the temporal period included in the Czech corpus). Information on linguistic annotation is available for all the corpora except for the Finnish, Greek, Estonian and the *Proceedings of Norwegian Parliamentary debates* corpora. Although the documentation on the three thematic corpora described in section 2.2 refers to "Annotation of conversation"[10], the information on levels of linguistic annotation (e.g. PoS-tagging) is not given.

3. Findings from the CLARIN Focus Groups

In addition to evaluating the existence, findability, documentation and accessibility of parliamentary corpora in the CLARIN infrastructure presented in Section 2, we wanted to better understand how users experience the digital research infrastructure that CLARIN provides. To this aim, we conducted two half-day focus group interviews (Sanders, 2017) with 11 researchers from different disciplines from 10 European countries who are interested in CLARIN's parliamentary resources, asking them to share their experiences with the CLARIN infrastructure, obstacles they encountered, suggestions for improvement and the support and training they need.

Results indicate that both Social Sciences and Humanities researchers and speech and language technology/IT experts need more guidance about the CLARIN datasets, corpora and tools relevant for parliamentary data. First and foremost, they expressed a need for a more explicit metadata policy to ensure that high quality materials are easily available and accessible. In addition to easy access and navigation towards the relevant resources and tools, they also recommended that thorough documentation, training materials and best practice use cases for parliamentary data be provided in an enhanced online research environment. They also called for more systematic promotion campaigns, as CLARIN and its resources and tools are still unknown in many relevant research communities in their opinion. In the long run, it was recommended that CLARIN develops procedures to guarantee and monitor the quality of not only corpus metadata but also the quality of data and tools and to offer clearly visible information on recent updates of resources and tools.

4. Recommendations towards improved visibility of CLARIN parliamentary corpora

Based on the results of the resource survey and the focus group on parliamentary data we propose below recommendations to increase the visibility of these corpora to the heterogeneous and international research community, to showcase their potential for interdisciplinary, trans-national and cross-cultural studies, and to alleviate the technical obstacles that are preventing the use of the resources on a larger scale. The recommendations are comprehensive in the sense that they address all stages in the lifecycle of a resource and involve all the key players, such as resource developers, curators, infrastructure providers, knowledge sharing experts, and funders. While some of the recommendations require minimal to moderate post-production or curation efforts that can be handled centrally, others would require a substantial investment and direct involvement of the developers and curators. Despite the fact that this might not be a feasible short-term goal, the recommendations

[8] http://www.statmt.org/europarl/.
[9] https://vlo.clarin.eu.
[10] https://hdl.handle.net/11403/fr-parl/v1.

could be implemented in stages in future extensions or refinements of the existing resources, as well as by initiatives that are building new parliamentary corpora.

Intended use and users. Hughes et al. (2016) point out that "we can no longer take the impact and value of our expensive digital resources for granted, and it is not sufficient to make assumptions about use and users of digital collections". This is why we need to sample, annotate, format, document and release parliamentary corpora in such a way that they will be valuable to scholars with diverse backgrounds beyond corpus and computational linguistics which is still the prevalent situation in the CLARIN community. This issue is very important because in other disciplines different research data sampling methodologies are required (controlling for sociodemographic features, or topic-, event- or concept-based filtering etc.). An obvious development in this respect would be comprehensive data inclusion policies and regular updates of corpora with new material so that researchers could analyse the most recent but also chronologically the most diverse parliamentary activities. A more ambitious development would be semantic integration within and across parliamentary corpora. This would enable researchers to track and compare the same concepts and topics in different parliaments. A major boost would also be achieved by cross-referencing parliamentary corpora with external knowledge bases, such as place-name gazetteers and biographical lexica as well as with external documents, such as legislation and media coverage.

User interfaces and documentation. The results of the focus groups systematically show that the developers of CLARIN's tools and resources are generally overestimating users in terms of technological solutions they are offering to the researchers but underestimating them in terms of documentation about the tools and resources they believe will be relevant for the researchers. Overall, easy access to resources and straightforward user interfaces were emphasized the most and seem to carry the most impact. In addition, researchers attempting comparative studies reported interface fatigue (especially when offered in a language researchers are not proficient in, only partially localized into English or run on different platforms, resulting in different functionality as well as different results of seemingly identical functionalities). This is why researchers have expressed a need to be able to use a single tool for all parliamentary corpora that would require less time and effort to master but would also ensure that quantitative results are comparable across corpora. Good documentation was also pointed out as prerequisite for resource and tool criticism and interpreting research results (e.g. speech transcription and editing policy). On the other hand, the most frequent users expressed a desire for more complex functionalities of the interfaces and access to more advanced tools, such as distant reading, text mining and visualization applications which are currently not offered for a large majority of the available parliamentary corpora. This suggests that a balanced development of both simple and advanced solutions might be the most successful long-term solution.

Data structure and annotation. A prerequisite for a successful integration of multilingual and multinational parliamentary information into a single research environment is a systematic, incremental roadmap which requires all corpus developers to comply with a set of mutually agreed upon building blocks and text annotation, corpus encoding and metadata encoding standards. This will make the data at least formally uniform and will enable exploration and comparison across corpora.

Outreach activities and knowledge sharing. On-going promotion of parliamentary resources is of paramount importance, which was also confirmed in our focus groups. Namely, researchers will be most likely to use a resource or a tool if it is recommended to them by a colleague or in a training event they attend. While this is positive, it is not enough to result in a significant increase in users, and may be insufficient to maintain existing numbers. This is a common problem with most resources developed within projects which are funded for limited periods. However, a research infrastructure such as CLARIN has the instruments to ensure a recurrent budget for the promotion of its resources. According to the focus group results, researchers should be provided with use cases that demonstrate the importance and potential of parliamentary corpora to investigate research questions in their discipline. In addition to merely showcasing examples of research questions that can successfully be answered with parliamentary resources, the use cases should also demonstrate how advanced ICT approaches can be utilized in these kinds of studies. Apart from the use cases aimed at professional researchers, the need for educational use cases that can be integrated into university curricula have also been highlighted.

5. Conclusion

In this paper we have presented the parliamentary corpora available via the CLARIN infrastructure and analysed the level of their integration into the infrastructure, the quality of the associated metadata and ease of access. In general, the numerous parliamentary corpora are well integrated within the CLARIN infrastructure, their metadata is of high quality and most of the corpora can be downloaded. In terms of user on-line interfaces, parliamentary corpora are offered through many different concordancers which is an obstacle for users from different research backgrounds, international users and for users embarking on comparable research. In the framework of our efforts to make the corpora more visible and readily available to researchers from digital humanities and social sciences and to promote interdisciplinary, trans-national and cross-cultural studies, we have proposed some recommendations to make corpora more universally useful research datasets, to overcome technical and documentation barriers and to showcase the potential of parliamentary resources in research and education. They range from low-lying fruit to long-term policies and call for centralized interventions as well as for direct involvement of the resource developers and curators the actions of which need to be carefully motivated, planned, co-ordinated, monitored and evaluated by a designated task force.

6. Acknowledgements

The work reported in this paper has been supported by the member countries and observers in the CLARIN ERIC, and it has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 676529 for project CLARIN-PLUS. We would like to thank all the national User Involvement Coordinators and researchers who have provided invaluable feedback on our surveys. We would also like to thank the reviewers for their valuable comments.

7. Bibliographical References

Bayley, P. (Ed.). (2004). *Cross-cultural perspectives on parliamentary discourse (Vol. 10)*. John Benjamins Publishing: Amsterdam.

Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., and Schumacher, A. (2016). Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *Sixth Swedish Language Technology Conference 2016*. http://people.cs.umu.se/johanna/sltc2016/abstracts/SLTC_2016_paper_31.pdf.

Cheng, J. E. (2015). Islamophobia, Muslimophobia or racism? Parliamentary discourses on Islam and Muslims in debates on the minaret ban in Switzerland. *Discourse & Society*, 26(5): 562-586.

Chester, D. N. and Bowring, N. (1962). *Questions in parliament*. Clarendon Press: London.

Franklin, M. and Norton, P. (1993). *Parliamentary questions*. Oxford University Press: Oxford.

Hinrichs, E. and Krauwer, S. (2014). The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. In N. Calzolari et. al. (Eds.), *Proceedings of LREC 2014 : 9th International Conference on Language Resources and Evaluation*, 1525–31.

Hirst, G., Feng, V. W., Cochrane, C., and Naderi, N. (2014). Argumentation, Ideology, and Issue Framing in Parliamentary Discourse. In E. Cabrio, S. Villata, A.Z. Wyner (Eds.), *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*.

Hughes, L. M., Ell, P. S., Knight, G. A., and Dobreva, M. (2013). Assessing and measuring impact of a digital collection in the humanities: An analysis of the SPHERE (Stormont Parliamentary Hansards: Embedded in Research and Education) Project. *Digital Scholarship in the Humanities*, 30(2): 183-198.

Ihalainen, P., Ilie, C., and Palonen, K. (2016). *Parliament and Parliamentarism: A Comparative History of a European Concept*. Berghahn Books: Oxford, NY.

Ilie, C. (2017). Parliamentary Debates. In R. Wodak and B. Forchtner (Eds.), *The Routledge Handbook of Language and Politics*.

Kapočiūtė-Dzikienė, J. and Krupavičius, A. (2014). Predicting Party Group from the Lithuanian Parliamentary Speeches. *Information Technology and Control*, 43(3): 321-332.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf.

Mandravickaitė, J. and Krilavičius, T. (2015). Language usage of members of the Lithuanian Parliament considering their political orientation. *Deeds and Days*, 64: 133-151.

Meurer, P. (2017). From LFG structures to dependency relations. *Bergen Language and Linguistic Studies*, 8: 183-201.

Oravecz C., Váradi, T., and Sass, B. (2014). "The Hungarian Gigaword Corpus." In N. Calzolari et al. (Eds.), *Proceedings of LREC 2014 : 9th International Conference on Language Resources and Evaluation*, 1719-23.

Pančur, A. and Šorn, M. (2016). Smart Big Data : Use of Slovenian Parliamentary Papers in Digital History. *Contributions to Contemporary History*, 56(3): 130-146.

Rayson, P., Baron, A., Piao, S., and Wattam, S. (2015). Large-scale Time-sensitive Semantic Analysis of Historical Corpora. In *Proceedings of the 36th Meeting of ICAME*.

Sanders, W. (2017). Focus Group on User Involvement conducted during the CLARIN-PLUS Workshop "Working with Parliamentary Records", Sofia, Bulgaria, 27 March 2017.

Sverredal, K. (2014). Obehöriga verb äga ej tillträde En undersökning av verbets pluralkongruens i svenska. http://uu.diva-portal.org/smash/get/diva2:850961/FULLTEXT01.pdf.

Van Dijk, T. A. (2010). Political identities in parliamentary debates. European Parliaments under Scrutiny. In C. Ilie (Ed.), *European Parliaments under Scrutiny: Discourse strategies and interaction practices*, 29-56.

Van Uytvanck, D., Stehouwer, H., and Lampen, L. (2012). Semantic metadata mapping in practice: The Virtual Language Observatory. In Calzolari et al. (Eds.), *Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation*, 1029-1034.

8. Language Resource References

Bartis, I. (2017a). Plenary Sessions of the Parliament of Finland, Downloadable Version 1. https://vlo.clarin.eu/record?4&docId=http_58__47__47_urn.fi_47_urn_58_nbn_58_fi_58_lb-2017030901&q=parliament+of+finland&index=1&count=1081450.

Bartis, I. (2017b). Plenary Sessions of the Parliament of Finland, Kielipankki Korp Version 1. https://vlo.clarin.eu/record?5&docId=http_58__47__47_urn.fi_47_urn_58_nbn_58_fi_58_lb-2017020202&q=Plenary+Sessions+of+the+Parliament+of+Finland&index=2&count=1096117.

Common Language Resources and Technology Infrastructure Norway. (2005). Proceedings of Norwegian parliamentary debates (2008-2015). http://clarino.uib.no/korpuskel/landing-page?resource=stortinget&view=short.

clarin:el. (2015). Hellenic Parliament Sittings (2011-2015). http://hdl.grnet.gr/11500/AEGEAN-0000-0000-2545-9

DK-CLARIN. (2011). DK-CLARIN Almensprogligt korpus - offentlig del: tekster fra Folketinget, CLARIN-DK repository. https://clarin.dk/clarindk/item.jsp?id=dkclarin:986010#body_short-0.

ELRA. (2008). PTPARL Corpus. http://catalog.elra.info/product_info.php?products_id=1179.

Kapočiūtė-Dzikienė, J., Šarkutė, L. & Utka, A. (2017). Lithuanian Parliament Corpus for Authorship Attribution, CLARIN-LT digital library in the Republic of Lithuania, http://hdl.handle.net/20.500.11821/17.

Lapponi, E. and Søyland, M. G. (2016). Talk of Norway, Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository, http://hdl.handle.net/11509/123.

Lennes, M. (2017). Plenary Sessions of the Parliament of Finland, Kielipankki LAT Version 1. https://vlo.clarin.eu/record?6&docId=http_58__47__47_urn.fi_47_urn_58_nbn_58_fi_58_lb-2017122021&q=Plenary+Sessions+of+the+Parliament+of+Finland&index=3&count=1096117.

Pančur, A., Šorn, M., and Erjavec, T. (2016). Slovenian parliamentary corpus SlovParl 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1075.

Pančur, A., Šorn, M., and Erjavec, T. (2017). Slovenian parliamentary corpus SlovParl 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1167.

Pražák, A. and Šmídl, L. (2012). Czech Parliament Meetings, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11858/00-097C-0000-0005-CF9C-4.

The SAMUELS project. (2016). The Hansard Corpus. https://www.hansard-corpus.org/.

Truan, N. (2016a). Parliamentary Debates on Europe at the House of Commons (1998-2015) [Corpus]. ORTOLANG (Open Resources and Tools for LANGuage). https://hdl.handle.net/11403/uk-parl.

Truan, N. (2016b). Parliamentary Debates on Europe at the assemblée nationale [Corpus]. ORTOLANG (Open Resources and Tools for LANGuage). https://hdl.handle.net/11403/fr-parl/v1/.

Truan, N. (2016c). Parliamentary Debates on Europe at the Bundestag [Corpus]. ORTOLANG (Open Resources and Tools for LANGuage). https://hdl.handle.net/11403/de-parl/v1/.

University of Tartu. (2014). Transcripts of Riigikogu (Estonian Parliament). http://www.cl.ut.ee/korpused/segakorpus/riigikogu/.

Váradi, T. (2005). Hungarian National Corpus. http://hdl.handle.net/11372/LRT-345

SlovParl 2.0: The Collection of Slovene Parliamentary Debates from the Period of Secession

Andrej Pančur[1], Mojca Šorn[2], Tomaž Erjavec[3]
Institute of Contemporary History[1,2], Jožef Stefan Institute[3]
Ljubljana Slovenia
{andrej.pancur,mojca.sorn}@inz.si, tomaz.erjavec@ijs.si

Abstract

The paper describes the process of acquisition, up-translation, encoding, and annotation of the collection of the parliamentary debates from the Assembly of the Republic of Slovenia from 1990-1992, covering the period before, during, and after Slovenia became an independent country in 1991. The entire collection, comprising 232 sessions, 58,813 speeches and 10.8 million words was uniformly encoded in accordance with the Text Encoding Initiative (TEI) Guidelines, using the TEI module for drama texts. The corpus contains extensive meta-data about the speakers, a typology of sessions etc. and structural and editorial annotations. The corpus was also converted to use the spoken corpus module of the TEI, and from this encoding automatically part-of-speech tagged and lemmatised. The corpus is maintained on GitHub and its major versions archived in the CLARIN.SI repository and is available for analysis under the CLARIN.SI concordancers, offering an invaluable resource for historians studying this watershed period of Slovenian history.

Keywords: Slovenia's independence, Text Encoding Initiative, Open Science

1. Introduction

Parliamentary papers are a rich source of data used by different academic disciplines, among others, historiography, sociology, political science, linguistics, economic and economic history. Parliamentary papers include transcriptions of parliamentary debates, debate reports, session papers, petitions, legal documents, amendments, statements, written questions, committee reports, transcription of committee debates, etc. In some European countries, a large part of parliamentary papers is already accessible in digital form, but mostly in PDF only (Benardou, et al., 2015).

This mostly also applies to Slovenia as transcriptions of parliamentary debates in PDF or HTML are available for different historical regional parliaments[1] and various national parliaments of the countries to which Slovenia belonged.[2] Only for the National Assembly of the Republic of Slovenia, all transcription of parliamentary debates after 1990 are available in the HTML format.[3] Other parliamentary papers, especially session papers,[4] are also increasingly available in digital form. Under a conservative estimate, the session papers and the transcriptions of parliamentary debates since 1945 alone have more than 170 million words (Pančur & Šorn, 2016).

It is clear that no researcher is able to read that much text in its entirety. Most researchers in the humanities understand such digital materials only in terms of easier and quicker access to desired information (Spiro, 2014). They have no intention of reading the whole text, but only to find what they are looking for in the text and typically use search engines to identify sources and do a full text search on the results. This also applies to researchers of Slovenian parliamentary history. However, this research method has its limitations. Inasmuch as the researcher does not carefully examine every search result, the results are always lacking their proper context (Robertson, 2016).

For these reason, a small group of historians decided to build a corpus of parliamentary debates that will capture as much contextual information as possible. We chose a relatively short but historically very interesting period of the National Assembly of the Republic of Slovenia between 1990 and 1992, covering the period before, during, and after Slovenia became an independent country in 1991. From 1945 to 1991 Slovenia was part of Yugoslavia, and the parliament reflected its socialist system. The first multi-party elections took place in April 1990. In the next two years, the socialist assembly served the democratically elected members as a framework that enabled them to change and adapt the legislation for the Slovene Republic, which then led to Slovenia's independence in June 1991. In 1992, the members of the assembly passed the new constitution, which formally ended the era of the Socialist Assembly of Slovenia and established the new classical parliament.

The first, pilot version of this corpus spanning 1990 to 1992 and containing 2.7 million words was released in 2016 (Pančur et al., 2016), and on its basis we have made experiments on the possible use of such corpora in historical research (Pančur & Šorn, 2016).

Building of annotated corpora of (historical) parliamentary debates has already been undertaken for a number of countries, e.g., United Kingdom from 1803 on (Alexander, et al., 2016), Netherlands from 1814 on (Marx & Schuth, 2010) and Canada from 1901 on (Beelen et al., 2017). The Dutch corpus has already been successfully used in historical research (Piersma et al., 2014). From the Slavic-speaking countries, we are aware of only one other available corpus of Parliament

[1] Representative assemblies of the Austrian crown lands Carniola (1861-1874) http://hdl.handle.net/11686/menu719 and Styria (1848-1914) http://www.landesarchiv.steiermark.at/cms/ziel/111284715. Assembly of the Yugoslav federal republic of Slovenia (1947-1990) http://hdl.handle.net/11686/menu407, http://hdl.handle.net/11686/menu407.
[2] Habsburg Monarchy (1861-1918), http://alex.onb.ac.at/sachlichegliederung.htm. Yugoslavia (1919-1939, 1942-1953), http://hdl.handle.net/11686/menu233, http://hdl.handle.net/11686/menu825, http://hdl.handle.net/11686/menu822.
[3] https://www.dz-rs.si.
[4] http://hdl.handle.net/11686/menu828.

Meetings, from the Czech republic (Pražák & Šmidl, 2012).

This paper documents the making, annotation and availability of the second, comprehensive (10.8 million words and 58,813 speeches) version of the SlovParl corpus and is structured as follows: Section 2 explains the process of compilation, Section 3 details its annotation, Section 4 focuses on its availability, Section 5 gives some possibilities of a quantitative analysis of the corpus, and Section 6 gives the conclusions and directions for further research.

2. Building the Corpus

2.1 Basic Principles

In the design of SlovParl corpus, we followed these basic principles:

1. **Multidisciplinary**: The corpus must be useful not only for historians, but also for other disciplines. That is why SlovParl corpus (and also this paper) was created in close cooperation between the Slovenian DARIAH[5] and CLARIN[6] communities.
2. **All-inclusive**: In addition to parliamentary debates, other types of parliamentary papers will eventually be included.
3. **Long-term**: Since such large-scale plans can't be realized during the period of a short-term research project, these activities should be financed as part of long-term research infrastructures.
4. **Open science**: All previous principles can be optimally realized only in accordance with the principles of open science.

2.2 Document structure

Parliamentary debates are typically published in a uniform format, which fluctuates very little in time (Marx, 2009). This also applies to Slovenian parliamentary debates. By analysing representative samples, we found the following structure of parliamentary proceedings (with minimal and maximal occurrences of structural elements):

- Document (1, n)
  - Table of contents (0, 1)
  - List of speakers (0, 1)
  - Index (0, 1)
  - Annex (0, n)
  - Meeting (1, n)
    - Non-verbal content (0, n)
    - Topic (1, n)
      - Non-verbal content (0, n)
      - Speech (1, n)
        - Non-verbal content (0,n)
        - Paragraph (1, n)
          - Non-verbal content (0, n)

The structure of individual documents is very flexible. They might contain all meetings of all parliamentary chambers in one year, one meeting that lasts for several days, or only one day of the meeting. The document may contain the table of contents, the list of speakers, the topic index and annexes (session papers, legislation), or these might be present in separate documents. Non-verbal content of parliamentary debates (information about the meeting and the chairperson, description of the outcome of a vote, description of actions like applause, etc.) can be present anywhere in the structure of the meeting. Transition from one topic to another can occur during the chairman's speech.

2.3 Source Files

Transcriptions of parliamentary debates are available as PDF or HTML files on the web portal SIstory – History of Slovenia and on the Web pages of the Slovenian parliament. PDFs contain either images or OCR scanned text, while HTML files contain the digitized analogue of paper transcripts or born-digital text. Furthermore, OCR produced at times high-quality results but also quite low-quality transcriptions, due to the low print quality of the original. The following conversion, transcription and annotation procedures have been developed for these different source file formats: PDF → DOCX → XML, HTML → XML (Pančur, 2016).

To build the SlovParl corpus we only needed the HTML → XML conversion path, as the transcriptions of parliamentary debates of the National Assembly of the Republic of Slovenia are available on their web portal in HTML. We originally scraped the wanted data from their website, but after 2016 the links to the HTML files, together with metadata, are openly accessible as XML files.[7] The information (such as transcriptions of parliamentary debates) from this web portal is regarded as information of public character, with the disclamer that it can be always altered.[8]

2.4 Transcription

Transcriptions of Slovene parliamentary debates from the period of secession (1990-1992) were initially published as analogue publications and were digitized a few years ago by the National Assembly. OCR errors have been in most cases corrected.

The uniform structure of documents with parliamentary debates is very well suited for automatic annotation. But because HTML files for the period 1990-1992 do not contain born-digital text, the document structure is not clearly marked. The layout and other typographical aspects of source text (bold, italic, underline, indent, uppercase, punctuation, spacing) are not always consistently applied. Therefore, when converting from HTML to XML, semi-automatic annotation was performed in several steps. Each step contained:

1. using an XSL stylesheet for automatic annotation;
2. searching for annotation errors (XPath and regular expression search);
3. additional manual annotation.

[5] http://www.dariah.si/en/.
[6] http://www.clarin.si/info/about/.
[7] https://www.dz-rs.si/wps/portal/Home/OpenData.
[8] https://www.dz-rs.si/wps/portal/en/Home/pravnoObvestilo.

3. Annotation

The SlovParl 2.0 corpus is encoded as one XML document. Ten years ago there was no special XML schema for parliamentary proceedings (Marx, 2009). Today, the situation is completely different, and one can choose between Political Mashups (Gielissen & Marx, 2009),[9] Parliamentary Metadata Language (PML) (Gartner, 2014), and, last but not least, the Akomo Ntoso[10] schema.

Despite these options esp. developed for annotating parliamentary proceedings, we decided to use the Text Encoding Initiative Guidelines (TEI Consortium, 2016) for encoding SlovParl. This decision was based on our first two basic principles: multidisciplinary and all-inclusive corpus design. TEI is not only the *de facto* standard for annotating electronic text in the humanities, but is also widely used in the Slovenian CLARIN community (Erjavec et al., 2016). TEI has community-based maintenance, extensive documentation and a number of supporting tools. A central aspect of TEI usage is customization and the TEI Guidelines are designed with customization in mind. Unlike Political Mashups and PML, TEI can be used not only for the annotation of parliamentary proceedings, but also for all other types of parliamentary papers. In this respect, only Akoma Ntoso is comparable with TEI, as it is specially designed for parliamentary, legislative and judiciary documents. It also allows customization. However, the TEI ODD (One Document Does it all) specification language can also be used as a powerful technical platform for customization, as it offers project and data specific customisations and documentation, comparison of TEI-based project through their ODDs and even ODD chaining.[11]

3.1 TEI drama and TEI speech

Each TEI document is rooted in the <TEI> elements, which first contains a <teiHeader> element with metadata (title, date, time period, parliamentary organization, licence, source, automatic annotation and revision description). The TEI header is followed by the <text>, which in our case contains the document structure described in Sec. 2.2 above. The table of contents, the list of speakers, Index and Annexes can be found as elements in <front> or <back>, while meetings are located in the <body> element. Topics are encoded as elements inside <body>. They bear the @corresp attribute with references to the table of contents.

Scenes, acts and speeches are structural features of performance text (Marx, 2009). We used the TEI module for Performance texts for implementing the analysis of the materials. These include elements for encoding the list of speakers as a cast list (<castList>), a speech (<sp>), the name of the speaker (<speaker>) and the "stage directions" (<stage>). Each speech element bears a @who attribute with a local reference to the <actor> element in the cast list. Different types of non-verbal content (<stage>) are annotated with the @type attribute, which can have the following values: location, time, vote, quorum, debate, comment, gap, vocal, kinesic, and incident.[12] The <timeline> element provides a set of ordered points in time which are linked to the <stage> element with information about the time of the beginning and end of the debate.

In the next phase, these TEI documents are included in the <teiCorpus> element. We made a common list of speakers and the index of topics for the entire corpus. In both cases, we encoded this data in separate TEI documents. In this way, we created a list of all MPs and other speakers (<listPerson>) and a list of all organizations (<listOrg>) whose members were these speakers. We used the TEI module for encoding persons (<person>), places and organizations (<org>). We took into account any changes to the names and structure of the organization. Through the attributes @ref and @ana of the <affiliation> element, persons are associated with the organizations (parliamentary chamber, political party, government institution) to which they belonged over different time periods. In the <speech> element, the local reference to the element <actor> was moved from @who to @corresp. Attribute @who now contains relative URI reference to a local document with <listPerson>.

For the next phase, we intended to carry out the linguistic annotation of the corpus. But within paragraphs (<p>) the speeches were very often interrupted by non-verbal <stage> elements. Therefore, we decided to break the existing paragraphs into verbal (utterance <u>) and non-verbal elements (<note>, <vocal>, <kinesic>, <incident> and <writing>) and these elements are defined in the TEI module for spoken corpora. An XSLT stylesheet was used to convert the source TEI drama-encoded corpus to the target TEI speech-encoded corpus. Local documents for the list of persons and the topic index have been included in <teiHeader> of the speech <teiCorpus>.

3.2 Linguistic annotation

The TEI-speech encoded corpus was tokenized, sentence segmented, tagged with morphosyntactic descriptions (MSDs) and lemmatised with the ReLDI tagger (Ljubešić & Erjavec 2016). The resulting corpus is encoded identically to the source one, but, as illustrated in Figure 1, with the added sentence (<s>) word (<w>), punctuation (<pc>) and whitespace (<c>) elements. The word elements also bear the @lemma attribute, while both word and punctuation elements are annotated with @ana, which gives the MSD of the token.

```
<s>
  <w lemma="2." ana="msd:Mdo">2.</w><c> </c>
  <w lemma="verifikacija" ana="msd:Ncfsn">Verifikacija</w>
  <c> </c>
  <w lemma="mandat" ana="msd:Ncmsg">mandata</w>
  <c> </c>
  <w lemma="v" ana="msd:Sl">v</w><c> </c>
  <w lemma="zbor" ana="msd:Ncmsl">zboru</w>
  <pc ana="msd:Z">.</pc>
</s>
```

Figure 1. Linguistic annotation of the corpus.

[9] http://schema.politicalmashup.nl/schemas.html.
[10] http://www.akomantoso.org/.
[11] https://wiki.tei-c.org/index.php/ODD_chaining.

[12] Those familiar with TEI will notice that the last four value are in fact also names of TEI elements. We used them as the values of stage/@type in order to have a uniform encoding of the "stage directions" as present in the original transcripts.

It should be noted that the MSDs are given using the <prefixDef> element in the TEI header, which defines the prefixing scheme used, showing how abbreviated URIs using the scheme may be expanded into full URIs. In the case of the SlovParl 2.0 corpus the "msd:" prefix is simply expanded to local reference (i.e. "#") with the definitions of the MSDs included in the <back> element of linguistically annotated corpus – there, each MSD is defined as a feature-structure giving the decomposition of the MSD into its features. It is thus a simple matter, using just the TEI encoded corpus, to move from "msd:Mdo" to "Category = Numeral, Form = digit, Type = ordinal".

4. Availability and maintenance

4.1 GitHub

In accordance with our fourth basic principle (open science), all TEI annotated versions of the corpus are accessible and maintained in GitHub repositories:

- https://github.com/SIstory/Sejni_zapiski (DOCX → TEI drama – Phase 1)
- https://github.com/SIstory/Seje_DZ (HTML → TEI drama – Phase 1)
- https://github.com/SIstory/SlovParl (TEI drama – Phase 2)
- https://github.com/DARIAH-SI/CLARIN.SI (TEI speech)

4.2 CLARIN.SI repository

The corpus from the last GitHub repository is made available under the Creative Commons CC BY licence in the CLARIN.SI repository, comprising 231 sessions, 58,813 speeches and 10.8 million words (Pančur et al., 2017).

This repository item comprises four datasets:

- the corpus in TEI (module Transcription of speech);[13]
- the corpus in TEI with added automatic linguistic annotation;
- the corpus in CSV for statistical analysis software;
- the corpus in vertical format used by various concordancers.

4.3 Concordancers

The linguistically annotated version of the SlovParl 2.0 corpus has also been mounted under the two concordancers recently installed at CLARIN.SI, namely KonText[14] and noSketch Engine[15], enabling on-line exploration of this and other corpora.

The two concordancers are open source[16] and both use the same Manatee back-end (Rychlý, 2007) and set of indexed corpora, but provide different front-ends. Apart from visual differences, KonText supports log-in via the authentication and authorization infrastructure (AAI), and, in fact, allows only basic functionality without logging in. However, log-in enables the user to personalise the visual appearance of the concordancer, save sub-corpora and the query history. On the other hand, noSketch Engine, does not support log-in, so all its functionality is available to anonymous users, however, this also has the disadvantage of not allowing personalisation of the interface etc. As both concordancers use the same back-end, they also support querying via the powerful CQL query language, enabling searching via logical combinations of annotations, using regular expression, etc.

In order for a corpus to be indexed by the concordancers it needs to be fist converted to the so called vertical file format. We down-converted the linguistically annotated TEI encoded corpus to this format, also flattening the structure of the original, so that the vertical file is structured into texts (corresponding to one session) and paragraphs (corresponding to one speech) and with non-verbal parts omitted. Both structures carry metadata on e.g. the title of the session and its date, the speaker name and sex, and the type and topic of the speech. As mentioned above, this encoding of the corpus is also available for download from the CLARIN.SI repository.

5. Quantitative analysis

As mentioned above, the original reason for building a corpus was its use in historical research. In order to obtain the desired statistical information from TEI documents, we used the XML Query Language (XQuery) and XSLT. As a programming language for transforming XML documents, XSLT is not really intended for use in quantitative analysis. On the other hand, as a group of digital humanist, we have a good knowledge of XSLT, which enabled us to quickly find interesting information in the corpus. (Pančur & Šorn, 2016) For example, at the longest session the total duration of speeches was more than 56 hours and 256,692 words were spoken. From the beginning to the end of this session three months passed, it lasted 13 days and was interrupted 36 times. On the other hand, the total duration of speeches at the briefest session was only 10 minutes (643 words).

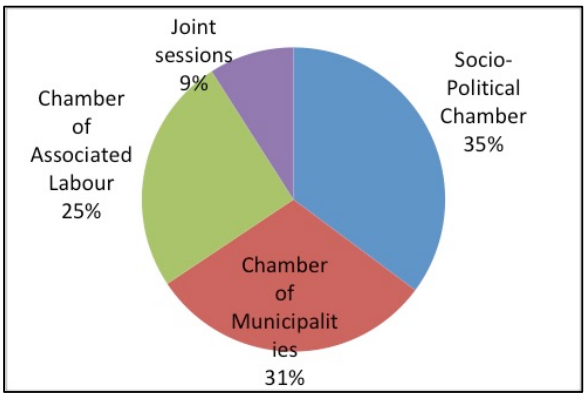


Figure 2: Number of words spoken in the chambers of the Assembly of the Republic of Slovenia (1990/92).

The Socialist Assembly of Slovenia comprised three chambers: the Socio-political chamber, the Chamber of Municipalities and the Chamber of Associated Labour.

[13] For researchers without XML knowledge this dataset is also available as a teiPublisher application. http://exist.sistory.si/exist/apps/parla/
[14] https://www.clarin.si/kontext/
[15] https://www.clarin.si/noske/
[16] The branch of KonText we use is available from https://github.com/ufal/lindat-kontext, while noSketch Engine can be downloaded via https://nlp.fi.muni.cz/trac/noske.

Joint sessions of all chambers represented less than a tenth of all parliamentary speeches (Figure 2). However, in previous research historians devoted almost exclusive attention only to some Joint sessions (Pesek, 2007). These researchers only read those small parts of the text that they considered relevant. Of course, such methods often yield useful results and a number of good studies have been created in such a manner using only pre-selected parts of parliamentary speeches. Why then would you need to build a corpus, if historians can still do well without it? Or the similar historian's question to the authors of an interdisciplinary book (corpus linguistics and historiography): "[…] what any quantification would actually show – it was clear that the corpus could quantify, but what was the purpose of that?" (McEnery an Baker, 2017, 200). We believe that it is the best to answer such a question with a concrete example:

After first multi-party elections (May 16, 1990 – May 14, 1992) the government consistent of newly established parties. Opposition parties stemmed from the former communist party and various socialist organizations. According to historians, because of its political inexperience, the coalition was relatively more silent compared to the opposition. (Pesek, 2007, p. 550) This finding was based on reading the speeches from some selected sessions (Gašparič, 2017). But corpus data show the opposite is true (Figure 3). The opposition numbered 36% of the MPs, who only had 20% of the speeches in which 32% of all words were spoken.

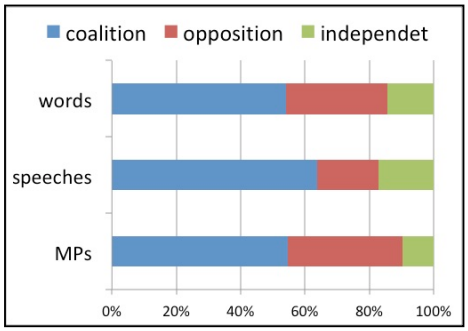


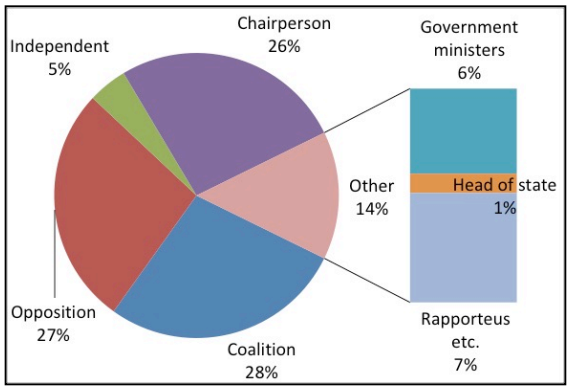
Figure 3: The percentage of members of parliament in coalition or opposition, the number of speeches and the number of spoken words; May 8, 1990 – May 14, 1992

But this simple analysis can also be misleading. All chairpersons were either coalition or independent MPs, and these chairpersons spoke as much as 26% of all words (Figure 4). If we exclude the speeches of chairmen, we find that the opposition and coalition MPs actually spoke about the same number of words.

On average, opposition MPs had more speeches than coalition MPs, which were also slightly longer. But this does not mean that the coalition as a whole was more silent than opposition. Both groups had outstanding speakers. Similarly, both groups had MPs who were almost completely silent (Figure 5).

The main question therefore is why the opposition on average had more MPs who were willing to speak more than coalition MPs? In addition to "political experience", an adequate answer to this question can only be given if other personal (gender, age etc.) and social factors (education, occupation, affiliation etc.) are taken into account.

Figure 4: Number of words spoken in the Assembly of the



Republic of Slovenia (May 8, 1990 – May 14, 1992) by organization membership.

We also made a set of CSV files containing various metadata from the corpus, appropriate for use with statistics-oriented software, such as R. This makes it easier for us to test new research hypotheses, as before, using only XSLT. At the same time, according to specific research needs, we can also easily add new metadata about persons and organizations.

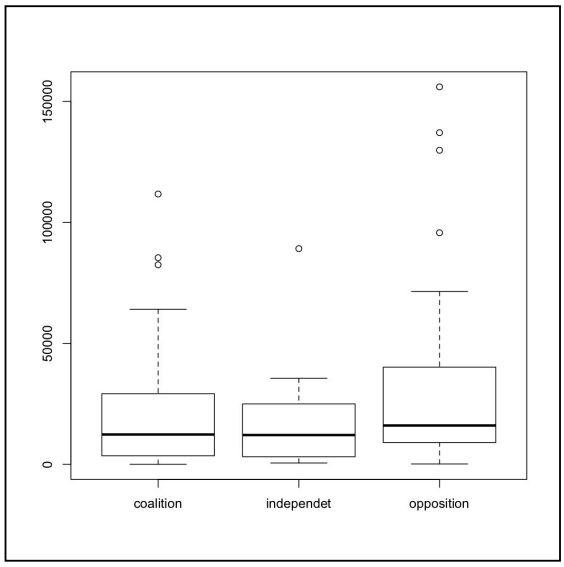


Figure 5: Number of words spoken in the Assembly of the Republic of Slovenia (May 8, 1990 – May 14, 1992) by organization membership.

6. Conclusions

The creation of SlovParl was in many aspects in accordance with good practices in the production of scholarly digital edition. This approach is particularly valuable in this regard:
"when digital editions are designed so that their textual data is captured using standards like TEI, this opens up

important opportunities for alternative deployments of the data." (MLA Commons, 2015)

In accordance with our basic principles, in the next years the corpus will not only be complemented with new parliamentary papers, but we will also pay special attention to research data reuse in different academic disciplines.

We hope that in this way we will be able to help other academic disciplines in tackling the shortage of not only these, but also related resources. At the moment, there are some larger projects aiming to collect and annotate similar political text resources. The Manifesto Project analyses parties' election manifestos[17] and the Comparative Agendas Project collects and organize data from archived sources to track policy outcomes across countries.[18] The results of these projects are of course also interesting for us. This is especially true for automatic topic classification of related language like Croatian (Karan et al., 2016). However, on the other hand, we believe that the topic classification from SlovParl can also provide a good basis for extension of contents analysis from only document titles to full text.

7. Acknowledgements

The work presented in this paper was supported by the Slovenian historiography research infrastructure (I0-0013), and the Slovenian ESFRI infrastructures DARIAH-SI and CLARIN.SI which are financially supported by the Slovenian Research Agency.

8. Bibliographical References

Beelen, K., Thijm, T.A., Cochrane, C., Halvemaan, K., Hirst, G., Kimmins, M., Lijbrink, S., Marx, M., Naderi, N., Rheault, L., Polyanovsky, R., and Whyte, T. (2017). Digitization of the Canadian Parliamentary Debates. *Canadian Journal of Political Science/Revue canadienne de science politique*, 50(3): 849-864.

Benardou, A., Dunning, A., Schaller, M., and Chatzidiakou, N. (2015). Research Themes for Aggregating Digital Content: Parliamentary Papers in Europa. Europeana Cloud – Work Package 1.

Erjavec, T., Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Fišer, D., Laskowski, C., and Zupan, K. (2016) Annotating CLARIN.SI TEI corpora with WebAnno. In *Proceedings*, pp. 1-5. https://www.clarin.eu/sites/default/files/erjavec-etal-CLARIN2016_paper_17.pdf.

Gartner, R. (2014). A metadata infrastructure for the analysis of parliamentary proceedings. In *Big Humanities Data, The Second IEEE Big Data 2014 Workshop*. Bethesda, Maryland, USA.

Gašparič, J. (2017). Parlament im Übergang: Versammlung der Republik Slowenien zum Zeitpunkt des Verfalls des Sozialismus und Jugoslawiens als Gegenstand einer historischen Analyse. Lecture, Kommission für Geschichte des Parlamentarismus und der politischen Parteien, Berlin, Germany, September 21.

Gielissen, T. and Marx, M. (2009). Exemplification of Parliamentary Debates. In *Proceedings of the 9th Dutch-Belgian Information Retrieval Workshop*, DIR 2009, pp. 19-25.

Karan, M., Šnajder, J., Širinić, D. and Glavaš, G. (2016) Analysis of Policy Agendas: Lessons Learned from Automatic Topic Classification of Croatian Political Text. *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Science, and Humanities (LaTeCH)*, pp. 12-21, Berlin, Germany, August 11.

Ljubešić, N. and Erjavec, T. (2016). Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Marx, M. (2009). Long, often quite boring, notes of meetings. In *ESAIR '09 Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pp. 46-53, Barcelona, Spain, February.

Marx, M. and Schuth, A. (2010). DutchParl: The Parliamentary Documents in Dutch. In *LREC 2010, Seventh International Conference on Language Resources and Evaluation*, pp. 3670–3677, Valleta, Malta, may. European Language Resource Association (ELRA).

McEnery, A. and Baker, H. (2017). *Corpus Linguistics and 17th-Century Prostitution: Computational Linguistics and History*. London and New York: Bloomsbury Academic.

MLA Commons (2015). Considering the Scholarly Edition in the Digital Age: A White Paper of the Modern Language Association's Committee on Scholarly Editions. https://scholarlyeditions.mla.hcommons.org/2015/09/02/cse-white-paper/.

Pančur, A. (2016) Označevanje zbirke zapisnikov sej slovenskega parlamenta s smernicami TEI = Encoding the Slovenian Parliament Session Minutes in Line with the TEI Guidelines. In *Proceedings of the Conference on Language Technologies & Digital Humanities*. Ljubljana: Ljubljana University Press, pp. 142-148.

Pančur, A. and Šorn, M. (2016) Smart Big Data: Use of Slovenian Parliamentary Papers in Digital History. *Prispevki za novejšo zgodovino*. 56 (3): 130-146. http://ojs.inz.si/pnz/article/view/193.

Pesek, R. (2007). *Osamosvojitev Slovenije: 'Ali naj Republika Slovenija postane samostojna in neodvisna država?'*. Ljubljana: Nova revija.

Pierska, H., Tames, I., Buitinck, L., Doornik, J., and Marx, M. (2014). War in Parliament: What a Digital Approach Can Add to the Study of Parliamentary History.Digital Humanities Quarterly, 8(1). .

Robertson, S. (2016). The Differences between Digital Humanities and Digital History. In *Debates in Digital Humanities 2016*. Minneapolis and London: University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/76.

Rychlý, P. (2007). Manatee/Bonito – A Modular Corpus Manager. Proceedings of the 1st Workshop on Recent Advances in Slavonic Natural Language Processing. 65–70. Brno, Masaryk University.

Spiro, L. (2014) Access, Explore, Converse: The Impact (and Potential Impact) of the Digital Humanities on

[17] https://manifesto-project.wzb.eu/.
[18] https://www.comparativeagendas.net/.

Scholarship. In *Keys for architectural history research in the digital era*. https://inha.revues.org/4925.

TEI Consortium. (2016). *Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/P5/.

9. Language Resource References

Alexander, M., Anderson, J., Archer, D., Baron, A., Davies, M., Hope, J., Jeffries, L., Kay, C., Rayson, P., and Walker, S. (2016). The Hansard Corpus: British Parliament. https://www.hansard-corpus.org.

Pančur, A., Šorn, M. and Erjavec, T. (2016). *Slovenian parliamentary corpus SlovParl 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1075.

Pančur, A., Šorn, M. and Erjavec, T. (2017). *Slovenian parliamentary corpus SlovParl 2.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1167.

Pražák, A. & Šmídl, L. (2012). Czech Parliament Meetings, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11858/00-097C-0000-0005-CF9C-4.

Polish Parliamentary Corpus

Maciej Ogrodniczuk
Institute of Computer Science, Polish Academy of Sciences
Warsaw, Poland
maciej.ogrodniczuk@ipipan.waw.pl

Abstract

This paper presents the Polish Parliamentary Corpus (PPC) — a new resource built upon the Polish Sejm Corpus and extended with current Senate proceedings and older (1918–1990) parliamentary transcripts. Corpus texts are automatically annotated with state-of-the-art language tools for Polish, resulting in a multi-layered stand-off sentence- and token-level segmentation, disambiguated morphosyntactic information, syntactic words and groups, named entities and coreference. The corpus is being constantly updated with new data from the current sittings. Currently the PPC is among the largest parliamentary corpora worldwide, amounting to approx. 300M words.

Keywords: written corpora, quasi-spoken data, parliament transcripts, Polish

1. Introduction

The idea of creating a separate text corpus based on Polish parliamentary data (Sejm and Senate, the lower and upper houses of the Polish parliament) appeared as early as 2010 when a paper outlining the resource was registered to SinFonIJA 3 conference[1]. The text build on the concepts introduced in the National Corpus of Polish (Przepiórkowski et al., 2010, NKJP)[2], pointing out availability of Sejm data in PDF format and suggesting further steps in the process: inclusion of Senate data and audio recordings. The first phase of the work could be completed only with an European CESAR project[3] in 2011, when all then available data was gathered[4] The data was retrieved from internal Sejm databases and compared to previously available NKJP data.

The resource has been made available as The Polish Sejm Corpus[5] (Ogrodniczuk, 2012). The texts have been encoded in NKJP-based TEI P5 format and following layers of linguistic annotation available in NKJP: paragraph-, sentence- and token-level segmentation, lemmatization, disambiguated morphosyntactic information, named entities, syntax words and groups. A searchable corpus version has also been made available as Poliqarp (Janus and Przepiórkowski, 2006) search engine binary (to be run on user's computer) and a Poliqarp-powered simple online search engine (`http://sejm.nlp.ipipan.waw.pl/`).

At the same time the texts were processed, although much more fragmentarily, by many other researchers in Poland. Parliamentary proceedings were included in the major written corpora such as the IPI PAN Corpus (Przepiórkowski, 2004), National Corpus of Polish (Przepiórkowski et al., 2010) with its distributable subcorpus[6], KPWr corpus (Broda et al., 2012) or the internal corpora of the Polish-Japanese Academy of Information Technology.

Over the years many new ideas were put forward calling for the update of the Sejm corpus. Apart from the constant flow of new data, language processing tools of much higher quality have been made available and new parliamentary resources have been produced by the Parliament itself, the most important of which were all transcripts of parliamentary proceedings from 1918–1990[7] digitized by the Sejm Library. Even though they were made available only in the form of images, this was a very important step towards the completion of the resource, now ready to include all 100 years of newest parliamentary history of Poland. Last but not least, the data from Polish Senate, similar in character but originally omitted from the corpus, were ready to be added.

The usage of the current corpus brought another motivation for maintenance of a separate parliamentary resource: the data has been widely popular among representatives of both the humanities and computational linguists[8]. As compared to other domains, the data features a broad spectrum of topics despite its controlled flavour. The usefulness of such setting also seems to be confirmed by several international initiatives related to parliamentary data such as the recent CLARIN-PLUS workshop "Working with Parliamentary Records" in Sofia[9] or user involvement queries summarized at the CLARIN conference in Budapest.

All these intermediary steps paved the way for the current version of the corpus which we call the Polish Parliamentary Corpus (PPC). The next sections of the paper describe its contents and construction principles.

[1] See `http://www.ung.si/~jezik/sinfon_3/program.html`.

[2] Pol. Narodowy Korpus Języka Polskiego, see `http://nkjp.pl`.

[3] Central and South-East European Resources, a CIP – ICT PSP grant 271022, February 2011 – January 2013.

[4] Sittings from terms of office 1–6 (1991–2011) and questions from terms of office 3–6 (1997–2011).

[5] See `http://clip.ipipan.waw.pl/PSC`

[6] See `http://zil.ipipan.waw.pl/DistrNKJP`

[7] See Parlamentaria website: `https://bs.sejm.gov.pl/F/?func=file&file_name=find-nowe&local_base=ars01`)

[8] See e.g. (Przybyła and Teisseyre, 2014; Marasek et al., 2015; Pęzik, 2015; Szela, 2016).

[9] See `http://www.clarin.eu/event/2017/clarin-plus-workshop-working-parliamentary-records`.

2. Corpus Data and Format

All data available in the Polish Sejm Corpus (stenographic records from 1991–2011, interpellations and questions from 1997–2011) has been included in the resulting resource. The Senate data from 1991–2011 has been retrieved from the distributable subcorpus of the National Corpus of Polish; later data has been newly harvested from the Senate website. The texts of interpellations and questions not available in the Polish Sejm Corpus have been acquired from Sejm website[10] using the P4 toolset by Daniel Janus.

However, the most interesting portion of Polish parliamentary transcripts dating to 1918–1990 which has never been made available before has been added to the resource only recently. The digitized data in the form of image-based PDF files has been retrieved from Parlamentaria website, passed through FineReader OCR tool and manually verified by human proof-readers.

The corpus follows the XML TEI P5-based annotation model put forward by the National Corpus of Polish[11] — a *de facto* standard for encoding and documenting Polish linguistic data. The format assumes stand-off linguistic annotation distributed over various layers represented in separate files:

- `header.xml`, covering detailed metadata of the sitting (sitting number/day, list of speakers etc.)
- `text_structure.xml`, the structure of the sitting split into utterances of the MPs grouped into continuous statements, created with dedicated scripts
- `ann_segmentation.xml.gz`, sentence- and token-level segmentation, created with Morfeusz SGJP (Woliński, 2006)
- `ann_morphosyntax.xml.gz`, disambiguated morphosyntactic annotation and lemma information, created with Morfeusz SGJP and Toygger tagger (Krasnowska-Kieraś, 2017)
- `ann_words.xml.gz`, syntactic words, created with Spejd (Buczyński and Przepiórkowski, 2009) shallow parser
- `ann_groups.xml.gz`, syntactic groups, created with Spejd
- `ann_named.xml.gz`, named entities, created with NERF (Savary et al., 2010)
- `ann_coreference.xml.gz`, mentions and coreference annotation, created with the newest neural system (Nitoń et al., 2018).

For detailed description of the format structure see (Ogrodniczuk, 2012).

[10] See `http://www.sejm.gov.pl`.

[11] See `http://nlp.ipipan.waw.pl/TEI4NKJP/` for samples of NKJP files.

3. Corpus Statistics and Availability

The current size of the corpus amounts to 194M segments with detailed distribution over houses and periods presented in Table 1. Apart from the stenographic records the corpus contains 104M segments of interpellations and questions.

Several interfaces have been made available to access corpus data such as a familiar Poliqarp-based (Janus and Przepiórkowski, 2006) interface at `http://sejm.nlp.ipipan.waw.pl/` (see Figure 1) or a more utterance-centric Smyrna-based (Janus, 2015) interface at `http://smyrna.sejm.nlp.ipipan.waw.pl/` (see Figure 2). Poliqarp binary package has also been made available to facilitate offline statistical queries, currently available only in the desktop version of the search engine.

4. Current and Future Work

Providing the corpus data together with some basic search capabilities is just the first step in the long process of making the data usable by representatives of the digital humanities, the most interested in the parliamentary resources. Apart from natural directions of development of the corpus (improving search and presentation, adding more annotation layers, richer metadata, audio/video linking) three of them are particularly worthy of note.

The first of them is inclusion of the remaining parliamentary data, now available only in the paper form: the first of them being the proceedings of committees, both standing and select. Their processing has already started and will be continued until the end of 2018.

The linguistic engineering tasks seem equally important. First type of them relates to improved processing of data with newest tools. Due to the deep neural network revolution new processing applications are being made available every year with improved accuracy reached on many levels of linguistic annotation[12] In case of large automatically tagged corpora even small progress results in much better quality of data. Another group of NLP tasks concerns the fact that the data span a long period of time with at least two important changes in the Polish orthography (a major reform in 1936 and some minor adaptations after 1956). Obviously, the accuracy of processing of such data with modern tools drops significantly with orthographic alterations which calls for development of separate tools for different historical periods. Such experiments for morphological analysis have already been carried out, cf. e.g. (Kieraś et al., 2017).

Last but not least, such a diverse dataset (as far as the topic and date span is concerned) would benefit from presentation interface resembling Google Books Ngram Viewer, capable of visualising term frequency differences. A similar solution has recently been adopted for Chronopress (Pawłowski, 2016), a corpus of post-war Polish press (until 1962).

Acknowledgements

The work reported here was financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

[12] See e.g. Toygger, a new tagger of Polish (Krasnowska-Kieraś, 2017).

Period	Years	Sejm	Sittings	Segments	Senate	Sittings	Segments
Second	1919–1922	Legislative Sejm	342	208 524	–	–	–
Polish	1922–1927	1st term of office	340	533 109	1st term	157	153 980
Republic	1928–1930	2nd	86	111 894	2nd	31	5 790
	1930–1935	3rd	148	139 166	3rd	80	123 651
	1935–1938	4th	90	141 387	4th	54	56 191
	1938–1939	5th	31	44 592	5th	20	38 776
	1943–1947	State National Council	11	11 671	–	–	–
People's	1947–1952	Legislative Sejm	108	514 572	–	–	–
Poland	1952–1956	1st term of office	39	76 244	–	–	–
	1957–1961	2nd	59	115 563	–	–	–
	1961–1965	3rd	32	62 799	–	–	–
	1965–1969	4th	23	45 233	–	–	–
	1969–1972	5th	19	33 492	–	–	–
	1972–1976	6th	32	63 155	–	–	–
	1976–1980	7th	29	57 352	–	–	–
	1980–1985	8th	79	132 845	–	–	–
	1985–1989	9th	50	89 436	–	–	–
Third	1989–1991	10th	79	157 225	1st term	61	111 450
Polish	1991–1993	1st term of office	45	7 803 935	2nd	38	1 461 165
Republic	1993–1997	2nd	115	22 299 861	3rd	102	5 057 468
	1997–2001	3rd	119	24 313 939	4th	90	8 261 548
	2001–2005	4th	109	28 986 555	5th	88	6 489 812
	2005–2007	5th	48	11 833 471	6th	39	3 573 955
	2007–2011	6th	100	22 682 341	7th	83	8 827 024
	2011–2015	7th	102	22 587 764	8th	82	7 110 114
	2015–	8th	54	5 905 461	9th	53	3 504 637

Table 1: Statistics of the Polish Parliamentary Corpus

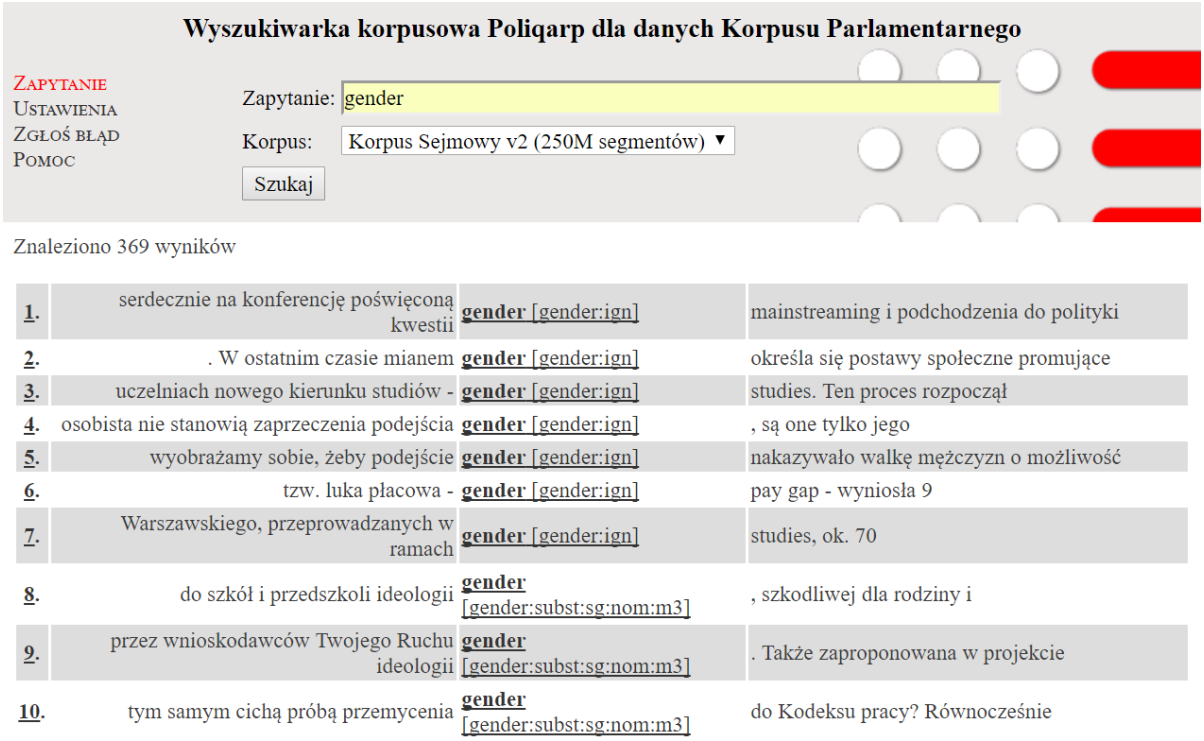
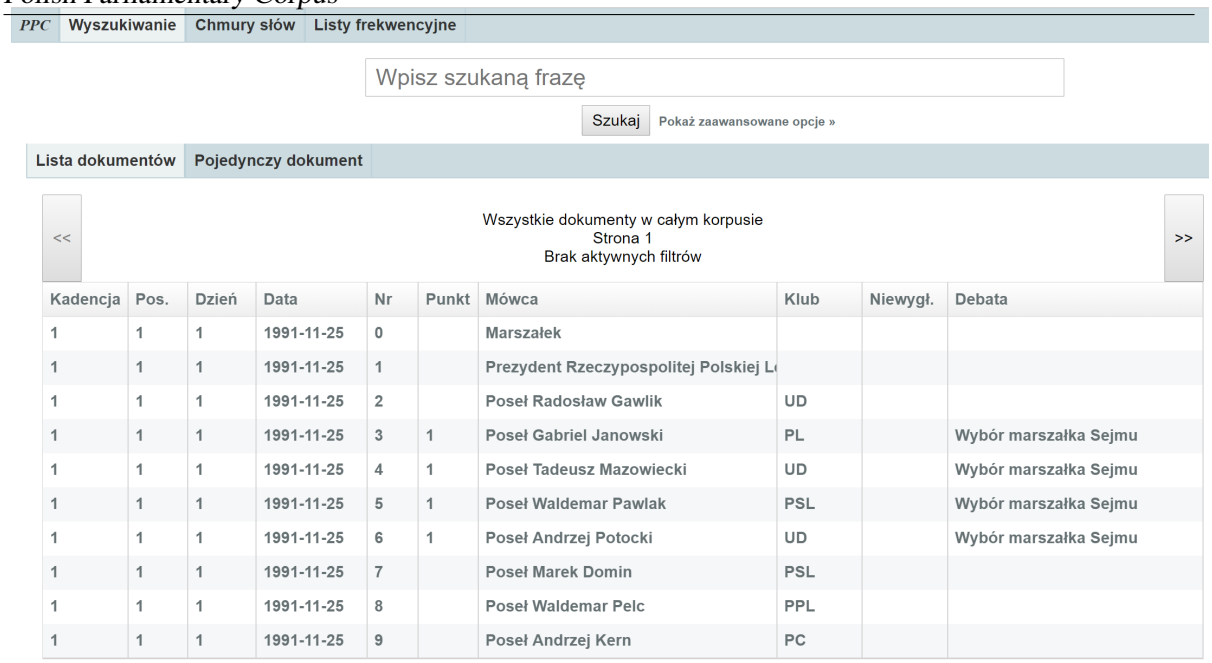


Figure 1: NKJP-based Poliqarp search in the corpus



Kadencja	Pos.	Dzień	Data	Nr	Punkt	Mówca	Klub	Niewygł.	Debata
1	1	1	1991-11-25	0		Marszałek			
1	1	1	1991-11-25	1		Prezydent Rzeczypospolitej Polskiej L			
1	1	1	1991-11-25	2		Poseł Radosław Gawlik	UD		
1	1	1	1991-11-25	3	1	Poseł Gabriel Janowski	PL		Wybór marszałka Sejmu
1	1	1	1991-11-25	4	1	Poseł Tadeusz Mazowiecki	UD		Wybór marszałka Sejmu
1	1	1	1991-11-25	5	1	Poseł Waldemar Pawlak	PSL		Wybór marszałka Sejmu
1	1	1	1991-11-25	6	1	Poseł Andrzej Potocki	UD		Wybór marszałka Sejmu
1	1	1	1991-11-25	7		Poseł Marek Domin	PSL		
1	1	1	1991-11-25	8		Poseł Waldemar Pelc	PPL		
1	1	1	1991-11-25	9		Poseł Andrzej Kern	PC		

Figure 2: Smyrna-based search in the corpus

Bibliographical References

Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski, A., and Wardyński, A. (2012). KPWr: Towards a Free Corpus of Polish. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 3218–3222, Istanbul, Turkey. European Language Resources Association (ELRA).

Buczyński, A. and Przepiórkowski, A. (2009). Spejd: A Shallow Processing and Morphological Disambiguation Tool. *Human Language Technology: Challenges of the Information Society. Vol. 5603*, pages 131–141.

Janus, D. and Przepiórkowski, A. (2006). Poliqarp 1.0: Some technical aspects of a linguistic search engine for large corpora. In Jacek Waliński, et al., editors, *The proceedings of Practical Applications of Linguistic Corpora 2005*, Frankfurt am Main. Peter Lang.

Kieraś, W., Komosińska, D., Modrzejewski, E., and Woliński, M. (2017). Morphosyntactic annotation of historical texts. The making of the baroque corpus of Polish. In Kamil Ekštein et al., editors, *Proceedings of the Twentieth International Conference Text, Speech, and Dialogue (TSD 2017)*, volume 10415 of *Lecture Notes in Computer Science*, pages 308–316. Springer International Publishing.

Krasnowska-Kieraś, K. (2017). Morphosyntactic disambiguation for Polish with bi-LSTM neural networks. In Zygmunt Vetulani et al., editors, *Proceedings of the Eighth Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 367–371, Poznań, Poland. Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Marasek, K., Koržinek, D., and Brocki, Ł. (2015). System for Automatic Transcription of Sessions of the Polish Senate. *Archives of Acoustics*, 39(4).

Nitoń, B., Morawiecki, P., and Ogrodniczuk, M. (2018). Deep Neural Networks for Coreference Resolution for Polish. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ogrodniczuk, M. (2012). The Polish Sejm Corpus. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 2219–2223, Istanbul, Turkey. European Language Resources Association (ELRA).

Przepiórkowski, A., Górski, R. L., Łaziński, M., and Pęzik, P. (2010). Recent developments in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta. ELRA.

Przepiórkowski, A. (2004). *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Przybyła, P. and Teisseyre, P. (2014). Analysing utterances in Polish parliament to predict speaker's background. *Journal of Quantitative Linguistics*, 21(4):350–376.

Pęzik, P. (2015). Spokes – a search and exploration service for conversational corpus data. In *Selected Papers from CLARIN 2014*, Linköping Electronic Conference Proceedings, pages 99–109. Linköping University Electronic Press, Linköpings universitet.

Savary, A., Waszczuk, J., and Przepiórkowski, A. (2010). Towards the Annotation of Named Entities in the National Corpus of Polish. In Nicoletta Calzolari, et al.,

editors, *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 3622–3629, Valletta. European Language Resources Association.

Szela, M. (2016). O wykorzystaniu angielsko-polskiego korpusu równoległego tekstów prawnych w badaniu cech języka tekstów tłumaczonych. In Agnieszka Gruszczyńska, Ewa; Leńko-Szymańska, editor, *Polsko-językowe korpusy równoległe*, pages 210–226. Instytut Lingwistyki Stosowanej, Warszawa.

Woliński, M. (2006). Morfeusz — a practical tool for the morphological analysis of Polish. In Mieczysław A. Kłopotek, et al., editors, *Proceedings of the International Intelligent Information Systems: Intelligent Information Processing and Web Mining'06 Conference*, pages 511–520, Wisła, Poland, June.

Language Resource References

Janus, D. (2015). *Smyrna: prosty konkordancer obsługujący język polski*. Available: `http://smyrna.danieljanus.pl`.

Pawłowski, A. (2016). *ChronoPress – Chronologica Corpus*. Copyright CC BY-NC-SA 3.0 (Attribution-NonCommercial-ShareAlike 3.0 Unported). Available in CLARIN-PL digital repository: `http://hdl.handle.net/11321/260`.

ParlAT beta
Corpus of Austrian Parliamentary Records

Tanja Wissik, Hannes Pirker
Austrian Centre for Digital Humanities, Austrian Academy of Sciences
Sonnenfelsgasse 19, 1010 Vienna, Austria
{tanja.wissik, hannes.pirker}@oeaw.ac.at

Abstract

The paper presents the beta Version of the ParlAT Corpus, a corpus of Austrian parliamentary records and its current state. The ParlAT project aims to create a corpus of all digitally available parliamentary records from the National council – one of the two chambers of the Austrian parliament – starting with 1945, i.e. for the period of the so called "Second Republic". The ParlAT beta contains parliamentary records for the last 21 years (i.e. between 1996 - 2017), that is 36% of the relevant digitally available parliamentary records. This paper will describe the data collection and data processing and give an outlook on future work.

Keywords: parliamentary records, corpus building, annotation, linking to external sources

1 Introduction

Parliamentary records are an interesting resource for various fields in the Humanities and Social Sciences, such as linguistics, political science, history, as well as for fields in the Information Sciences such as NLP or information retrieval. As a consequence, there are many initiatives, on the national and international level, that aim at compiling and analysing parliamentary records. Examples for monolingual corpora are the Hansard Corpus, the collection of the parliamentary records of the British Parliament between 1803 and 2005 (Alexander and Davies, 2015) or the Talk of Norway, a collection of the Norwegian parliamentary data (Lapponi and Søyland, 2016), examples for multilingual corpora are the ECPC Corpus, the European Parliamentary Comparable and Parallel Corpora for Spanish and English (Calzada Pérez, Marín Cucala and Martínez Martínez, 2006). A recent survey on parliamentary data in the context of the research infrastructure CLARIN (Fišer and Lenardič, 2017) has identified over 20 corpora of parliamentary data. However, the available corpus for Austrian parliamentary records, listed in the survey only covered a short time period, from 2013 to 2015 (Sippl et al. 2016), and is therefore the corpus not suitable for all research questions, especially not for diachronic analysis.F

The aim of the ParlAT project is to fill this gap and create a corpus of all digitally available parliamentary records from the National Council ("Nationalrat"), one of the two chambers of the Austrian Parliament for the "Second Republic." i.e. for the historic period starting in 1945 until today.

The verbatim records, in German called "Stenographische Protokolle" ("shorthand record") are available from the website of the Austrian Parliament[1] starting from the V legislation period. Prior to this, for the legislation period I-IV ("First Republic"), only scans are available via the platform ALEX at the Austrian National Library[2]. For the legislation period V to XIX (from 1945 to 1995) the documents are only available in pdf format, for the legislation period XX – XXV (starting from 1996) the documents are also available in html format. While parliamentary records are highly standardised and structured texts, changes to the structure over time can be observed (see also Figure 1).

The so called "Geschäftsordnungsgesetz 1975" law dictates that all public sessions of the National Council are recorded verbatim and in their entirety. However, it is also true that these records are not live recordings, and that the speakers get the verbatim records prior to publication and can make, for example, stylistic changes. In case of doubt, the President of the National Council decides if a change is admissible or not.

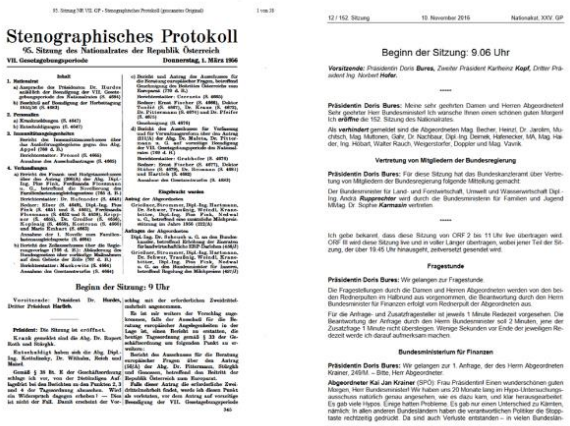


Figure 1: Comparison of verbatim records over time: left from 1950 in column style and right from 2016

All the texts are downloadable from the website of the Austrian Parliament and parliamentary documents can be searched by keyword or by identification number in the search interface. Besides the verbatim records, other documents, such as requests and inquiries, are available. However, the search interface does not support a full-text search or linguistically motivated queries nor is it possible to query texts across different legislative periods.

[1] https://www.parlament.gv.at/PAKT/STPROT/
[2] http://alex.onb.ac.at/spe.htm

Legislation period	Beginning	End	Nr. of documents	Format
V	19.12.1945	08.11.1949	117	pdf
VI	08.11.1949	18.03.1953	103	pdf
VII	18.03.1953	08.06.1956	95	pdf
VIII	08.06.1956	09.06.1959	84	pdf
IX	09.06.1959	14.12.1962	109	pdf
X	14.12.1962	30.03.1966	95	pdf
XI	30.03.1966	31.03.1970	175	pdf
XII	31.03.1970	04.11.1971	52	pdf
XIII	04.11.1971	04.11.1975	151	pdf
XIV	04.11.1975	04.06.1979	123	pdf
XV	05.06.1979	18.05.1983	149	pdf
XVI	19.05.1983	16.12.1968	175	pdf
XVII	17.12.1968	04.11.1990	52	pdf
XVIII	05.11.1990	06.11.1994	151	pdf
XIX	07.11.1994	14.01.1996	57	pdf
XX	15.01.1996	28.10.1999	183	pdf, html
XXI	29.10.1999	19.12.2002	118	pdf, html
XXII	20.12.2002	29.10.2006	164	pdf, html
XXIII	30.10.2006	27.10.2008	76	pdf, html
XXIV	28.10.2008	28.10.2013	220	pdf, html
XXV	29.10.2013	08.11.2017	199	pdf, html
Total number of documents			2648	

Table 1: Availability of Austrian parliamentary records for the time period 1945 – 2017.

2 ParlAT – Corpus of Austrian Parliamentary Records

The present version of the ParlAT corpus covers the parliamentary records from the XX to the XXV legislative period (1996 – 2017). The legislative period that has started in November 2017 is not yet included.

However, the ParlAT is planned as a monitor corpus and new material will be added over time. In this phase of the project, we focused on the documents that are available in html. However, we are testing and establishing a workflow to also include the documents in pdf format, once several OCR-related issues have been resolved.

2.1 Coverage and size of the ParlAT beta

The ParlAT beta contains the parliamentary records from 1996 – 2017. However, out of the 960 documents available for this period only 952 documents are included in the corpus, as eight documents are only preliminary verbatim records in pdf format which could not be processed for the first version of the corpus. Therefore, 36% of the available documents have already been processed and are available in the corpus query system. The corpus size is 75,222,970 tokens.

Number of documents	952
Number of tokens	75,222,970
Number of types	585,628
Number of lemmas	123,894

Table 2: Size of the ParlAT beta.

3 Data collection and processing

In the following section we will describe the data collection and data processing for this project. There are a lot of different workflow described in literature for example inter alia Marx 2009, Marx and Schuth 2010, Blessing et al 2015, Blätte 2016.

3.1 Data format

The html documents were scraped from the Austrian parliamentary website and transformed into so called "vertical" or "word-per-line (WPL)" text, because the corpus query system we are using (see section 3.3), requires this input format. In this format, words are written one word per line, so each line contains one word, number or punctuation mark. The "vertical" is a plain text file without any formatting. In this format, the part-of-speech tagging and lemmatization are provided in two additional columns, separated by tabs (Kilgarriff et al.,2004; SketchEngine 2017).

3.2 Metadata, annotation and markup

For the parliamentary records, we only used a reduced set of metadata: type of document, legislative period, date, year and where the original file is stored.

The parliamentary records were part-of-speech tagged and lemmatized using both the TreeTagger and the RFTagger.. Moreover, basic structural markup in form of xml tags were added. Structural information is recorded in the xml element <section> with a @type attribute that can take three different values: "preamble", "sitzung" and "postfix". The section type "preamble" contains general information on the parliamentary session such as legislation period, date, agenda items, notification of sickness or absence of delegates, request or inquires to be treated during the parliamentary session. The section type "sitzung" contains the actual parliamentary session, recording the speakers and the speeches, but also interjections and heckling. The section type "postfix" contains the imprint information.

Another structural element is the <comment> element. Text passages that were set in italics and between brackets in the original documents and contained information other than the utterances of the speakers, were annotated using this element. An example for such 'comments' would be notes on occurrences such as applause from a specific party "*(Beifall bei der SPÖ)*" or interjections from members of a specific party *"(Zwischenruf bei der SPÖ.)" or* information on procedural elements like the president taking the chair *"(Präsidentin Bures übernimmt den Vorsitz).*

The original html files also contain references and links to other documents (such as motions, reports etc.). Furthermore, persons appearing in the documents, such as delegates, are annotated and linked to their profile on the website of the Austrian Parliament. In the transformation into a "vertical" text to be processed in a corpus query tool a lot of this linking information is lost because the corpus query tool cannot process complex markup (Kilgarriff 2004; SketchEngine 2017). However, we tried to keep information such as speaker identification by annotating it with the element <person>. Since the project is in its preliminary phase, we have not yet annotated information on turn-taking, and onset and endings of the utterances of speakers – so when one speaker starts and ends his speech in front of the parliament. However, this is planned as a next step as well as the enrichment with biographical metadata for the speakers.

3.3 Corpus query system and interface

At the moment, internally, we use the SketchEngine as corpus query system.

Figure 2: Concordance of "Antrag" (Motion) in the ParlAT beta within the SketchEngine interface.

However, it is not the most suitable tool when using source text with complex xml markup and integrated links.

4 Intertextuality and referencing

One of the characteristics of parliamentary records is the complex intertextuality with a high frequency of cross-references. These cross references can refer to other parliamentary documents such as motions, reports etc. or they can refer to legal texts and legislation. Therefore, the annotation of these cross-references and linking to the external document is one of the prospects of this project. It has already been mentioned that, while the complex linking system would already be in place in the html documents, it is lost when processing the text in a corpus query system. For this reason, we are testing different components to configure a system where both approaches can be used in parallel, to establish and maintain the links to external documents and to visualise networks, for instance, but also to conduct linguistic queries on the material like in a corpus environment at the same time.

5 Discussion and further work

As stated at the beginning, we are building a corpus of Austrian parliamentary records for different user scenarios within linguistics, political science or history. The reported work is in progress. After finishing the speaker annotation a first version of the corpus – including the documents from 1996 to 2017 – will be published in ARCHE[3] and will made available through the CLARIN infrastructure.

In the second phase, we will start processing the pdf files and we will expand our work on the semantic annotation. Furthermore, we will look into the issue of cross referencing to external documents and to combine the two approaches into one interface: the linking to external resources and the corpus query paradigm.

[3] ARCHE (A Resource Centre for Humanities) is the depositing service of the Austrian Centre for Digital Humanities.

6 Acknowledgments

This work has been partly funded by the Nationalstiftung für Forschung, Technologie und Entwicklung in Österreich.

7 Bibliographical References

Calzada Pérez, María; Marín Cucala, Noemí; Martínez Martínez, José Manuel (2006). ECPC: European Parliamentary Comparable and Parallel Corpora / Corpus Comparables y Paralelos de Discursos Parlamentarios Europeos. *Procesamiento del Lenguaje Natural* 37, 349-350.

Blätte, Andreas (2016). Integrationspolitik im Bundesländervergleich: Die Analyse thematischer Verknüpfungen von Integration auf Basis der PolMine-Plenarprotokollkorpora. Presentation at the the Forum CA[3].CLARIN-D, Hamburg. Available at https://www.clarin-d.de/images/forumca3/4_5_blaette_clarin_hamburg.pdf

Blessing, André; Kliche, Fritz; Heid, Ulrich; Kantner, Cathleen; Kuhn, Jonas. (2015). Computerlinguistische Werkzeuge zur Erschließung und Exploration großer Textsammlungen aus der Perspektive fachspezifischer Theorien. In: Baum Constanze; Stäcker, Thomas (Hrsg.). Grenzen und Möglichkeiten der Digital Humanities. (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1). text/html Format. DOI: 10.17175/sb001_013

Calzada Pérez, María; Marín Cucala, Noemí; Martínez Martínez, José Manuel (2006). ECPC: European Parliamentary Comparable and Parallel Corpora / Corpus Comparables y Paralelos de Discursos Parlamentarios Europeos. *Procesamiento del Lenguaje Natural* 37, 349-350.

Fišer, Darja and Lenardič, Jakob (2017). Overview of Parliamentary Data and Corpora. Available at https://office.clarin.eu/v/CE-2017-1019-Parliamentary-data-report-version-2.pdf

Kilgarriff, Adam, Rychlý, Pavel, Smrž, Pavel and Tugwell, David (2004). The sketch engine. *Information Technology*. 105-116. Available at https://www.sketchengine.co.uk/wp-content/uploads/The_Sketch_Engine_2004.pdf

Marx, Maarten (2010). Advanced Information Access to Parliamentary Debates. *Journal of Digital Information* Vol 10 Nr. 6, Available at https://journals.tdl.org/jodi/index.php/jodi/article/view/668

Marx, Maarten and Schuth Anne (2010). DutchParl: A corpus of parliamentary documents in Dutch. In: Calzolari , Nicoletta et al. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). 3680-3677.

Available at http://www.lrec-conf.org/proceedings/lrec2010/pdf/263_Paper.pdf

SketEngine (2017). Documentation. Available at https://www.sketchengine.co.uk/documentation/

Sippl, Colin; Burghardt, Manuel; Wolff, Christian; Mielke, Bettina (2016). Korpusbasierte Analyse Österreichischer Parlamentsreden. In: Jusletter IT, 25. Februar 2016.

8 Language Resource References

Alexander, Marc and Mark Davies. (2015) Hansard Corpus 1803-2005. Available online at http://www.hansard-corpus.org.

Lapponi, Emanuele and Søyland, Martin G. (2016). Talk of Norway. Available at https://github.com/ltgoslo/talk-of-norway (2016-10-29).

Wissik, Tanja and Pirker, Hannes (2018). ParlAT Corpus. Austrian Centre for Digital Humanities. https://id.acdh.oeaw.ac.at/parlat

A Corpus of Grand National Assembly of Turkish Parliament's Transcripts

Onur Güngör[1], **Mert Tiftikci**[2], **Çağıl Sönmez**[3]
Department of Computer Engineering,
Bogazici University[1,2,3], Istanbul, Turkey
Huawei R&D Center[1], Istanbul, Turkey
onurgu@boun.edu.tr[1], mert.tiftikci@boun.edu.tr[2], cagil.ulusahin@boun.edu.tr[3]

Abstract

In parliaments throughout the world, decisions that are taken directly or indirectly lead to events that affect the society. Eventually, these decisions affect other societies, countries and the world. Thus, transcriptions of these are important to people who want to understand the world, namely historians, political scientists and social scientists in general. Compiling these transcripts as a corpus and providing a convenient way to query the contents is also important from the point of linguists and NLP researchers. Currently, many parliaments provide these transcriptions as free text in PDF or HTML form. However, it is not easy to obtain these documents and search the interested subject. In this paper, we describe our efforts for compiling the transcripts of Grand National Assembly of Turkish Parliament (TBMM) meetings which span nearly a century between 1920 and 2015. We have processed the documents provided by the parliament to the public and transformed them into a single collection of text in universal character coding. We also offer an easy to use interface for researchers to launch custom queries on the corpus on their own. To demonstrate the potential of the corpus, we present several analyses that give quick insights into some of the linguistic changes in Turkish and in Turkish daily life over the years.

Keywords: parliamentary text, corpus, historical records, transcriptions

1. Introduction

As the technological tools for archiving and disseminating text proliferate, we see an increasing number of parliaments across the world that share the transcripts of legislative meetings with the public (Fišer, 2017). This enables a new line of research for humanities and social sciences (Bayley et al., 2004; Cheng, 2015; Georgalidou, 2017; Pančur and Šorn, 2016) and computational linguistics (Mandravickaite and Krilavičius, 2017; Høyland et al., 2014; Grijzenhout et al., 2014; Rheault et al., 2016).

Although parliamentary data is shared with the public, conducting statistical analysis on them is cumbersome in general. This is mainly because they are usually accessed through a search engine where the common workflow is to search for a specific keyword and use search results to investigate the evidence to the specific research question. If the only way of access is through a search engine, it is not possible to calculate statistics of word usage frequency across time or to employ word clustering algorithms besides others which require access to the whole set of documents at once.

In this work, we present our work to address this issue by crawling, processing and combining the transcripts of Grand National Assembly of Turkish Parliament into a single corpus. Our contributions include easier programmatic access to the corpus and several methods to calculate NLP related statistics over the corpus.

The remainder of this paper is organized as follows: in Section 2. we compiled a summary of related work. We explain the details of the process of building the corpus in Section 3. Then, we present a simple analysis of the corpus in Section 4. Finally we conclude in Section 5.

2. Related Work

Although there have been studies in the literature that employ parliamentary data (Vives-Cases and Casado, 2008), studies that compile and process the transcripts to be accessed in a straightforward manner are relatively scarce and do not follow a single format (Verdonik et al., 2013; Graën et al., 2014; Marx et al., 2010).

A recent workshop organized by CLARIN programme aimed to join forces and motivate research on using NLP technologies to make parliamentary data accessible for humanities and social sciences research. The initiative published a report which summarizes the parliamentary corpora in the CLARIN infrastructure (Fišer, 2017).

3. Building the corpus

In this section, we will give details of data preparation and preprocessing phases.

The members of parliament were elected each five years beginning from 1920 until 2007. After 2007, the elections were made every four years. The time period that a parliament is functional after each election is said to form a "term" and is made up of several "lawmaking year"s depending on the actual duration of the "term" which can change due to unscheduled elections or other unexpected events. Every "lawmaking year" is conducted as a series of meetings which we call "sessions". These "sessions" are transcribed by clerks present in the hall in real time. These transcriptions were traditionally published as a periodical called 'Tutanak Dergisi'[1]. With the introduction of digital media, the transcriptions are published on the web as soon as they are redacted in digitized form by the Library, Documentation and Translation Department and Information Technologies Department[2] of the parliament. However, the transcriptions of the sessions before 20th "term" are not published in digitized form, only as scanned images of 'Tutanak Dergisi'. On the other hand, scanned images of Tutanak Dergisi is available for the first 25 terms (1920-2015).

[1] literally 'Journal of Minutes of the Meeting' in English
[2] https://www.tbmm.gov.tr/kutuphane/

So we have chosen to base our work on these scanned images of Tutanak Dergisi to have a corpus which spans 95 years of transcriptions.

The data preparation process can be summarized as the following: We start by crawling the web pages of TBMM. In this phase, we extract the locations of PDF files which contain the transcriptions. We use a command-line tool to extract text from downloaded PDF files. Then we apply a very simple preprocessing operation in which we only get rid of unprintable characters introduced by the text extraction process. We then tokenize the resulting text and obtain the final version of the transcripts. Finally, we compile every document into a single corpus in a reusable format.

3.1. Crawling

Even though we share the scripts which we used for crawling and downloading, we also give the details of the crawling process here for others to replicate.

We used manual labor to obtain the URLs pointing to the scanned images of 'Tutanak Dergisi'. Our effort started with a single visit to a page which contains pointers to every "term" page. After this, we visited every "lawmaking year" page which is accessed through each "term" page. We used a simple browser extension to extract the URLs found in a 'lawmaking year' page.

3.2. Processing

We used `pdftotext` to extract the text contained in each PDF file. `pdftotext` is a tool which is part of `poppler`[3] PDF rendering library. It produced good results in general but sometimes this approach produced erroneous results or no results at all. This is mainly due to the quality of the scanning done when the parliament publishes these files. We only remove spurious characters at the end of lines to obtain the text in free form. We continue with a simple tokenization and conclude our processing by coding the words using a dictionary. We do not strip out or reorder any word during this process.

Our corpus in its current form only records the date of the session. We did not extract the speaker, the context, the political party the speaker belongs to or try to identify other people during the session as suggested in the literature (Marx et al., 2010). However, we made our file format so simple that it is both human and machine readable. Our corpus file basically contains a single document in each line with words in the document in the order as they appeared in the source documents.

The code that is used to crawl and process the corpus can be found in our Github repository[4].

4. Analyzing the corpus

The resulting corpus contains 208 million tokens in 12645 documents which are derived from transcriptions of general assembly sessions between 1920 and 2015. Each document includes data from a session which usually spans a day. We do not include the transcriptions between 2015 and 2018 in this study as they were provided in HTML format as part of a different mode of distribution.

The total number of unique tokens is 619,505, but if we only consider tokens which are found more than 10 times, this figure decreases to 358,286. The number of unique tokens in a given Turkish corpus is usually more than the expected number for other languages which are not morphologically rich thus do not exhibit extensive inflection and derivation. We tested the coverage of our corpus by looking up these unique tokens in a decent Turkish language dictionary[5] from Turkish Language Institute ("Türk Dil Kurumu"). As a result, we found out that about 70% of all unique tokens can be found in the Turkish dictionary. The median number of tokens in a document is 9,642. We give figures summarizing the total number of words and sessions held per year in Figure 1a and 1b. The distribution of document lengths follow an exponential pattern as can be seen in Figure 1c.

The total size of PDF files is 3.9 gigabytes. After we process and encode the words with a dictionary, the size of the resulting file decreases to 1.2 gigabytes. We share the corpus in a compatible format with the scientific community in our source code repository. Alternatively, we will share the corpus through the Virtual Language Observatory (VLO) in the CLARIN infrastructure and as a shared LREC resource.

4.1. Access to the corpus

In addition to serving the processed corpus as a downloadable file, we provide an offline interactive interface suited for use by linguists or social scientists[6].

For implementing such an interface, we employed Project Jupyter[7] and created a Jupyter notebook. A Jupyter notebook is a special file which can be used to mix documentation and sample code. This enables the user to run simple queries and write their own exploration scripts.

4.2. Word and Topic Distributions

We employed several analyses on the corpus to demonstrate the potential areas for research. First, we have employed latent Dirichlet allocation model (Blei et al., 2003), to discriminate words into a predefined number of topics. We set the number of topics to 20. Using this allocation, we could interpret the topic distribution of a given transcription. We examine these allocations to interpret the representation quality of the topics. For example, in Figure 2, we plot the average weights of selected topics in a year. We chose to present these topics because topics 4, 9 and 12 contain words that are considered as old in Turkish. This is validated in the figure. On the other hand, topic 15 and 16 are two topics that can be used to mark transcriptions recorded in 1990's and beyond.

A similar observation can be also done by looking at the plot of the word usage frequencies of the Turkish word 'mebus'. This word of Arabic origin is used to refer to a member of the parliament in old Turkish. It was adopted

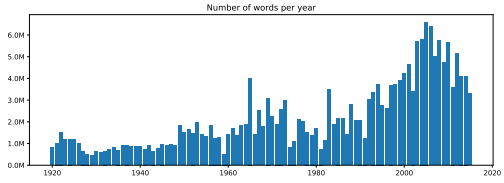
[3] `https://poppler.freedesktop.org/`

[4] `https://github.com/onurgu/turkish-parliament-texts`

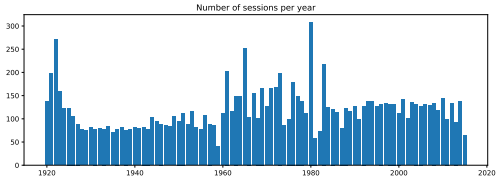
[5] `http://www.tdk.gov.tr/`

[6] Both can be accessed at `https://github.com/onurgu/turkish-parliament-texts/releases`

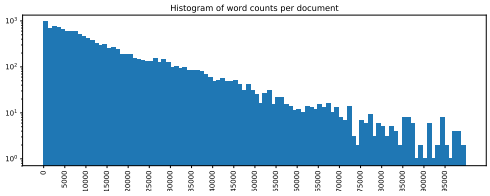
[7] `http://jupyter.org/`



(a) Number of words per year.



(b) Number of sessions per year.



(c) Histogram of word counts per document.

Figure 1: Figures that summarize several statistics about the corpus.

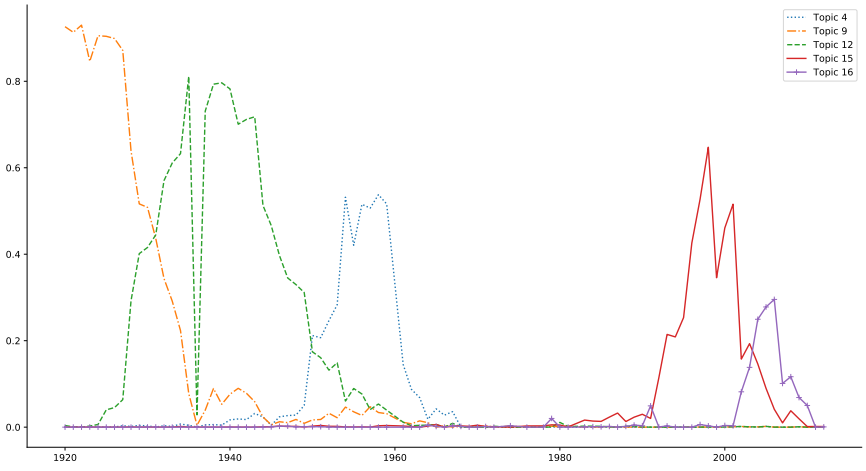


Figure 2: Distribution of selected topics across time.

during the Ottoman era and continued to be used in the republic. It was known that its use decreased through time. We also validate this with a simple query of word counts for the word 'mebus' across the corpus. As can be seen in Figure 3, the usage of 'mebus' is nearly always before 1960. The alternative word 'milletvekili' is mainly used after 1960. This basic plot itself provides an analytical tool for comparing different eras of culture and linguistics. Thus, this shows a good example of the potential of the information contained in the corpus.

In this corpus, there are also traces of introduction of technology in daily life of Turkish citizens. To demonstrate this, we present a comparison between electronic communication devices across the entire corpus ordered by time in Figure 4. The curves in the figure show that television did not become a frequent concept of debate until 1980's. On the other hand, telephone network related issues lost a considerable weight in 1960's. Lastly, we observe the introduction of internet in the parliamentary transcriptions at an increasing pace.

We defer further analytical research to future work. However, we have to note that these analyzes are only scratching the surface. Firstly, due to the high volume of meaningful historical text, we believe that it is possible to conduct comparative linguistics research in Turkish. Second, a wide range of discourse analysis can be done as we can relate every sentence to a specific person belonging to a political party. Moreover, these sentences can be part of a dialogue adding more value to the utterance.

5. Conclusions

In this paper, we present our work on creating a corpus of Grand National Assembly of Turkish Parliament which

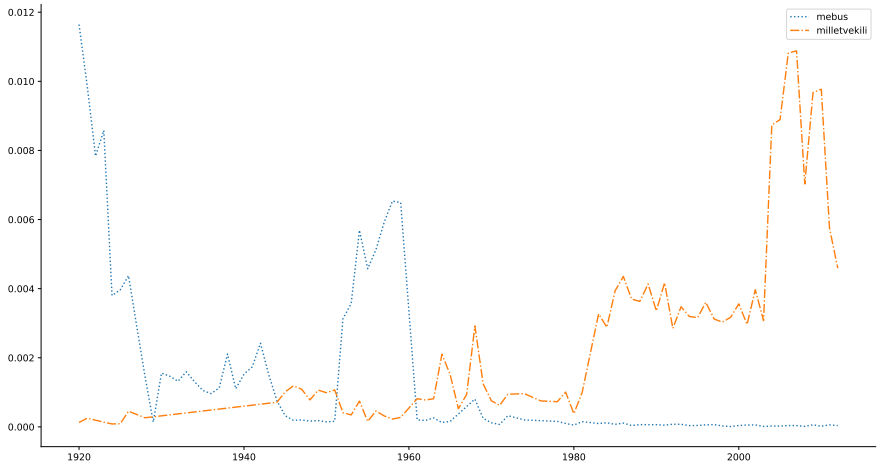


Figure 3: Yearly usage frequencies for 'mebus' and 'milletvekili' across whole corpus. The yearly usage frequency is normalized over all words in a year.

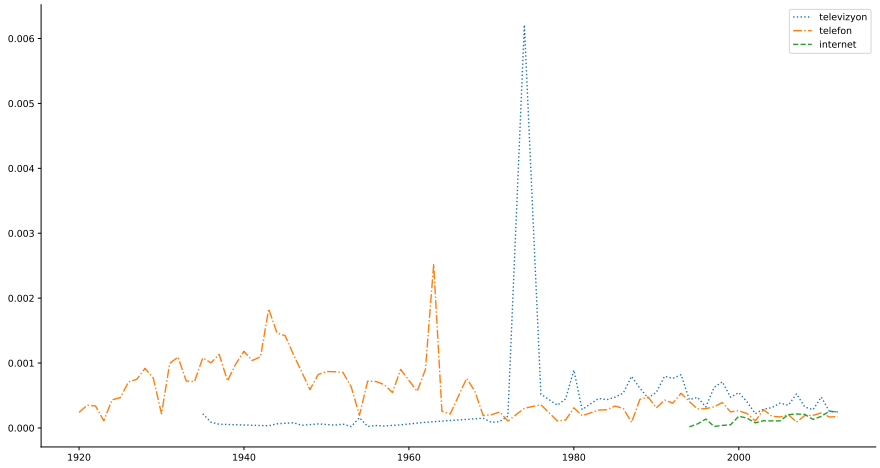


Figure 4: Yearly usage frequencies for 'televizyon', 'telefon' and 'internet' across whole corpus. These words are Turkish translations of 'television', 'telephone' and 'internet' respectively. The yearly usage frequency is normalized over all words in a year.

is intended to be used by social scientists and computational linguists while conducting research on transcriptions of parliamentary sessions. We provide the corpus in digitized form as a single file which can be explored easily with fixed or custom investigative functions.

However, due to the vast amount of work required, we postponed further work such as extensive visualization of documents, extracting person names, political party memberships, mentions of geographical places or buildings and dialogues during the sessions in a structured manner. Also, we omitted the parliamentary sessions between 2015 and today as they were provided in a different format. Future parliamentary sessions will be published in this format. Thus there is work to be done for combining the current version of our corpus with this new source of parliamentary session transcriptions and automatically updating the corpus continuously. Additionally, further spelling correction techniques can be employed to increase the quality of digi-

tization.

6. Acknowledgements

This work is partly supported by the Turkish Ministry of Development under the TAM Project number DPT2007K120610.

7. Bibliographical References

Bayley, P., Bevitori, C., and Zoni, E. (2004). Threat and fear in parliamentary debates in Britain, Germany and Italy. *Cross Cultural Perspectives on Parliamentary Discourse. Amsterdam: John Benjamins*, pages 185–236.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Cheng, J. E. (2015). Islamophobia, Muslimophobia or racism? parliamentary discourses on Islam and Muslims in debates on the minaret ban in Switzerland. *Discourse & Society*, 26(5):562–586.

Fišer, Darja, L. J. (2017). Parliamentary corpora in the CLARIN infrastructure. *CLARIN Annual Conference 2017*.

Georgalidou, M. (2017). Using the Greek parliamentary speech corpus for the study of aggressive political discourse. *CLARIN-PLUS Workshop "Working with Parliamentary Records"*.

Graën, J., Batinic, D., and Volk, M. (2014). Cleaning the Europarl corpus for linguistic applications. In *KONVENS*.

Grijzenhout, S., Marx, M., and Jijkoun, V. (2014). Sentiment analysis in parliamentary proceedings. *From Text to Political Positions: Text analysis across disciplines*, 55:117.

Høyland, B., Godbout, J.-F., Lapponi, E., and Velldal, E. (2014). Predicting party affiliations from European parliament debates. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 56–60. Association for Computational Linguistics.

Mandravickaite, J. and Krilavičius, T. (2017). Stylometric analysis of parliamentary speeches: Gender dimension. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 102–107, Valencia, Spain, April. Association for Computational Linguistics.

Marx, M., Aders, N., and Schuth, A. (2010). Digital sustainable publication of legacy parliamentary proceedings. In *Proceedings of the 11th Annual International Digital Government Research Conference on Public Administration Online: Challenges and Opportunities*, dg.o '10, pages 99–104. Digital Government Society of North America.

Pančur, A. and Šorn, M. (2016). Smart big data: Use of Slovenian parliamentary papers in digital history. *Prispevki za novejšo zgodovino/Contributions to Contemporary History*, 56(3):130–146.

Rheault, L., Beelen, K., Cochrane, C., and Hirst, G. (2016). Measuring emotion in parliamentary debates with automated textual analysis. *PLOS ONE*, 11(12):1–18, 12.

Verdonik, D., Kosem, I., Vitez, A. Z., Krek, S., and Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language Resources and Evaluation*, 47:1031–1048.

Vives-Cases, C. and Casado, D. L. P. (2008). Spanish politicians discourse about the responses to violence against women. *Gaceta sanitaria*, 22 5:451–6.

UKParl: A Data Set for Topic Detection with Semantically Annotated Text

Federico Nanni[1], Mahmoud Osman[1], Yi-Ru Cheng[1], Simone Paolo Ponzetto[1], Laura Dietz[2]
[1]Data and Web Science Group - University of Mannheim, Germany
[2]Computer Science Department - University of New Hampshire, USA

Abstract

We present a dataset created from the Hansard House of Commons archived debates of the UK parliament (2013-2016). The resource includes fine-grained topic annotations at the document level and is enriched with additional semantic information such as the one provided by entity links. We assess the quality and usefulness of this corpus with two benchmarks on topic classification and ranking.

Keywords: topic detection, text as data, text classification, entity linking, ranking

1. Introduction

In recent years, the prompt availability of digital collection of political texts (Koehn, 2005; Vinciarelli et al., 2009; Bachmann, 2011; Cullen et al., 2014; Merz et al., 2016; van Aggelen et al., 2017) has fostered much work in the field of computational social science (CSS), an interdisciplinary field where political science scholars adopt – among other methodologies – Natural Language Processing (NLP) approaches for studying the act and content of political communication (Grimmer and Stewart, 2013).

A task that has attracted large interest in the Computational Social Science community (CSS) is the automatic detection of topics in unstructured text, since this can, in turn, support higher-level tasks such as fine-grained political campaign analyses (Nanni et al., 2016), measuring the agreement between political leaders (Menini et al., 2017) and quantify political attention (Quinn et al., 2010), to name a few.

However, while there is such large availability of digital collections of transcript of campaign speeches and parliamentary debates, social media posts on political events or datasets of party manifestos, most of these collections lack fine-grained annotations of the topics they cover. This limits both the types of analysis that researchers can conduct employing such corpora and the development of benchmarks and evaluation campaigns for testing topic detection algorithms in the political science domain.

Contributions. Consequently, in order to address these issues, we provide the research community with: *a)* a political corpus that we have constructed from the UK parliament Hansard House of Commons archived debates (2013-2016), including fine-grained topic annotations at the document level and entity links; *b)* two different topic prediction benchmarks, in order to foster further research on textual topic detection in the political domain.

2. Related Corpora

One of the first machine-readable resources of transcript of political speeches available to the research community is the well-known *EuroParl* corpus (Koehn, 2005), a collection of parallel texts in 11 languages (later extended to 21 languages (Islam and Mehler, 2012)) created from the proceedings of the European Parliament (EP)[1]. The same collection has been recently made available as linked open data (van Aggelen et al., 2017): *LinkedEP*[2] offers translation of the reports of the plenary meetings of the EP, together with additional metadata information such as the political affiliation of the parliament members, for instance, which is organized in over 25 million triples. Similar resources can be found on the government websites of the United Kingdom[3] and of Italy[4]; regarding the case of the United States, Thomas et al. (2006) presented a corpus of speeches from the US Congress. However, despite thee availability and usefulness for NLP research of such collections (cf. *EuroParl* historically being a core resource for the development of statistical machine translation systems), none of these resources offer fine-grained annotations of the topics addressed in the speeches.

Apart from transcripts of parliamentary debates, another relevant collection of political text is the Manifesto Corpus (Merz et al., 2016)[5], a resource presenting digitized and topically annotated electoral programs that is based on the coding of the Manifesto Project (7 broad categories and more than a hundred fine-grained type of annotations). While researchers have pointed out inconsistencies in the annotations (Mikhaylov et al., 2012), this resource is considered to be one of the biggest human-coded, multilingual, cross-national, open-access corpora in the field of political science. The corpus provides more than 1,800 machine-readable documents, containing more than 600,000 annotated statements as well as metadata like political party affiliations and election year. However, for evaluating a topic detection system the topical annotations of the Manifesto Project remain too coarse-grained: as a matter of fact, instead of describing directly the topic addressed in text (e.g., "refugee crisis"), they map the content to a pre-defined fine-grained category like, for instance "freedom and human rights".

The work closest to ours is that of Bachmann (2011), where the authors conduct a corpus-driven semantic analysis of discourses about same-sex relationships in the UK Parliament. To this end, they create a corpus from the UK Hansard

[1] http://www.statmt.org/europarl/

[2] http://purl.org/linkedpolitics

[3] http://lda.data.parliament.uk/

[4] http://dati.camera.it/

[5] https://manifesto-project.wzb.eu/

Table 1: Corpus Statistics.

Session	# Speech	# Topic	# Token	# Entity
2013-14	23,935	2,343	175,604	72,791
2014-15	19,439	1,987	166,777	72,248
2015-16	26,605	1,923	169,119	74,678
Total	69,979	5,634	354,403	125,886

Archives[6] consisting of 16 electronic debates transcripts from both houses of the parliament: 9 debates from the House of Lords, and 7 from the House of Commons. In our work, we consider the same archive, but we collected all materials available between 2013 and 2016, which sum up to around 70,000 speeches and more than 5,600 topics, as presented in Table 1.

3. Corpus Overview

In order to create the corpus, we collected all transcripts of speeches made on the House of Commons floor between 2013 and 2016. Speeches have been manually associated with a single topic (e.g., 'Isis', 'Zika Virus', 'Greece Financial Crisis', etc.) by the curators of the corpus. In order to enable analyses leveraging background knowledge, we additionally aligned each topic, whenever possible, with the related Wikipedia page, for instance 'Isis' with /wiki/Islamic_State_of_Iraq_and_the_Levant. Given the large number of speeches in the corpus and the fact that the associated topics are often clearly defined (e.g., 'EU Sanctions (Russia)', 'Northern Ireland Political Situation'), this was done automatically by employing the topic as a query and matching it with the first retrieved page, using the Wikipedia search-tool. However, we are aware that for potentially ambiguous topics (e.g., 'Voting System', 'Foreign Students') or topics without a related Wikipedia page (e.g., 'Wi-fi in Hospitals') this approach could generate inconsistencies. We aim to address this issue in the future with the support of human annotators.

The dataset[7] follows the structure of the original collection and it is organized in three sessions: 2013-14, 2014-15 and 2015-16. Each session is divided into a set of topics, where for each topic-speech pair we provide i) the original text of the speech; and ii) the list of entities that were identified in text (we use TagMe (Ferragina and Scaiella, 2010) with standard settings). The number of unique speeches, topics, tokens and entities in the corpus are presented in Table 1. The alignment between the topic and the related Wikipedia page is provided in an accompanying file.

4. Topic Classification Benchmark

Numerous supervised models have been proposed in the past for the classification of political text (Purpura and Hillard, 2006; Stewart and Zhukov, 2009; Verberne et al., 2014; Karan et al., 2016; Zirn et al., 2016; Glavaš et al., 2017a, *inter alia*). Inspired by these works, we test different feature vector representations of text and classification algorithms to provide a benchmark for this task on our corpus.

[6] http://www.parliament.uk/business/publications/

[7] http://federiconanni.com/ukparl

4.1. Feature vector representations

We compare four different ways of processing the text and transforming it into feature vectors.

TF-IDF (words). Standard TF-IDF (logarithmic, L2-normalised variant) vectors of documents (tokenized and lemmatized).

TF-IDF (entities). We used TagMe! to identify, disambiguate and link entities in text. We then compute entity-based TF-IDF vectors by considering each document as a bag of entities.

Word embeddings. As in previous work (Glavaš et al., 2017b), the document embedding representation of each speech is computed as the element-wise average of the embeddings of the words in the text. Let W be the set of unique words in a document D. The embedding of D is then computed as:

$$\frac{1}{N}\sum_{w\in W}\text{freq}(w)\cdot\vec{v}_w$$

where $\text{freq}(w)$ is the frequency of word w, $\vec{v}_w$ is its embedding vector, and N is the total number of unique words in D. For this, we use the state-of-the-art pre-computed GloVe word embeddings (300d)[8].

Entity embeddings. As in the case of word embeddings, we computed the vector as the element-wise average average of the embeddings of the unique entities in the text. We use state-of-the-art pre-computed RDF entities embeddings (Ristoski et al., 2016).

4.2. Classifiers

We compare the performance of four different classifiers all implemented in the Python library Scikit-Learn[9].

NB. A standard multinomial Naive-Bayes classifier.

Nearest Centroid. This memory-based classifier first creates a centroid for each topic, and then assigns each example to the topic whose centroid is closest, based on the euclidean distance between the feature vectors.

***k*-NN.** A standard k-Nearest Neighbors classifier that labels each example with the majority class of the k[10] most similar labeled documents, based on the euclidean distance between the feature vectors.

SVM. A Support Vector Machine using a linear kernel, with standard parameters (C=1.0).

4.3. Dataset

We evaluate the performance of each pair of document representation and classifier on two different sets of speeches.

2015-16. We first select for testing the largest subset of our collection, namely all speeches addressed in the session 2015-16. Among the most relevant topics there are

[8] https://nlp.stanford.edu/projects/glove/

[9] http://scikit-learn.org/

[10] During testing we obtain consistently good performance using 10 neighbors.

Table 2: Results on topic prediction (2015-2016 subset)

Doc. Representation	Classifier	Topic Prediction			
		Macro			Micro
		P	R	F_1	F_1
TF-IDF (words)	NB	0.18	0.14	0.15	0.17
	NearestCentroid	**0.52**	**0.49**	**0.50**	**0.46**
	k-NN	0.41	0.42	0.41	0.42
	SVM	0.49	0.39	0.43	0.44
TF-IDF (entities)	NB	0.10	0.09	0.09	0.10
	NearestCentroid	**0.30**	**0.30**	**0.30**	**0.28**
	k-NN	0.22	0.23	0.22	0.24
	SVM	0.27	0.25	0.25	**0.28**
Word embeddings	NB	0.31	0.28	0.29	0.24
	NearestCentroid	0.33	**0.33**	**0.33**	0.33
	k-NN	0.26	0.27	0.26	0.29
	SVM	**0.36**	0.31	**0.33**	**0.38**
Entity embeddings	NB	0.16	0.18	0.16	0.15
	NearestCentroid	0.23	0.23	0.23	0.21
	k-NN	0.17	0.18	0.17	0.20
	SVM	**0.27**	**0.25**	**0.25**	**0.28**

Table 3: Results on topic prediction (complete corpus).

Doc. Representation	Classifier	Topic Prediction			
		Macro			Micro
		P	R	F_1	F_1
TF-IDF (words)	NB	0.12	0.10	0.10	0.13
	NearestCentroid	**0.36**	0.33	**0.34**	0.36
	k-NN	0.22	0.22	0.22	0.27
	SVM	0.34	**0.35**	**0.34**	**0.38**
TF-IDF (entities)	NB	0.06	0.06	0.06	0.07
	NearestCentroid	**0.17**	**0.16**	**0.16**	**0.17**
	k-NN	0.10	0.11	0.10	0.13
	SVM	0.15	0.16	0.15	0.16
Word embeddings	NB	0.16	0.16	0.16	0.17
	NearestCentroid	**0.22**	**0.22**	**0.22**	0.22
	k-NN	0.15	0.16	0.15	0.19
	SVM	0.19	0.17	0.16	**0.25**
Entity embeddings	NB	0.09	0.10	0.08	0.09
	NearestCentroid	0.13	0.13	0.13	0.13
	k-NN	0.10	0.11	0.09	0.13
	SVM	**0.14**	**0.14**	**0.14**	**0.19**

the Scotland Bill, Brexit, the war in Syria, immigration in UK and the economic crisis in Greece. We excluded the general topics 'Topical Questions', 'Business of the House' and 'Engagements' and topics with less than 10 speeches; the final dataset is composed by 490 topics and more than 15,000 speeches.

All. The second benchmark is instead composed of speeches from all four sessions. Here we also removed the general topics mentioned above and those with less than 10 speeches. The final data collection consists of a total of 1,341 topics and more than 41,000 speeches.

4.4. Results

The results of our benchmark (precision, recall and F1-Score) are presented in Table 2 and 3. As it can be seen, in both cases the use of lexical features (TF-IDF) outperforms semantic approaches based on word embeddings or the use of entity links. This is mainly due to the size of the documents analyzed, which makes it difficult to represent them with a single embedding vector maintaining their meaning. Among the different classifiers that we tested, the best performance have been achieved in both datasets by the Nearest Centroid and the Support Vector Machine.

Table 4: Topical Ranking task on dataset.

	MAP	P@1
Baseline (Random)	0.13	0.04
Entity frequency	**0.37**	**0.24**
Entity TF-IDF	**0.37**	**0.24**
Centroid (embeddings)	0.20	0.10
Position (doc. order)	0.23	0.09
Position + frequency	0.24	0.14
Position + TF-IDF	0.22	0.11
Position + centroid	0.22	0.12

5. Topic Ranking Benchmark

There are many different ways of predicting in an unsupervised way the topic addressed in a political text (Grimmer and Stewart, 2013). In our setting, the topic of each document is represented by its aligned (Wikipedia) entity, such as, for instance, /wiki/European_Migrant_Crisis. This task has been already approached by the NLP community, for example in Hulpus et al. (2013) and in Lauscher et al. (2016) by combining entity linking and topic models. As already noticed in previous work (Hulpus et al., 2013), it is often the case that the topic of the document is not directly mentioned in the text. In our case we noticed that only 22% of the documents (15,581 documents) in our collection mention the entity that is assigned as its topic label. When considering this subset, the task of predicting the topic is similar to that of the entity salience (Dunietz and Gillick, 2014).

5.1. Ranking Approaches

Inspired by previous works, we present the results of our evaluation regarding topic-label ranking comparing different baseline approaches over the Topic Ranking benchmark.

Entity frequency. We rank entities in the document by their frequency of mentions. This follows the intuition that the topic of a document is probably often mentioned in a text.

Entity TF-IDF. Following previous work (Lauscher et al., 2016), we additionally weight the raw frequency of entities by their inverse document frequency (i.e., standard TF-IDF).

Centroid (embeddings). We compute for each document its centroid on the basis of its entity embeddings (Ristoski et al., 2016). Entities are ranked by their distance to the centroid.

Position-based ranking. Inspired by Dunietz and Gillick (2014), we consider entities mentioned at the beginning of the document (in our case the first 10 entities), and rank them by their order of appearance (**Position**). We additionally experiment with alternative ranking functions, namely on the basis of raw frequency of occurrence (**Position + frequency**), a standard TF-IDF weighting scheme (**Position + TF-IDF**), or distance to their centroid computed on the basis of the entity embeddings (**Position + Centroid**).

5.2. Results

We present the results of our benchmark on topic ranking in Table 4, where we quantify performance using standard ranking-sensitive metrics like Mean Average Precision (MAP) and Precision@1. As it can be noticed, for both

metrics the best baseline approaches rely on ranking entities based on raw or weighted (TF-IDF) frequency. Instead, the use of a centroid as well as the adoption of the heuristic presented in Dunietz and Gillick (2014) do not lead to good results, showing the complexity of the task. Based on these initial findings, we will explore in future works how to identify the topic of a document when this is not explicitly mentioned in the content. A possible approach could, for instance, employ relatedness measures to retrieve additional entities from the knowledge base, as already done in similar tasks by Hulpus et al. (2013) and Weiland et al. (2016).

6. Conclusion

In this paper we presented a dataset of political speeches addressed at the UK House of Commons (2013-2016), with fine-grained topic annotations at the document level and enriched with entity links. The corpus is accompanied by two benchmarks on topic classification and ranking.
We envision the use of this dataset and benchmarks for supporting future interactions between the NLP and CSS communities in developing and testing together new algorithms for addressing the topic detection task.

Acknowledgments

This work was funded in part by a scholarship of the Eliteprogramm for Postdocs of the Baden-Württemberg Stiftung (project "Knowledge Consolidation and Organization for Query-specific Wikipedia Construction") and was also supported by the SFB 884 on the Political Economy of Reforms at the University of Mannheim (project C4), funded by the German Research Foundation (DFG).

7. Bibliographical References

Bachmann, I. (2011). Civil partnership - "Gay marriage in all but name". *Corpora*, 6(1).

Cullen, A., Hines, A., and Harte, N. (2014). Building a Database of Political Speech: Does Culture Matter in Charisma Annotations? *Proc. of Audio/Visual Emotion Challenge*.

Dunietz, J. and Gillick, D. (2014). A new entity salience task with millions of training examples. In *Proc. of EACL*, pages 205–209.

Ferragina, P. and Scaiella, U. (2010). TAGME: One-the-fly Annotation of Short Text Fragmetns (by Wikiepdia Entities). *Proc. of CIKM*.

Glavaš, G., Nanni, F., and Ponzetto, S. P. (2017a). Cross-lingual classification of topics in political texts. In *Proc. of NLP+CSS*.

Glavaš, G., Nanni, F., and Ponzetto, S. P. (2017b). Unsupervised cross-lingual scaling of political texts. In *Proc. of EACL*.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Hulpus, I., Hayes, C., Karnstedt, M., and Greene, D. (2013). Unsupervised graph-based topic labelling using dbpedia. In *Proc. of WSDM*, pages 465–474.

Islam, Z. and Mehler, A. (2012). Customization of the europarl corpus for translation studies. In *Proc. of LREC*.

Karan, M., Širinić, D., Šnajder, J., and Glavaš, G. (2016). Analysis of policy agendas: Lessons learned from automatic topic classification of Croatian political texts. In *Proc. of LaTeCH*.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the MT Summit*, pages 79–86.

Lauscher, A., Nanni, F., Ruiz Fabo, P., and Ponzetto, S. P. (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):67–88.

Menini, S., Nanni, F., Ponzetto, S. P., and Tonelli, S. (2017). Topic-based agreement and disagreement in us electoral manifestos. In *Proc. of EMNLP*.

Merz, N., Regel, S., and Lewandowski, J. (2016). The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*, 3(2):1–8.

Mikhaylov, S., Laver, M., and Benoit, K. R. (2012). Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis*, 20(1):78–91.

Nanni, F., Zirn, C., Glavaš, G., Eichorst, J., and Ponzetto, S. P. (2016). TopFish: topic-based analysis of political position in US electoral campaigns. In *Proc. of PolText*.

Purpura, S. and Hillard, D. (2006). Automated classification of congressional legislation. In *Proc. of dg.o*, pages 219–225.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., and Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Ristoski, P., Rosati, J., Di, T., Leone, R. D., and Paulheim, H. (2016). RDF2Vec : RDF Graph Embeddings and Their Applications. *IOS Press*.

Stewart, B. M. and Zhukov, Y. M. (2009). Use of force and civil–military relations in russia: an automated content analysis. *Small Wars & Insurgencies*, 20(2):319–343.

Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proc. of EMNLP*.

van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., and Beunders, H. (2017). The debates of the european parliament as linked open data. *Semantic Web*, 8(2).

Verberne, S., Dhondt, E., van den Bosch, A., and Marx, M. (2014). Automatic thematic classification of election manifestos. *Information Processing & Management*, 50(4):554–567.

Vinciarelli, A., Dielmann, A., Favre, S., and Salamin, H. (2009). Canal9: A database of political debates for analysis of social interactions. In *Proc. of ACII*, pages 1–4.

Weiland, L., Hulpus, I., Ponzetto, S. P., and Dietz, L. (2016). Understanding the message of images with knowledge base traversals. In *Proc. of ICTIR*, pages 199–208. ACM.

Zirn, C., Glavaš, G., Nanni, F., Eichorts, J., and Stuckenschmidt, H. (2016). Classifying topics and detecting topic shifts in political manifestos. In *Proc. of PolText*.

EuroParl-UdS: Preserving and Extending Metadata in Parliamentary Debates

Alina Karakanta, Mihaela Vela, Elke Teich
Department of Language Science and Technology, Saarland University
alina.karakanta@uni-saarland.de, {m.vela, e.teich}@mx.uni-saarland.de

Abstract

Multilingual parliaments have been a useful source for monolingual and multilingual corpus collection. However, extra-textual information about speakers is often absent, and as a result, these resources cannot be fully used in translation studies. In this paper we present a method for processing and building a parallel corpus consisting of parliamentary debates of the European Parliament for English into German and English into Spanish, where original language and native speaker information is available as metadata. The paper documents all necessary (pre- and post-) processing steps for creating such a valuable resource. In addition to the parallel corpora, we collect monolingual comparable corpora for English, German and Spanish using the same method.

Keywords: parallel corpus, comparable corpus, European Parliament, metadata, multilingual

1. Introduction

Multilingual parliaments have been a useful source for monolingual and multilingual corpus collection. However, it is often the case that the compilation of these corpora is not transparent or that useful information about speakers and the status of a given speech (original vs. translation) is absent. Consequently, parliamentary corpora cannot be directly used for research on translation.

An important and probably the earliest attempt to create a parallel corpus from parliamentary proceedings is the Canadian Hansard corpus. It consists of transcripts of debates of the Canadian Parliament (annotated with metadata about the original language) in the two official languages of Canada, English and French. Similarly, several attempts have been made to collect and structure the proceedings of the European Parliament. One of the most popular collections of European parliamentary proceedings is EuroParl (Koehn, 2005), which has been widely used for machine translation[1] and cross-lingual research (Cartoni et al., 2013). It consists of transcribed and revised spoken utterances by speakers of the European Parliament (EP), translated into several languages. Although the monolingual subcorpora of EuroParl often include metadata about the original language of the sentences, this information is not always consistent and it is completely absent from the bilingual corpora. For this reason, EuroParl might be suitable for training MT systems, but for other tasks manipulation of the data is often required. For this reason, other projects have focused on correcting and structuring EP proceedings for linguistic applications (cf. Corrected and Structured EuroParl corpus (Graën et al., 2014), European Comparable and Parallel Corpora (Calzada Pérez et al., 2006), Digital Corpus of the European Parliament (Hajlaoui et al., 2014), Talk of Europe – Travelling CLARIN Campus/LinkedEP (van Aggelen et al., 2017)).

For translationese research, parliamentary proceedings have to be structured as parallel corpora where the translation direction is known. Most of the previous projects on this field rely on the "language" tag to extract sentences produced in the original language from EuroParl (Lembersky et al., 2012b), even though this information is scarce and sometimes inconsistent, as shown by Cartoni and Meyer (2012). Rabinovich et al. (2015) compile a cross-domain corpus for translationese research annotated with metadata about the translation direction. In later work, Rabinovich et al. (2017) attempt to preserve the traits of the original author in the extracted corpus in order to measure the signals left by the author's gender in original and translated text. Nisioi et al. (2016) create a monolingual English corpus of native, non-native and (human) translated texts extracted from the EP proceedings. The corpus is a subset of the corpus collected by Rabinovich et al. (2015) and preserves, similar to our corpus, metadata about the speaker.

Contrary to these approaches, we provide a complete pipeline to collect and compile European Parliament debates into a high-quality, metadata-rich corpus, with accurate speaker and language information, useful for a variety of natural language processing (NLP) tasks.

The paper is structured as follows. Section 2. presents the motivation for building such a resource. Section 3. describes the processing steps, including crawling the web, sorting and filtering the crawled data. In this section we also give an overview on the metadata as well as the corpus structure and statistics. Section 4. discusses possible applications of such a corpus in the field of translation studies and Section 5. provides a brief summary and conclusion.

2. Motivation

Motivations for building a resource as the one described lie in the intended context of use. Machine translation can profit from such a resource, since it has been shown that for statistical machine translation (SMT) direction-aware translation models yield better translation quality than models trained on texts in the opposite direction (Kurokawa et al., 2009; Lembersky et al., 2012a).

Translation studies, in particular research on the specific properties of translations, is a research field that can profit from such a resource. Research on (human and machine) translations has shown that translations exhibit specific properties, such as simplification, explicitation, normaliza-

[1] http://statmt.org/moses/

tion, shining-through etc., also known as "translationese" (cf. Baker (1995; Laviosa (1998; Teich (2003; Volansky et al. (2015)). The only factor taken into consideration in this kind of studies is, by now, translation direction. As shown by Koppel and Ordan (2011) translationese research should incorporate other relevant factors, too, including information on the speaker (native vs. non-native) or production mode (written vs. spoken).
Other NLP research fields such as gender identification (Koppel et al., 2002) or topic detection (Yang et al., 2011; Blei, 2012) might also benefit from metadata-rich corpora. For example, information about the affiliation of a speaker to a specific party or parliamentary group interconnected with information about the country they represent, allows for detecting common (or different) topics at party, group, national or European level.

3. Corpus Processing

In this section, we describe a pipeline for building a comparable/parallel corpus from European Parliament debates. It is based on meta-information on the proceedings and the Members of the European Parliament (MEP). Our final goal is to obtain:

(i) a **parallel corpus** where the source language (SL) sentences come from native SL speakers and are aligned to sentences in the required target language (TL) and

(ii) a **comparable monolingual corpus** of the target language, where the sentences come from native TL speakers.

The process of building the corpus can be described in the following steps:

1. Download proceedings in HTML
2. Download MEPs' metadata in HTML
3. Extract MEPs' information in a CSV file
4. Model proceedings as XML
5. Filter out text units not in the expected language
6. Add MEPs' metadata to proceedings
7. Add sentence boundaries
8. Annotate token, lemma, Part-of-Speech
9. Separate originals from translations and filter by native speakers
10. Extract text into raw format
11. Sentence-align the resulting corpus

Even though this is an end-to-end pipeline, some steps are independent from each other. For example, step 11 applies only to create a raw sentence-aligned parallel corpus, suitable for MT experiments, while step 8 is optional and can be applied at any point.

3.1. Crawling the Data

The data to compile the corpus was collected from the official website of the European Parliament[2]. A typical URL for the proceedings of a given day consists of the base URL, a date and the language version. To date our method provides support only for English, German and Spanish, but it can be easily localized by simply translating the roles (e.g. president, commissioner) in the required language. It is also possible to determine a specific date range.

```
language_version = en #choose language
if list_of_dates is True:
    read(list_dates)
else:
    generate_range_of_dates(list_dates)
for date in list_dates:
    generate(URL)
    request(URL)
    if URL is True:
        download(document)
    else:
        proceed_next_date(date)
```

Figure 1: Pseudocode for crawling the proceedings

Following the process shown in Figure 1, we collected URLs with dates between 20/07/1999 and 18/01/2018. The format of the obtained data is HTML, which allows us later to preserve meta-information as XML. In addition to the proceedings, the European Parliament website maintains a database with all MEPs[3]. We obtained MEPs' information, such as basic information about the speaker and their history record, also in HTML.

3.2. Metadata

There are two types of metadata collected for the purpose of this corpus:

(a) **Proceedings' metadata**
Proceedings' metadata is basic metadata about the parliamentary session. As depicted in Figure 2, a session is divided into several sections, i.e. agenda items, which are then subdivided into interventions. Information is also obtained about the speakers and the source language of the text. Lastly, the metadata contains the actual text of the proceedings as paragraphs.

(b) **MEPs' metadata**
Basic metadata is extracted about each MEP, such as nationality, political affiliation with the European Parliament and with the national parties. As shown in Figure 3, the information is split into 3 categories:

- meps.csv: basic information about the MEP
- national_parties.csv: political affiliation of the MEP in his/her country
- political_groups.csv: political affiliation at the European Parliament

[2] http://www.europarl.europa.eu/
[3] http://www.europarl.europa.eu/meps/en/map.html

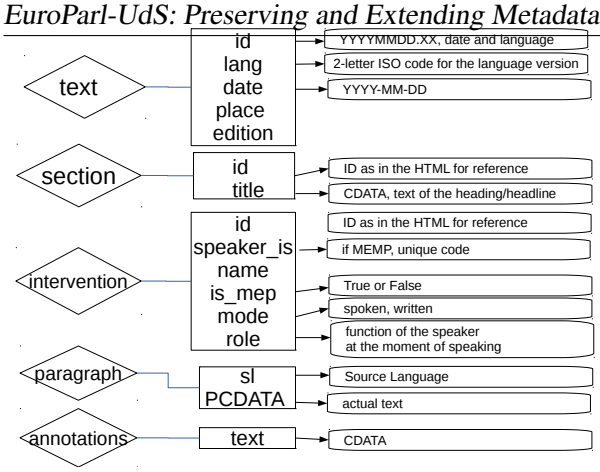


Figure 2: Metadata structure for the proceedings. The words in the diamonds represent the tags and the words in the squares the attributes under each tag. The third column contains the description of each attribute.

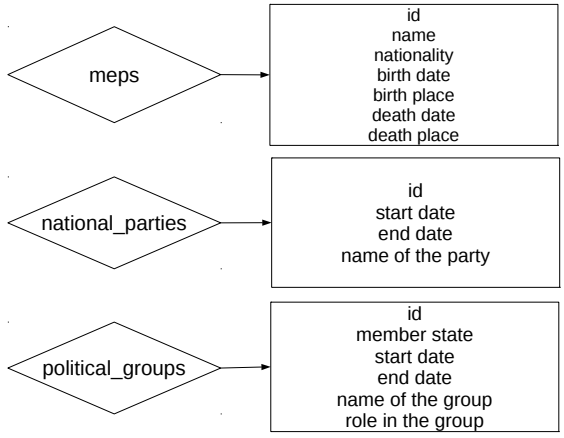


Figure 3: Metadata structure for information about MEPs. The words in the diamonds represent the tags and the words in the squares the attributes under each tag.

It should be noted that not all speakers before the European Parliament are MEPs. There are also members of other European institutions, representatives of national institutions, guests, etc. There is currently no metadata for them, but the information about these speakers is extracted from the proceedings. For each proceeding in XML we retrieve all interventions whose speaker is an MEP. Then we add relevant speaker's metadata to the intervention.

3.3. Sorting the Data

Since our goal is to achieve maximum quality of the data obtained, we employ a series of sorting and filtering techniques to clean the data and preserve the utterances that best serve our tasks.

As a first step, we filter out text units not in the expected language. Interventions sometimes remain untranslated and thus their text appears in their original language. In order to avoid this noise, we identify the most probable language of each text unit and remove the paragraphs which are not in the expected language (e.g. Bulgarian fragments found in the English version) using the Python language identifiers *langdetect*[4] and *langid*[5] and a series of heuristics.

Secondly, we filter out interventions to preserve only sentences by native speakers. A native speaker is defined here as someone holding the nationality of a country with the source language as official language. For English we filter MEPs whose origin is United Kingdom, Ireland or Malta, for German Germany, Austria, Belgium, Luxembourg and Italy, and for Spanish Spain.

An optional step is to perform Part-of-Speech tagging and lemmatization using TreeTagger (Schmid, 1994).

3.4. Sentence Alignment

The creation of a parallel corpus requires sentenced-aligned data with one sentence per line. For this task we employ hunalign (Varga et al., 2005), an automatic sentence aligner. First, we split the text into sentences using NLTK's Punkt tokenizer. Then, we extract the text from the XML files and write to files one sentence per line based on intervention, in the format of filename.intervention_id.lang e.g. *1999720.2-202.en*. This is particularly important for alignment quality as each intervention is a small text unit and aligning a few sentences per time yields higher accuracy than aligning a full text. Before aligning the sentences, we tokenized the text using Moses tokenizer with the specific setting for each language. Then, the interventions are aligned. Since we wish to obtain the highest possible quality, we set a confidence threshold of 30 for the aligned sentences and rerun the alignment based on the dictionary built in the first alignment round. A numerical ladder file is created, based on which we perform the final alignment on the untokenized files. Finally, the resulting alignments are concatenated in one file for the source and one for the target language to create the parallel corpus.

3.5. Corpus Structure and Statistics

The corpus is structured according to the steps followed for its compilation. For every step, the files generated are stored in a specified folder so that they can be used for any suitable task. At the time of compilation, the corpus consists of 1077 files for English, German and Spanish, while the parallel and the comparable corpora are in a one-file raw format that can be used directly for training an MT system. The final corpus structure is shown in Table 1.

The statistics for the comparable and the parallel corpus for the three supported languages are presented in Table 2 and Table 3. In Table 2, the language identifier method filters out texts not in the required language, while still preserving a large amount of data. The application of factors relevant for translation, both for the comparable and the parallel corpora provides us useful information about the language preferences of the speakers in the Parliament. Of course, neither all sentences in a specific language are produced by native speakers of this language, nor all sentences are translated into all languages. For this reason, filtering

[4]https://pypi.python.org/pypi/langdetect
[5]https://pypi.python.org/pypi/langid

Directory	Description
html	The crawled proceedings and MEPs' information in HTML
metadata	MEPs' metadata in CSV
txt	Raw text of the proceedings
xml	Proceedings transformed from HTML to XML
xml_langid	Proceedings in XML where the text not in the expected language is filtered out
xml_metadata	Proceedings in XML with added MEPs' metadata
xml_sentences	Proceedings in XML where text is split into sentences
xml_translationese	Proceedings in XML filtered by factors relevant for translation – original, translation, native speaker For each language *a*, it contains · the originals in *a*, · the originals in *a* only by native speakers, · all translations from any language into *a* and · all translations into *a* from a specific *SL* where the speakers are native speakers of the *SL*
xml_ttg	PoS-tagged and lemmatized proceedings in XML
raw_parallel	For each language the corresponding parallel corpora
raw_comparable	For each language the comparable corpus of original texts by native speakers

Table 1: Corpus structure

	EN		DE		ES	
	words	sents	words	sents	words	sents
html	95.21 M	5.11 M	91.48 M	5.25 M	97.08 M	5.19 M
xml	95.60 M	5.11 M	92.43 M	5.27 M	97.33 M	5.17 M
langidfilter	65.55 M	3.23 M	40.23 M	2.63 M	51.32 M	2.49 M
translationese_orig	19.69 M	0.84 M	11.74 M	0.68 M	10.75 M	0.37 M
translationese_native	8.67 M	0.37 M	7.86 M	0.42 M	5.66 M	0.18 M

Table 2: Statistics of the comparable corpora after every processing step

	EN→DE		EN→ES	
	words	sents	words	sents
all	42.08 M/38.93 M	1.91 M	42.11 M/44.21 M	1.87 M
translationese_orig	6.43 M/6.22 M	296.7 K	5.75 M/6.18 M	249 K
translationese_native	3.18 M/3.10 M	137 K	2.93 M/3.15 M	125 K

Table 3: Statistics of the parallel corpora after every processing step

sentences produced in the original language (non-translated texts) shows that only 20%-30% of the sentences in the supported languages are originals, while around 50% of the originals are produced by native speakers. In spite of this, the pipeline described above still provides us with a high quality and significant in size dataset, useful for a variety of applications.

4. Possible Applications

A corpus as described in this paper is a valuable resource for various kinds of applications. One application is machine translation, for which a metadata-rich corpus allows a more principled data selection, which in some cases has been shown to be more beneficial than using all the data available both for phrase-based as well as neural machine translation (Axelrod et al., 2011; Gascó et al., 2012; van der Wees et al., 2017).

Another application is human translation, e.g. modelling translational choice. Using the EuroParl-UdS, in our ongoing research we employ the noisy channel model as commonly applied in machine translation. According to Equation 1)

$$\arg\max_t p(t|s) = \arg\max_t p(s|t)p(t) \quad (1)$$

translation is described by maximizing the product of the probability of a TL expression t given a SL expression s by maximizing

(i) the probability of a SL expression s given a TL expression t and

(ii) the probability of a TL expression t on its own, i.e. without being conditioned by s.

This matches exactly the human translator's goal of reaching a high level of translation adequacy by maximizing the

fidelity to SL (i.e. high likelihood that the SL expression is a match for a particular TL expression) and the conformity with TL expectations (i.e. high probability of the chosen translation solution in the context of the TL) and can therefore be taken as a basis for modelling human translational choice (Teich and Martínez Martínez, forthcoming). Furthermore, we employ the corpus in studies of translation entropy, comparing the range and distribution of translation options in professional productions in EuroParl-UdS with learner translations for analysis of translation difficulty in different translation learner groups (Martínez Martínez and Teich, 2017).

5. Summary and Conclusion

We have presented an approach to building and processing parallel corpora consisting of parliamentary debates of the European Parliament (EP) harvesting valuable metadata such as speaker status and translation direction. Existing corpora built from EP proceedings do not contain such metadata, which impedes their use in translation studies or variationist linguistic analysis. We have shown our approach at work for English into German and English into Spanish parallel corpora as well as corresponding monolingual comparable corpora, but the approach is generic and can be applied to any language pair.

A metadata-rich resource such as the EuroParl-UdS is valuable for various NLP tasks and it is crucial for the advancement of insights into the process of human translation, where we need to know as much as we can about the production conditions, including the status of a given text (original vs. translation) and information about the speaker. In addition, our complete and fully documented pipeline can be easily used to compile metadata-rich or raw, parallel and comparable corpora for various linguistic applications. The corpus is available at CLARIN-PID[6] under licence CC-BY-SA-NC-4.0; the scripts are available on GitHub at https://github.com/hut-b7/europarl-uds.

6. Acknowledgements

We would like to thank our colleague José Manuel Martínez Martínez for his support - by providing us his scripts (https://github.com/chozelinek/europarl) for crawling the data - while building this resource.

7. References

Axelrod, A., He, X., and Gao, J. (2011). Domain Adaptation via Pseudo In-domain Data Selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, PA, USA. Association for Computational Linguistics.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2):223–245.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Calzada Pérez, M., Marín Cucala, N., and Martínez Martínez, J. M. (2006). ECPC: European Parliamentary Comparable and Parallel Corpora / Corpus Comparables y Paralelos de Discursos Parlamentarios Europeos.

Cartoni, B. and Meyer, T. (2012). Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies. In *LREC, 2012*, page 6, Istanbul.

Cartoni, B., Zufferey, S., and Meyer, T. (2013). Using the europarl corpus for cross-linguistic research. *Belgian Journal of Linguistics*, 27(1):23–42.

Gascó, G., Rocha, M.-A., Sanchis-Trilles, G., Andrés-Ferrer, J., and Casacuberta, F. (2012). Does More Data Always Yield Better Translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 152–161, Stroudsburg, PA, USA. Association for Computational Linguistics.

Graën, J., Batinic, D., and Volk, M. (2014). Cleaning the europarl corpus for linguistic applications.

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., and Varga, D. (2014). Dcep-digital corpus of the european parliament. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, volume 5, pages 79–86, Phuket, Thailand. AAMT.

Koppel, M. and Ordan, N. (2011). Translationese and Its Dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326. Association for Computational Linguistics.

Koppel, M., Argamon, S., and Shimoni, A. R. (2002). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4):401–412.

Kurokawa, D., Goutte, C., and Isabelle, P. (2009). Automatic detection of translated text and its impact on machine translation. In *In Proceedings of MT-Summit XII*, pages 81–88.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*, 43(4):557–570.

Lembersky, G., Ordan, N., and Wintner, S. (2012a). Adapting translation models to translationese improves smt. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–265. Association for Computational Linguistics.

Lembersky, G., Ordan, N., and Wintner, S. (2012b). Language Models for Machine Translation: Original vs. Translated Texts. *Comput. Linguist.*, 38(4):799–825, December.

Martínez Martínez, J. and Teich, E. (2017). Modeling routine in translation with entropy and surprisal: A comparison of learner and professional translations. In L. Cercel,

[6] http://hdl.handle.net/21.11119/0000-0000-D5EE-4

et al., editors, *Kreativität und Hermeneutik in der Translation*, pages 403–426.

Nisioi, S., Rabinovich, E., Dinu, L. P., and Wintner, S. (2016). A Corpus of Native, Non-native and Translated Texts. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Rabinovich, E., Wintner, S., and Lewinsohn, O. L. (2015). The Haifa Corpus of Translationese. *CoRR*, abs/1509.03611.

Rabinovich, E., Patel, R. N., Mirkin, S., Specia, L., and Wintner, S. (2017). Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084. Association for Computational Linguistics.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees.

Teich, E. and Martínez Martínez, J. (forthcoming). Translation, entropy, cognition. In Arnt Lykke Jakobsen et al., editors, *Routledge Handbook of Translation and Cognition*.

Teich, E. (2003). *Cross-Linguistic Variation in System und Text. A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin.

van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., and Beunders, H. (2017). The debates of the European Parliament as Linked Open Data. *Semantic Web*, 8(2):271–281.

van der Wees, M., Bisazza, A., and Monz, C. (2017). Dynamic Data Selection for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1400–1410.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, pages 590–596.

Volansky, V., Ordan, N., and Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Yang, T.-I., Torget, A. J., and Mihalcea, R. (2011). Topic Modeling on Historical Newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104, Stroudsburg, PA, USA. Association for Computational Linguistics.

Annotation of the Corpus of the Saeima with Multilingual Standards

Roberts Darģis[1], Ilze Auziņa[2], Uldis Bojārs[1,2], Pēteris Paikens[2], Artūrs Znotiņš[2]
Faculty of Computing, University of Latvia[1], Institute of Mathematics and Computer Science, University of Latvia[2]
Raina bulvaris 19, Riga, LV-1459, Latvia[1], Raina bulvaris 29, Riga, LV-1459, Latvia[2]
{ roberts.dargis, ilze.auzina, arturs.znotins, peteris.paikens }@lumii.lv, uldis.bojars@lu.lv

Abstract

This paper describes a release of corpus of the Saeima (parliament of Latvia) as open data resources for multidisciplinary research. The corpus consists of the transcription of Latvian parliamentary debates from 1993 until 2017, containing 38 million tokens from 468 speakers. Current comparative research of parliamentary debate is not sufficiently facilitated by simply providing unannotated corpora and results mostly in monolingual research by local researchers. We propose that augmenting such corpora with extra layers according to commonly used multilingual standards would make it easier to analyze and compare multiple corpora in different languages. In this regard, we believe that the key factors that need to be added are identifiers of entities mentioned in each utterance, and morphosyntactic information for linguistic analysis. For these reasons, the provided corpus is augmented with named entity linking to the Wikidata knowledge base (provided as linked data), automated translations to English, and morphological and syntactic annotations in Universal Dependency format. A part of this corpus is the LinkedSaeima dataset containing structured information about the Saeima debates published as Linked Open Data.

Keywords: syntactic annotation, entity linking, linked data, corpus, RDF, open government data

1. Introduction

The Corpus of the Saeima (Parliament of Latvia) was first published in 2016 (Dargis et al., 2016). At the time, the published corpus was in plain text format with speaker annotations and other metadata. With the increasing availability of corpora in different languages, we realized that unannotated corpora are not enough to facilitate comparative research across multiple language.

To enable researchers to conduct a comparative research across multiple languages without the need to know any of the languages, we propose augmenting corpora with extra layers according to commonly used multilingual standards.

In the paper we describe a new release of the Corpus of the Saeima. The new release contains multiple additional annotation layers:

- Morphosyntactic information for linguistic analysis (lemmas, morphological tags, syntactic).
- Automated translations to English.
- Named entity mentions with links to the Wikidata knowledge base.

The new release of the Corpus of the Saeima is published in multiple commonly used formats:

- A searchable text corpus in NoSketch query software (Rychlý, 2007).
- Syntactically parsed data according to the Universal Dependency standard (Nivre et al., 2016), containing morphological and syntactical annotations.
- LinkedSaeima – Linked Data representation of the corpus with structured information about Saeima proceedings and the entities mentioned in the corpus, represented in the dataset using Wikidata knowledge base identifiers.

To aid searchability for international researchers, the Linked Data format also contains text that was machine-translated to English. Speakers and roles are also linked to Wikidata entities where applicable.

2. The Data of the Corpus of the Saeima

The source data for this corpus was crawled from the Saeima's website[1] where verbatim reports of all the sessions of the Saeima are published in text format. The texts are processed using a semi-automatic pipeline to identify the boundaries of speeches and the speakers. The text is split into utterances, where each utterance contains a speech from only one speaker.

The Corpus of the Saeima includes transcriptions of parliamentary debate from 7 parliamentary terms (5th–12th), covering years 1993–2017. The transcriptions of the Corpus of the Saeima contain 38 million tokens, 497 thousand utterances and 468 speakers.

The available metadata for each utterance includes the date and type of the parliamentary session, speaker's name and affiliation.

3. Annotation Layers

3.1 Morphological and Syntactical Annotations

Morphological and syntactical annotation enables researchers to carry out quantitative analysis of different characteristics of the Corpus of the Saeima, for example:

- The use of gender pronouns in speech, depending on the gender of the speaker.
- The use of active and passive voice.
- The size of the vocabulary of different speakers.

[1] Saeima's website: http://saeima.lv/

The annotations contain lemma, part of speech, morphological features and syntactic dependencies according to the Universal Dependencies standard format.

To aid searching, texts are automatically tokenized, lemmatized and morphologically analyzed and tagged using CMM based tagger (Paikens et al., 2013). Syntactic dependencies are inferred by neural transition-based dependency parser (Znotins, 2016) trained on Latvian Universal Dependencies corpus version 2.1 (Pretkalniņa et al., 2016)[2].

3.2 Translation

The speeches from Latvian are translated to English using a neural machine translation system (Barone et al., 2017). The unreviewed machine-generated translation is provided for quantitative analysis and to aid searchability and understanding for international researchers. However, the text quality of automated translation is lacking, so for qualitative analysis a professional translator should be used.

3.3 Named Entities

For the purposes of this analysis, we developed a named entity linking system based on earlier research for news corpora analysis (Paikens, 2014). In this approach, we used the structured Wikidata information extracts provided by (Ismayilov et al, 2016) as the entity knowledge base. The Wikidata entity alias information is extended with Latvian morphological inflections and automatically generated variants for people and organization names to link the corpus mentions to Wikidata identifiers.

In the Corpus of the Saeima we identified 393 thousand mentions of 3 thousand unique entities. 165 thousand out of 497 thousand utterances contained entity mentions.

4. Available datasets

4.1 Universal Dependencies (CoNLL-U)

Automatic tokenization, morphological and syntactic annotations are published in CoNLL-U data format[3] with simple plain text based encoding, as shown in Figure 1. Columns contain word index, word form, lemma, part-of-speech tag, full morphological tag, morphological features, head of current word, universal dependency relation to head, and spacing information.

The CoNLL-U dataset is published as a language resource alongside this paper[4].

```
# newdoc id = 2016_03_31_355.txt_seq17
# newpar id = 2016_03_31_355.txt_seq17-p1
# sent_id = 2016_03_31_355.txt_seq17-p1s1
# text = Turpinām ar iesniegtajām izmaiņām Saeimas Prezidija apstiprinātajā
sēdes darba kārtībā.
 1  Turpinām     turpināt     _  vmnipt31pan  _  0  root   _  _
 2  ar           ar           _  sppd         _  4  case   _  _
 3  iesniegtajām iesniegt     _  vmnpdfpdpsyp _  4  amod   _  _
 4  izmaiņām     izmaiņa      _  ncfpd4       _  1  iobj   _  _
 5  Saeimas      saeima       _  ncfsg4       _  6  nmod   _  _
 6  Prezidija    prezidijs    _  ncmsg1       _  8  nmod   _  _
 7  apstiprinātajā apstiprināt _ vmnpdfslpsyp _ 8  amod   _  _
 8  sēdes        sēde         _  ncfsg5       _  10 nmod   _  _
 9  darba        darbs        _  ncmsg1       _  10 nmod   _  _
10  kārtībā      kārtība      _  ncfsl4       _  1  obl    _  SpaceAfter=No
11  .            .            _  zs           _  1  punct  _  _
```

Figure 1: A sample from CoNLL-U corpus.

4.2 Bonito corpus browser

The speeches from deputies of the Saeima are published in text corpus query software – NoSketch engine (Rychlý, 2007). The interface provides powerful corpus query system. Query can include words, lemmas, morphological tags and meta data. The result can be further filtered using positive or negative filters. The query result is displayed in concordances. From the result, frequencies and collocations can be computed in the NoSketch as well (Figure 2). The NoSketch query interface is available online with open access[5].

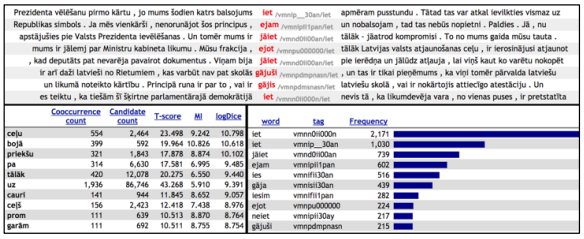


Figure 2: Screenshot of the NoSketch Engine.

4.3 Linked Data

Linked Data allows us to represent structured information about parliamentary debates by describing the properties of the objects from the domain of parliamentary meetings and relations between these objects. According to Linked Data principles, this information is represented using Resource Description Framework (RDF) (Berners-Lee, 2006).

The types of objects in the LinkedSaeima dataset[6] are:

- Meeting – a top-level concept representing one parliament meeting (a plenary) usually consisting of multiple Speeches;
- Speech – an individual speech given at a Meeting by a particular Speaker in some Role;
- Speaker – a person giving a speech;
- Role – a role (e.g. Prime Minister) which the person represented when giving a Speech.

[2] Universal Dependencies corpus version 2.1: https://github.com/UniversalDependencies/UD_Latvian
[3] CoNLL-U data format: http://universaldependencies.org/format.html
[4] The Corpus of the Saeima in CoNLL-U data format: http://saeima.korpuss.lv/datasets/ud/
[5] NoSketch server interface for the Corpus of the Saeima: dati.saeima.korpuss.lv/nosketch/
[6] LinkedSaeima dataset index page: http://dati.saeima.korpuss.lv/

For data modelling we reuse the work of the LinkedEP project (European Parliament debates as Linked Data) and their Linkedpolitics vocabulary, referenced in RDF data using prefixes *lpv* and *lpv_eu* (van Aggelen et al., 2017).

For example, a Speech is represented by *lpv_eu:Speech*, its properties include date (*dc:date*), sequence number and spoken text (*lpv:spokenText*), and it is related to the Meeting it is a part of (*dct:isPartOf*), to the Speaker (*lpv:speaker*) and its Role (*lpv:spokenAs*), and to the named entities mentioned in the text (*schema:mentions*).

The dataset is published as Linked Data and information about its objects is accessible by looking up relevant Linked Data URIs (Berners-Lee, 2006). All dataset objects have HTTP URI identifiers. The implementation uses LodLive[7] linked data browser to serve the data in HTML, RDF and multiple other formats (Figure 3).

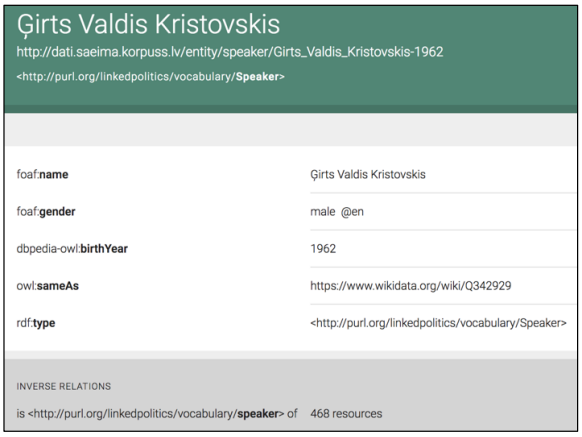


Figure 3: Screenshot of a LinkedSaeima entity in LodView.

A triple pattern fragments server and user interface[8] is published to make LinkedSaeima dataset queryable. Triple fragments server is a lightweight way for querying RDF datasets (Verborgh et al., 2016). The triple pattern fragments server can be used to query RDF dataset for RDF triplets according to any combination of subject, predicate and object (Figure 4).

The dataset is also released alongside this paper as a single RDF file that researchers can use to run more complex analysis[9].

Main innovation of this dataset, relative to the LinkedEP project, is the addition of named entity information, represented in RDF using *schema:mentions* property pointing to relevant Wikidata URI indentifiers. Another difference is that we "materialize" speaker Roles extracted from the corpus by giving them URI identifiers that can be used for querying the dataset (e.g. for speeches by presidents of the European Commission[10]).

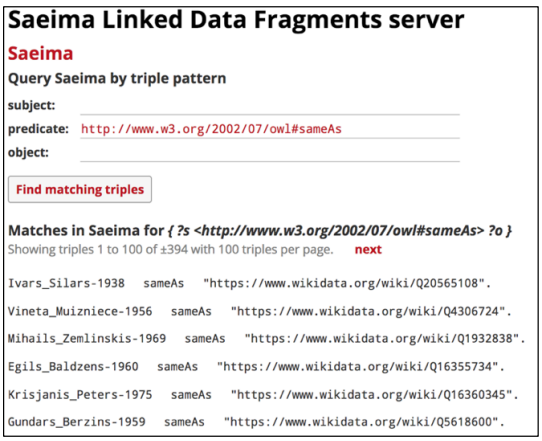


Figure 4: Screenshot of the LinkedSaeima triple pattern fragments server.

Directions for further development of the LinkedSaeima dataset include adding richer information about entity references and introducing new types of information related to the Saeima proceedings (e.g. extracting voting data). In this version, entities are represented by a property linking Speech objects to Wikidata entity identifiers. An alternative approach is to represent entity mentions as separate objects (e.g. by adapting W3C Web Annotation standard for entity references (Bojars et al., 2017). A benefit of this approach is that it can represent additional information such as the text position of the entity reference. Its downside is a larger and more complex dataset.

5. Expected use cases

Initially the corpus of the Saeima was created to facilitate the process of research for political and social scientists. The scientists have used this corpus for discourse analysis (Kruk 2007, Auzina 2007) and to oversee political and social processes in Latvia (Chojnicka 2013). It is also used by linguists as a corpus for language research (Treimane 2014).

The new annotation levels (especially named entities and translation) and its Linked Data representation will make it possible to compare Latvian parliamentary data with other national parliamentary data and to provide users with new ways for exploring this information. The described datasets have been used for different purposes:

- Annotation representation across languages for Named Entity Recognition (Ehrmann et al. 2011);
- Training and testing information extraction software;
- To produce bilingual or even multilingual cross-language resources such as dictionaries, or applications, for example, cross-lingual word sense disambiguation, cross-lingual information retrieval.

[7] LodLive linked data browser: https://github.com/dvcama/LodLive/
[8] LinkedSaeima triple pattern fragments server and user interface: http://dati.saeima.korpuss.lv/ldf/saeima
[9] LinkedSaeima RDF dump: http://saeima.korpuss.lv/datasets/rdf/
[10] URI for the President of the European Commission: http://dati.saeima.korpuss.lv/entity/role/89

6. Conclusions and further work

In conclusion, we have described a new dataset of parliamentary debate with extended annotations that should make it more useful for research and analysis.

We'd like to call upon this research community to extend their resources while keeping in mind multilingual applications. While currently parliamentary discourse analysis is fragmented, we believe that using standards that are common in NLP field we can pave the road for easy multilingual comparative analysis of many parliamentary corpora. Each country has similar data, but the language diversity and differences in technical format makes it difficult for researchers to summarize many corpora. We suggest others to investigate the possibility of providing their data in commonly used international formats, which in our opinion are Universal Dependencies for morphological and syntactic analysis, and RDF and Linked Data for entity information, with the hope of enabling new areas of research comparing parliamentary discourse of many countries.

Expected future work includes continuous processing of new debate data, improvements to entity linking and disambiguation, and extending the LinkedSaeima dataset with additional types of structured information e.g. voting data.

7. Acknowledgements

This work has been partially supported by the European Regional Development Fund (ERDF) project No. 201X/0020/2DP/2.1.1.1.0/14/ APIA/VIAA/000 at the Faculty of Computing, University of Latvia.

This work has been partially supported by the University of Latvia project AAP2016/B032 "Innovative information technologies".

The data collection process has been supported by Latvian National research program EKOSOC-LV No. 5.2.5.

The tools developed and used in this project have received financial support from the European Regional Development Fund under the grant agreement No. 1.1.1.1/16/A/219.

8. Bibliographical References

Barone, A. V. M., Helcl, J., Sennrich, R., Haddow, B., & Birch, A. (2017). Deep architectures for neural machine translation. arXiv preprint arXiv:1707.07631.

Berners-Lee, T. (2006). Linked Data – Design Issues. http://www.w3.org/DesignIssues/LinkedData.html

Bojars, U., Rasmane, A., Zogla, A. The Requirements for Semantic Annotation of Cultural Heritage Content. Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II) co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 22, 2017. CEUR Workshop Proceedings, vol. 2014.

Dargis, R., Rabante-Busa, G., Auzina, I., & Kruks, S. (2016, October). ParliSearch-A System for Large Text Corpus Discourse Analysis. In Baltic HLT (pp. 115-121).

Ismayilov, A., Kontokostas, D., Auer, S., Lehmann, J., & Hellmann, S. (2016). Wikidata through the Eyes of DBpedia. Semantic Web, (Preprint), 1-11.

Nivre, J., de Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., ... & Tsarfaty, R. (2016, May). Universal Dependencies v1: A Multilingual Treebank Collection. In LREC.

Paikens, P. (2014, September). Latvian Newswire Information Extraction System and Entity Knowledge Base. In Baltic HLT (pp. 119-125).

Paikens, P., Rituma, L., & Pretkalnina, L. (2013, May). Morphological analysis with limited resources: Latvian example. In Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16 (No. 085, pp. 267-277). Linköping University Electronic Press.

Pretkalniņa, L., Rituma, L., & Saulīte, B. (2016, October). Universal Dependency Treebank for Latvian: a Pilot. In Human Language Technologies-The Baltic Perspective: Proceedings of the Seventh International Conference Baltic HLT 2016 (Vol. 289). IOS Press.

Rychlý, P. (2007, December). Manatee/bonito-a modular corpus manager. In 1st Workshop on Recent Advances in Slavonic Natural Language Processing (pp. 65-70).

Skadina, I., & Rozis, R. (2016, October). Word Embeddings for Latvian Natural Language Processing Tools. In Human Language Technologies-The Baltic Perspective: Proceedings of the Seventh International Conference Baltic HLT 2016 (Vol. 289, p. 167). IOS Press.

van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., & Beunders, H. (2017). The debates of the European parliament as linked open data. Semantic Web, 8(2), 271-281.

Verborgh, R., Vander Sande, M., Hartig, O., Van Herwegen, J., De Vocht, L., De Meester, B., ... & Colpaert, P. (2016). Triple Pattern Fragments: A low-cost knowledge graph interface for the Web. Web Semantics: Science, Services and Agents on the World Wide Web, 37, 184-206.

Znotins A. (2016). Word embeddings for Latvian natural language processing tools, Proceedings of Human Language Technologies -- The Baltic Perspective, IOS Press.

A Sentiment-labelled Corpus of Hansard Parliamentary Debate Speeches

Gavin Abercrombie and **Riza Batista-Navarro**
School of Computer Science, University of Manchester, Kilburn Building, Manchester M13 9PL
gavin.abercrombie@postgrad.manchester.ac.uk, riza.batista@manchester.ac.uk

Abstract

Hansard transcripts provide access to Members of Parliament's opinions on many important issues, but are difficult for people to process. Existing corpora for sentiment analysis in Hansard debates rely on speakers' votes as sentiment labels, but these votes are known to be constrained by speakers' party affiliations. We develop an annotation scheme and create a novel corpus designed for use in the evaluation of sentiment analysis systems using automatically and manually applied speech labels. Observing the effects on speech sentiment of differing sentiment polarities in debate motions (proposals), we also apply sentiment labels to these motions. We find that humans are able to reach high agreement in identifying sentiment polarity in these debates, and that manually applied and automatically retrieved class labels differ somewhat, suggesting that speech content does not always reflect the voting behaviour of Members of Parliament.

Keywords: Hansard, UK Parliament, sentiment analysis

1. Introduction

Hansard transcripts of debates from the United Kingdom Parliament provide access to the opinions and attitudes of Members of Parliament (MPs) and their parties towards many important topics facing society. However, the large quantity of recorded material combined with the esoteric speaking style and opaque procedural language used in Parliament makes manual interpretation of information from these data a daunting task for the non-expert citizen.

Sentiment analysis is the task of automatically identifying the polarity (*positive* or *negative*) of the position a person takes towards an entity, such as an organisation, a policy, a movement, a situation, or a product. Automatic detection of MPs' sentiment towards the topics that they discuss in debates has applications in tasks such as information retrieval and question answering, and could allow the public to more easily assess and aggregate the contributions that their elected representatives make in Parliament.

Existing datasets for sentiment analysis of Hansard rely on speakers' votes as sentiment polarity labels (Onyimadu et al., 2013; Salah, 2014). However, it is widely recognised that MPs are to a large extent constrained in their voting behaviour, and often under pressure to vote along party lines irrespective of their personal opinion (Searing, 1994; Norton, 1997). For instance, in Example 1, the speaker appears to be against the motion, yet votes in support of it:

(1) **Motion:** That there shall be an early parliamentary general election.
Speech: Does my right hon. Friend agree that the Prime Minister, in calling this election, has essentially said that she does not have confidence in her own Government to deliver a Brexit deal for Britain? One way in which she could secure my vote and the votes of my hon. Friends is to table a motion of no confidence in her Government, which I would happily vote for.
Vote: 'Aye' (*positive*).

On top of this, MPs may change their mind between speech and vote, and are even known to vote erroneously on occasion.[1] These vote labels may not therefore be accurate reflections of the opinions displayed in the content of MPs' debate speeches, and an alternative form of class labelling may be required for effective sentiment classification using supervised machine learning methods.

Our contribution In this paper, we present Hansard Debates with Sentiment Tags (HanDeSeT), a novel corpus of manually labelled Parliamentary debates for use in the evaluation of automatic Parliamentary speech-level sentiment analysis systems. These consist of proposed *motions* and the associated *speeches* of Members of the House.

2. Related Work

Sentiment analysis has long been one of the most active areas of research in natural language processing (NLP), where attention has been focussed to a large extent on the domains of online reviews (e.g., Pang et al. (2002)) and social media (e.g., Pak and Paroubek (2010)).

For similar tasks in the legislative debate domain, Thomas et al. (2006) use crowdsourced annotations to build a dataset of speech segments from US congressional debates, for which they attempt to automatically determine whether the speakers support or oppose the proposed legislation. Meanwhile, Grijzenhout et al. (2010) create a corpus of Dutch parliamentary debates annotated for *positive* or *negative* 'semantic orientation' at the paragraph level.

In the field of political science, Schwarz et al. (2017) analyse debates from the Swiss parliament, comparing speech content with votes, and find that legislators speak with more freedom than they are able to exercise in their voting behaviour, further motivating our approach.

In the most similar work to ours, Salah (2014) collects a dataset of parliamentary debates comprised of 2,068 speeches in order to perform sentiment analysis on UK Hansard transcripts. Under the assumption that MPs' votes reflect the sentiment of their speeches, these votes are used to label speeches as having *positive* or *negative* polarity.

3. Hansard UK Parliamentary Debates

Hansard transcripts are largely-verbatim records of the speeches made in both chambers of the UK Parliament, in which repetitions and disfluencies are omitted, while supplementary information such as speaker names are added. As the superior legislative body, the House of Commons is

[1] As described by Paul Flynn, MP (Flynn, 2012).

generally of greater interest to the public and media, and is therefore the focus of this study.

3.1. Composition of House of Commons Debates

House of Commons debates consist of these elements:

Motions Debates are initiated with a *motion*—a proposal made by an MP. These motions can be either 'substantive'—requiring the House to support or oppose a policy, piece of legislation, or state of affairs—or 'general'—asking MPs to merely acknowledge that a particular topic has been considered by the House, regardless of their opinions towards it.[2]

Speeches When invited by the *Speaker* (the presiding officer of the chamber), other MPs may respond to the motion, one or more times. Each speaking turn may be comprised of a short statement or question, or a longer passage, which is divided into paragraphs in the transcript.

Divisions At any time (typically at the end of the debate) the Speaker may call a *division*, whereby MPs vote by physically moving to either the 'Aye' or 'No' lobby of the chamber. There may be more than one vote on each motion.[3]

3.2. Semantic Structure of House of Commons Debates

During data collection and initial experiments, we observed certain characteristics of the stucture of these debates which are likely to have a bearing on the sentiment detection task:

Motion sentiment Sentiment polarity is present in both debate speeches *and* motions. In proposing a motion, an MP expresses sentiment towards the policy, piece of legislation, or state of affairs in question.

Double negative effect The language used to express *positive* or *negative* speech sentiment is radically altered depending on the sentiment polarity of the motion. A sort of double negative effect is created, whereby speakers may use typically negative language to demonstrate positive sentiment and vice versa.

For example, if a motion praises the actions of the Government, speeches in support of the motion will likely contain positive language, while those opposing it will be characterised by negative language. If, however, a motion condemns Government policy, supporting speeches are also likely to contain negative language, and opposing ones positive language, as in Example 2:

(2) **Motion:** That an humble Address be presented to Her Majesty, praying that the Local Authorities (England) Regulations 2000 be annulled.
Speech: ... there are deep reservations in the county about all the proposals. I am particularly alarmed about the impact of key decisions. An enormous electoral ward such as Bowbrook or Inkberrow, where huge decisions could be taken affecting communities, will not be subject to openness under the proposals. Why are huge electoral divisions excluded in that monstrous way?

Based on these observations, Abercrombie and Batista-Navarro (2018b) propose a two-stage sentiment analysis model, in which opinions expressed in both motions and speeches are analysed. For this reason, we include manually annotated labels for motions as well as for speeches. Noting that speeches are often made in either attack or defence of the Government's actions, we also include a motion sentiment label derived from the party affiliation of the MP who proposes the motion: *positive* if they are a member of the governing party or coalition, *negative* if not.

4. Corpus Construction

We create and make available a corpus of labelled debates for speech-level sentiment analysis on Hansard debate transcripts from the House of Commons of the UK Parliament.

4.1. Data Collection

Debate transcripts from 1935 onwards are available in XML format on the parliamentary monitoring website TheyWorkForYou.com.[4] In order to obtain a sufficient quantity of speeches for which there are associated division votes, we downloaded the records of all debates in the House of Commons from May 1997 to July 2017.

Each file contains transcipts of a number of debates. We selected all debates under 'major-heading' elements in the XML files—debates which often culminate in *divisions*, or votes. We retained only debates that contain a motion and precisely one *division*, under the assumption that each member's vote represents their sentiment towards the motion under debate. We included only debates with substantive (rather than general) motions, as, by their nature, these demand polarised stances to be taken by MPs.

4.2. Data Processing

Parliamentery speeches incorporate much set, formulaic discourse related to the operational procedures of the chamber, which we automatically removed as it does not concern the motion or the speakers' opinions towards it. These include speech segments such as those used to thank the *Speaker*, or to cede the floor, as well as descriptions of activity in the chamber inserted into the transcripts by the reporters, for example showing that a member rose from their seat or indicated assent by nodding.[5] Additionally, we removed all utterances produced after a division is made, as these are generally procedural matters related to the running of Parliament and/or off-topic.

As in Salah (2014), we consider a member's *speech* to be the concatanated content of *all* their *utterances* (individual speech segments or paragraphs). For comparison of manual and vote labelling methods, we retained all speeches made by MPs who appear in the division of the given debate along with a record of their vote. We omit speeches made by the member of the assembly that proposes the motion, as, by definition, they speak in support of the proposal.

[2] See www.parliament.uk/about/how/business/debates.

[3] For example, several clauses or ammendments to a Bill or Paper discussed in a motion may be voted on individually.

[4] https://www.theyworkforyou.com/pwdata/scrapedxml/debates/

[5] We automatically remove the following procedural language: names of MPs mentioned in speeches (inserted by the reporters), utterances solely concerned with 'giving way' or making interventions, utterances concerning *points of order*.

Also following Salah (2014), we removed speeches totalling fewer than 50 words. In order to facilitate manual labelling, we restrict the quantity of text to be read by human annotators by including only those speeches comprised of five utterances or fewer. Finally, quotations within speeches were removed (and replaced with the word 'QUOTATION'), as these can reflect opposing or different points of view to those of the speaker, and represent confounding features for automatic sentiment classification.

5. Annotation

Annotation guidelines were developed in a two-round cycle using a randomly selected subsection (20%) of the corpus and three annotators—all L1 English speakers, university graduates, UK citizens, and self-reporting as being familiar with British politics and the UK parliament. The principal annotator (*annotator 1*, an author of this paper) then produced the gold standard labels for the complete corpus following the final version of these guidelines.

5.1. Development of Annotation Guidelines

Manual sentiment labelling was carried out on the corpus subsection (250 speech units) in two rounds of the following cycle:

1. Annotation guidelines produced/updated.
2. Two annotators labelled the corpus subsection.
3. Inter-annotator agreement calculated and disagreement analysis performed.

Finally, the principal annotator labelled the full corpus.

5.2. Annotation Guidelines

Following the final annotation guidelines,[6] the job of the annotator can be summarised as follows:

For each *unit* (motion plus speech) in the dataset, the annotator reads the motion carefully, makes a decision on its sentiment polarity towards the subject of the debate, and assigns it the corresponding label: '`1`' for *positive*, '`0`' for *negative*. The annotator then reads the speaker's utterances, considering their overall sentiment polarity, and assigns a label for the sentiment polarity of the speech in question towards the motion (again '`1`' or '`0`').

6. Analysis of the Annotations

To assess the validity of the manually applied labels, we calculated Cohen's kappa (κ) after each round of annotations. We then performed a systematic manual analysis of cases on which the annotators disagreed, identified measures that could be taken to improve agreement, and updated the annotation guidelines accordingly.

Annotation guidelines used	Motion κ	Speech κ
Version 1 (annotators 1 & 2)	0.56	0.57
Version 2 (annotators 1 & 3)	0.91	0.85

Table 1: Inter-annotator agreement (Cohen's kappa) for motion and speech sentiment polarity labels following the first and second versions of the annotation guidelines.

Identified causes of disagreement are presented in Table 2.

Cause of disagreement	Cases (%)
Motions	
Motion calls for action (*positive*), but opposes the target (*negative*)	85.0
Annotator error: same motion labelled differently in other examples	5.0
Annotator error: possible missed negation in motion	5.0
Possible misinterpretation: motion sentiment is against previous, not current Govt.	5.0
Speeches	
Off-topic speech content	16.7
Contextual information required	13.0
Procedural (i.e., long, detailed) motion	13.0
Motion IA disagreement	9.3

Table 2: Causes of annotator disagreement for round 1.

Round one Inter-annotator agreement on the first round of annotation was found to be 'moderate'[7] for both motions and speeches. This was poorer than expected, as intuitively the task seemed relatively straightforward, particularly for labelling of motions, which by definition in these substantive debates are proposed in favour of, or against something.

Round two To address the issues raised by this analysis, we updated the annotation guidelines, clarifying the instructions and adding further example cases. In particular, we defined a protocol for handling motions which call on the Government for action, but which can be seen as attacking its position. These are common in the corpus and had accounted for 85% of annotator disagreement on motion sentiment. We also provided the annotators with more contextual information, by including the MPs' party affiliation. This process resulted in 'very good' agreement on both motions and speeches for the second round of annotations, a considerable improvement on the first round. Given sufficiently clear instructions, humans appear to be capable of high levels of agreement in recognising sentiment polarity in parliamentary debates. As anticipated, sentiment identification in motions seems to be particularly straightforward. We manually analysed cases of disagreement in the second round of annotation, and found that the only two cases of disagreement over motion sentiment were probably caused by error or misinterpretation by one of the annotators. The same can be said for many cases of speech sentiment disagreement, although some were identified as being either off-topic or highly ambiguous, as in Example 3:

(3) **Motion:** That this House believes that the UK needs to stay in the EU because it offers the best framework for trade, manufacturing, employment rights and cooperation to meet the challenges the UK faces in the world in the twenty-first century; and notes that tens of billions of pounds worth of investment and millions of jobs are linked to the UK's membership of the EU, the biggest market in the world.

[6] Available in Abercrombie and Batista-Navarro (2018a).

[7] As described by Landis and Koch (1977).

> **Speech:** My hon. Friend is making a powerful speech and makes an important point about patriotism. Does he agree that key to Britain's national security is our economic security, and at a time when we are still borrowing as a nation more than the entire defence budget we need every single penny of public revenue to ensure our economy is strong, our finances are strong and our country is strong?

Here, without access to information about the speaker's views on a range of issues (e.g., the UK's membership of the EU), the speaker's sentiment towards the motion is likely to seem ambiguous. The presence of such speeches in House of Commons debates makes it unlikely that 100% agreement could be achieved on this task without further contextual clues.

7. Corpus Description

The corpus is available at `https://data.mendeley.com/datasets/xsvp45cbt4/`. It consists of 1251 motion-speech units taken from 129 separate debates, with each unit comprising a parliamentary speech of up to five utterances and an associated motion. Debates comprise between one and 30 speeches, and speeches range in length from 31 to 1049 words, with a mean of 167.8 words. The debates cover a two decade period from 1997 to 2017 and a wide range of topics from domestic and foreign affairs to procedural matters concerning the running of the House.

Each motion has both a manually applied and a *Government/opposition* sentiment label and each speech also has two sentiment polarity labels, produced with different labelling methods for comparison: (1) A speaker-vote label extracted from the division associated with the corresponding debate; and: (2) A manually assigned label.

In addition, the following metadata is included with each unit: *debate id*, *speaker party affiliation*, *motion party affiliation*, *speaker name*, and *speaker rebellion rate*.[8]

Manually applied motion labels are approximately evenly balanced; the other labels are slightly skewed towards the positive class (See Table 3).

Target	Label type	Positive	Negative
Motion	Govt./opp.	71 (55.0%)	58 (45.0%)
	Manual	67 (51.9%)	62 (48.1%)
Speech	Vote	713 (57.0%)	537 (43.0%)
	Manual	702 (56.5%)	544 (43.5%)

Table 3: Occurrences of sentiment labels in the corpus.

Concurrence between the vote labels and manually annotated labels is 92.8%. This indicates that, while the majority of speeches reflect the voting behaviour of the speaker, a significant number do not, and that division votes may not therefore be reliable sentiment polarity labels.

In general, MPs both speak and vote along party lines. Of the seven parties that have had more than one sitting MP at a time, the SNP, and the three smaller parties (DUP, UUP, Green) always vote as a block and are assigned the same manual sentiment label. The major UK-wide parties exhibit rather more rebellion, for speech sentiment (Lab: 8.2%, Con: 6.6%, LD: 2.8%) and vote (Lab: 4.2%, Con: 1.1%, LD: 0.0%).

Examining these 'rebel' speeches, we find that they tend to occur in debates that concern (a) topics of local interest, such as local government finance, in which MPs' loyalties may be divided between party and constituency, (b) matters of conscience, such as stem cell research, or (c) issues that are known to divide parties, such as membership of the EU. In several speeches, a speaker states explicitly that they will vote one way, only to vote for the opposing side, confirming the unreliabilty of votes as sentiment labels.

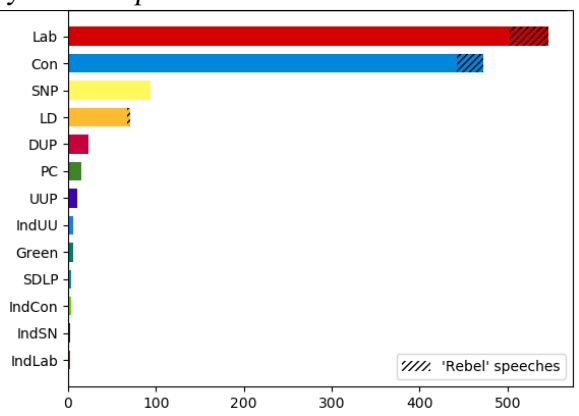


Figure 1: Number of speeches in the corpus by party, and number of 'rebel' speeches in which the manual sentiment label opposes the majority of the speaker's own party.

8. Conclusion

This paper presents a corpus of parliamentary debates from the UK House of Commons, manually annotated and vote-labelled for sentiment at the speech level and with additional sentiment labels applied to debate motions. In order to create this corpus, we developed a set of annotation guidelines, and demonstrated that, using these instructions, agreement on this sentiment identification task can be relatively straightforward for humans, although some ambiguous cases remain challenging. We obtained satisfactory inter-annotator agreement scores, which validate the corpus, and created gold standard labels for use in the evaluation of automatic sentiment analysis systems.

While the majority of manually annotated and automatically applied labels in the corpus agree, a number differ. This indicates that MPs may be freer to express personal opinion in their speeches than in their voting behaviour, and has implications for automatic sentiment analysis, where division votes are perhaps not the best labels for this task.

9. Acknowledgements

We would like to thank Loren Hosein, Kieran Lilwall, and Anthony Chambers for their contributions, and the anonymous reviewers for their invaluable comments.

[8] Rate of MPs' votes against the majority of their party in the current parliament, extracted from http://www.publicwhip.org.uk/.

10. Bibliographical References

Abercrombie, G. and Batista-Navarro, R. (2018a). *HanDeSeT: Hansard Debates with Sentiment Tags*. Mendeley Data.

Abercrombie, G. and Batista-Navarro, R. (2018b). 'Aye' or 'no'? Speech-level sentiment analysis on Hansard UK Parliamentary debate transcripts. In *Language Resources and Evaluation Conference*. LREC.

Flynn, P. (2012). *How to be an MP*. Biteback.

Grijzenhout, S., Jijkoun, V., Marx, M., et al. (2010). Opinion mining in Dutch Hansards. In *Proceedings of the Workshop From Text to Political Positions, Free University of Amsterdam*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Norton, P. (1997). Roles and behaviour of British MPs. *The Journal of Legislative Studies*, 3(1):17–31.

Onyimadu, O., Nakata, K., Wilson, T., Macken, D., and Liu, K. (2013). Towards sentiment analysis on Parliamentary debates in Hansard. In *Joint International Semantic Technology Conference*, pages 48–50. Springer.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Salah, Z. (2014). *Machine learning and sentiment analysis approaches for the analysis of Parliamentary debates.* Ph.D. thesis, University of Liverpool, UK.

Schwarz, D., Traber, D., and Benoit, K. (2017). Estimating intra-party preferences: comparing speeches to votes. *Political Science Research and Methods*, 5(2):379–396.

Searing, D. (1994). *Westminster's world: understanding political roles*. Harvard University Press.

Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 327–335. Association for Computational Linguistics.

Automatically Labeled Data Generation for Classification of Reputation Defence Strategies

Nona Naderi and Graeme Hirst
Department of Computer Science
University of Toronto
Toronto, ON, M5S 3G4, Canada
{nona,gh}@cs.toronto.edu

Abstract

Reputation defence is a form of persuasive tactic that is used in various social settings, especially in political situations. Detection of reputation defence strategies is a novel task that could help in argument reasoning. Here, we propose an approach to automatically label training data for reputation defence strategies. We experimented with over 14,000 pairs of questions and answers from the Canadian Parliament, and automatically created a corpus of questions and answers annotated with four reputation strategies. We further assessed the quality of the automatically labeled data.

1. Introduction

Maintaining good reputation is important in almost all social settings. Losing one's reputation can affect competitiveness, trust, position, and relations. Individuals, businesses, and institutions try to manage reputation threat or the danger of losing their reputation by using various persuasive defence strategies (Benoit, 1995). A recent prevailing example of reputation threat and defence is various sexual assault allegations and the use of strategies, such as *denial,* i.e., denying the situation, and *mortification,* i.e., the admission of guilt and apologizing and asking for forgiveness, in response to these allegations.

Maintaining good reputation is particularly important to politicians, as often the most acceptable political images are voted for and chosen by the electorate. Politicians who choose policies that impact citizens are more concerned with their reputations because they are held responsible for their actions by both citizens and other political parties.

One example of a reputation defence strategy is the expression of *mortification* in a statement that was issued by the U.S. Secretary of Health regarding the expenses of his travel on private planes:

"I regret the concerns this has raised regarding the use of taxpayer dollars. All of my political career I've fought for the taxpayers. It is clear to me that in this case, I was not sensitive enough to my concern for the taxpayer."[1]

While reputation defence strategies and their effectiveness have been extensively studied (Coombs and Holladay, 2008; Sheldon and Sallot, 2008; Burns and Bruner, 2000; Sheldon and Sallot, 2008; Lyon and Cameron, 2004), most of these studies are qualitative in nature. One exception is that of Naderi and Hirst (2017), who proposed a computational approach to identify reputation defence strategies from parliamentary debates. Here, we propose two semi-supervised approaches for identifying persuasive reputation defence strategies. One approach uses the observed word pairs from both reputation threat and reputation defence, and the other uses pattern-based representations of reputation defence.

We evaluated a subset of the automatically labeled data against crowd-sourced annotations. We further assessed the impact of the extended dataset in a multi-class classification task. We found that the approach based on the observed word pairs yields higher-quality labels for the reputation defence strategies.

2. Related Work

Ethos i.e., one's credibility has been considered as one of the important means of persuasion in Aristotle's rhetoric (Aristotle, 2007). In danger of losing credibility, one may prepare apologia that is a self-defence speech in response to the criticism or attack. According to Downey (1993), apologia has taken various functions and styles over time, for example, early contemporary apologia resembled classical apologia and used causal reasoning and detailed evidence; however, after 1960, apologia has been altering into "misleading narratives and dishonest apologies", replete with discrepancies. Similar to Downey's study, most previous work on persuasive reputation defence strategies focused on a few case studies (Brinson and Benoit, 1999; Benoit and Henson, 2009; Zhang and Benoit, 2009; Harlow et al., 2011) with the exception of one study. Naderi and Hirst (2017) created a corpus of parliamentary question and answers annotated with four reputation strategies and proposed a feature-based approach to detect these strategies (see Table 1). Parliamentary question periods provide a rich dataset to study various crises and the face-saving strategies that are used to manage these crises. Parliamentary question periods have previously been studied for analysing rhetorical aspects of questions (Zhang et al., 2017), interruption behaviour (Whyte, 2014), determining ideologies using party-membership (Hirst et al., 2014), and measuring emotions (Rheault et al., 2016).

While the task of automatic detection of reputation defence strategies is closely related to argumentation mining tasks (Stab and Gurevych, 2014; Nguyen and Litman, 2016; Biran and Rambow, 2011; Wang and Cardie, 2014; Peldszus, 2014), it differs in that it focuses on relations between arguments of reputation threat (questions)

[1] Health secretary Tom Price apologizes for taking private flights for work, *The Guardian,* 2017-09-28

Reputation defence strategies
Denial: 1. The government denies that the situation in question occurred. 2. The government denies causing the situation in question.
Excuse (evading responsibility): 1. The situation in question occurred in response to some other situations. 2. The situation in question occurred because of lack of information or control over important factors. 3. Some accidents caused the situation. 4. The motives or intentions of the government were good.
Justification (reducing offensiveness): 1. The government tries to increase positive feeling towards it (for example by mentioning positive actions the government performed in the past). 2. The government tries to convince the audience that the situation is not as bad they say. 3. The government tries to distinguish the situation in question from similar but less desirable situations. 4. The government tries to place the situation in a different or broader context. 5. The government attacks the opposition or questions their credibility. 6. The government offers compensation for the situation.
Concession (corrective actions): 1. The government promises to restore the situation to what it was before. 2. The government promises to make changes (for example to prevent the recurrence of the situation).
None of these strategies

Table 1: Conditions for each reputation defence strategy. This table is taken from the study by Naderi and Hirst (2017).

and reputation defence (answers). Previous studies on argumentation have shown that manually annotating argument-related information is difficult and results in moderate agreement (Habernal et al., 2014; Wachsmuth et al., 2017; Naderi and Hirst, 2017). Here, we aim to automatically create a large corpus of reputation defence strategies. We propose two approaches and examine the quality of the extracted data using these approaches.

3. Method

For our analysis, we used a dataset described by Naderi and Hirst (2017). This dataset consists of 493 pairs of questions and answers from Oral Question period from Canadian parliamentary proceedings, manually annotated with four reputation defence strategies (170 pairs of questions and answers are annotated as denial, 36 pairs as excuse, 173 pairs as justification, 95 pairs as concession, and 19 as none of these strategies). Here, we removed 19 pairs that were annotated as being *none* of these strategies, and focused on the remaining pairs. We refer to this corpus as the reputation defence strategy dataset throughout the paper. Given these manually labeled examples, we extracted a set of features to assign scores to unlabeled pairs of questions and answers and automatically expanded the training set.

3.1 Preprocessing of data

Here, we used the Lipad[2] (Linked PArliamentary Data) dataset (Beelen et al., 2017). This dataset consists of Canadian Hansards since 1901. We extracted 14,134 pairs of questions and answers from Oral Question period (1994–2014) as our unlabeled data. Since the questions asked by the government backbenchers are generally friendly and intended for clarification, we only focused on the questions asked by the opposition members and their respective answers by the government ministers. An example of a reputation defence is presented in Table 2. Furthermore, we extracted only the first question and answer pairs of each topic of discussion, because the remaining pairs require the context. We made sure that the pairs of questions and an-

Q. Mr. Speaker, we now know that the Prime Minister announced a $600,000 grant in his riding months before the project had been approved, and coincidentally just weeks before the federal election. Since only the Prime Minister knows when an election will be called, it is clearly and simply a case of announcing pre-election goodies. The Prime Minister would have us believe the grant was awarded after careful review, but program officer Lionel Bergeron thought differently when he said in a memo "This project has been announced by the Prime Minister. Its approval is urgent". How could the Prime Minister deny that he was just trying to influence voters in his riding by getting this grant before it went through the proper circle?
A. Mr. Speaker, this project had been discussed for years in Shawinigan. It is the kind of project that is badly needed in a district where unemployment is very high in the Saint-Maurice riding. Everyone had been talking about it. Everyone supported the project, including the hon. member for Saint-Maurice who has done his job as the local member for Saint-Maurice. We are very pleased that the project has worked and has indeed created the jobs that it was supposed to bring to the region.

Table 2: An example of reputation defence strategy; 1999-05-25, Chuck Strahl (Q) and Pierre S. Pettigrew (A).

[2] https://WWW.lipad.ca

swers from the reputation defence dataset were not included in our unlabeled dataset. We extracted two sets of features to assign scores to unlabeled question and answer pairs: (1) observed word pairs, (2) surface patterns. We will discuss these features in the following sections.

3.2 Pairs of words

Word pairs from a pair of arguments have been shown to be informative features in identifying implicit discourse relations between the two arguments (Marcu and Echihabi, 2002; Pitler et al., 2009; Biran and McKeown, 2013).
Additionally, Naderi and Hirst (2017) have shown that discourse relations between the question and answer sentences can help in capturing the relations between reputation threat and defence instances, and they can be informative features for the detection of reputation defence strategies. Therefore, we considered all the possible word pairs extracted from the cross-product of the question and answer. To represent the relevance of each word pair to each reputation defence strategy, we computed a correlation score using our seed examples. A score is assigned to each question and answer based on simple occurrences:

$$\left(\frac{\text{Count unique word pairs of Label}_i}{\text{Count total unique word pairs}}\right)$$

The raw score was then normalized by dividing by the sum of raw scores of all four strategies.

3.3 Pattern extraction

For extracting the surface patterns, we took an approach similar to that of Tsur et al., (2010). Using the extracted unlabeled question and answer pairs, we divided the words into frequent and infrequent words (IFW) according to their relative frequency in the unlabeled corpus and a specified threshold. This threshold was set to 1000 per million. The length of patterns was set to be 5 to 7 words with only 3 to 5 slots for infrequent words. Multiple patterns were extracted from each reputation defence answer. We then computed a score for each question and answer pair according to the exact matches of the patterns of each reputation defence strategy. For example, from the *denial* answer *Mr. Speaker, at no time have we interfered with the operations of Air Canada, and I stand by my answer of yesterday*, the following example patterns were extracted:

- *at no time have we* IFW *with*
- *no time have we* IFW *with the*
- *have we* IFW *with the*
- *i* IFW *by my* IFW *of yesterday*

Each question and answer pair was first assigned a raw score for each strategy, and then the score was normalized by the sum of all strategy scores (similar to the approach in Section 3.2):

$$\frac{\sum_k \text{Length}(\text{pattern}_k) \times \text{Count}(\text{pattern}_k)}{\sum_i \text{Score of Label}_i}$$

Q. Mr. Speaker, my question is for the Minister of Human Resources Development. It concerns the government's plans for the end to the TAGS program. How could the minister expect Canadians to take him seriously when he says that the government is working on plans to help out the affected communities after TAGS is finished and we know he is telling the RCMP and his own officials they should get ready for the fact that they will be doing nothing? The minister now has a copy of the leaked document before him. Will he explain why the government is making plans for a social disaster in fishing communities instead of preventing the end of assistance for fishing communities and the people in those areas?
A. Mr. Speaker, I have never asked the RCMP to do the sorts of things he said in his question. I understand that some of our officials need some training to be able to cope with confrontational situations and to handle more difficult situations on an individual basis. It has happened not only in relation to TAGS but across Canada. This is the way it works. Our government is doing the right thing by conducting a review of the post-TAGS situation. We are not particularly worried because we trust Canadians and we know Canadians behave properly all the time.

Table 3: An example of the *denial* strategy used together with the *justification* strategy; 1997-11-21, Peter Stoffer (Q) and Pierre S. Pettigrew (A).

Score of Label$_i$ is a raw score of strategy i.
The extracted word pairs that were assigned highest scores based on the sets of features, patterns, or observed pairs of words were considered as candidates to be added to the training set.

4. Evaluation

In order to be able to examine the quality of the extracted candidates, we used a five-fold cross-validation approach for the extension and evaluation of the data. In each fold, we used 94 instances of the reputation defence dataset (Naderi and Hirst, 2017) for test, and the remaining for data extension (extracting patterns and observed word pairs from question and answer pairs) and classification task. We extended the training data once with only the observed word pairs, and once with only the pattern features. In each fold, the size of the training set varies according to the assigned scores. Since each answer can express multiple reputation strategies (see the example in Table 3) or none, we used a threshold value to decide whether to add the candidate pair to the training set or not. We examined various threshold values for each approach.
The quality of the extracted pairs was evaluated in two ways: (1) comparison with manual annotation, and (2) the contribution of the added training data to the classification of reputation strategies.

4.1 Inter-annotator agreement

To examine whether the assigned labels are of high quality, we conducted a study with 180 random question and answer pairs on the CrowdFlower platform. The ques-

(a) Does the answer express Concession?	(b) Does the answer express Justification?
Q. Mr. Speaker, my question is for the Minister of Labour. Former workers at Singer are arguing that the federal government did not fulfill its contract obligations toward them because it gave the company, instead of them, the Government Annuities Account surplus, that is a part of their pension funds that it was responsible for administering. Does the Minister of Labour not agree that the contract binding the parties between 1946 and 1957 is abundantly clear and that the federal government had an obligation to pay the surplus out to the workers and not to Singer? **A.** Mr. Speaker, all the federal regulations have been applied in this matter.	**Q.** Mr. Speaker, if we understand this correctly, 72% of Canada's refugee claimants have entered Canada from the United States of America, which means that 28% of refugees obviously come from refugee camps. Is the minister telling us that we are only accepting 28% of legitimate refugees to this country who actually deserve to be raised to higher levels? **A.** Mr. Speaker, the member is telling us that legitimate refugees are only people who we picked up, that everyone crossing our borders or arriving at our airports are not legitimate. He should be ashamed of himself.

Table 4: (a) Disagreement among six annotators, two of whom annotated it as *concession* and three as not *concession;* 1995-06-01, Claude Bachand (Q) and Lucienne Robillard (A). (b) Three of the annotators confirmed the answer as *justification* strategy and two as not *justification;* 2002-04-30, Rahim Jaffer (Q) and Denis Coderre (A).

All crowdsourced annotations			
(a) Observed word pairs			
t > .33	**t > .32**	**t > .31**	**t > .30**
.60	.71	.73	.70
(b) Extracted patterns			
t > .90	**t > .80**	**t > .70**	–
.41	.43	.43	–
Crowdsourced annotations with confidence > 80%			
(c) Observed word pairs			
t > .33	**t > .32**	**t > .31**	**t > .30**
.80	**.85**	.77	.76
(d) Extracted patterns			
t > .90	**t > .80**	**t > .70**	–
.41	.39	.38	–

Table 5: (a) Evaluation of automatically assigned strategies using observed word pairs against all crowd annotations; (b) Evaluation of automatically assigned strategies using extracted patterns against all crowd annotations; (c) Evaluation of automatically assigned strategies using observed word pairs against crowd annotations with confidence $> 80\%$; (d) Evaluation of automatically assigned strategies using extracted patterns against crowd annotations with confidence $> 80\%$. t is the threshold used for accepting the candidate labels.

tion and answer pairs were sampled from a pool of pairs that were assigned a reputation strategy label using the two approaches that were described earlier (see Sections 3.2 and 3.3).

Contributors were shown a question and answer pair with the assigned reputation defence strategy, as well as the description and conditions of the assigned strategy from Table 1. The contributors were then asked whether the assigned strategy was correct or not. We asked for at least five annotations per pair from the English-speaking countries. The contributors were presented with one test pair of question and answer and three other pairs on each page, and had to maintain 80% accuracy throughout the job. In total, the task included 66 *denial,* 5 *excuse,* 79 *justification,* and 30 *concession* questions. 81 of 180 were agreed by all 5 annotators. Only 59 answers were annotated with a confidence score below 80%. The confidence score is the agreement of the five annotators weighted by the annotators' trust scores.[3] Trust scores are determined by the annotators' accuracy on the test questions they have seen. Table 4 shows two examples of disagreement by the annotators. Most of the answers that caused disagreement among annotators evaded providing a response to the given question.

Table 5 shows what percentage of the automatically assigned strategies using word pairs and pattern acquisition approaches were correct compared to the crowdsourced annotations. We once considered all the crowdsourced data. We further removed the crowdsourced annotations with the confidence scores lower than 80%, and assessed the quality of the automatically assigned labels against higher-quality crowdsourced annotations. When compared with the crowdsourced annotations with a confidence score of at least 80%, the labels that were extracted using the observed word pairs approach with the threshold $t > .32$ shows the highest agreement. The automatically assigned labels using pattern acquisition approach show low agreement with the crowdsourced annotations.

4.2 Five-fold cross-validation

We further evaluated the quality of the data by assessing its contribution to the classification task. As mentioned earlier, we performed a five-fold cross-validation using the reputation defence dataset. The test set always came from the reputation defence dataset. We performed a multi-class classification using a class-weighted Support Vector Ma-

[3] https://success.crowdflower.com/hc/en-us/articles/201855939-How-to-Calculate-a-Confidence-Score

	Original	t > .33	t > .32	t > .31	t > .30
Train	379	512	1238	3797	8495
			BOW		
F_1	51.32	54.65	55.39	52.61	55.28
Accuracy	53.35	56.74	59.10	56.32	62.00
Denial	62.40	64.86	65.69	63.29	75.77
Excuse	13.60	17.00	13.64	13.64	3.64
Justification	55.60	62.42	66.39	63.50	67.14
Concession	36.40	32.00	25.00	14.32	11.02
			BOW+Negation+VerbNet+Similarity+Senti.+Disc.		
F_1	56.92	55.62	54.83	51.86	56.42
Accuracy	57.59	57.37	57.58	55.48	62.85
Denial	65.00	64.73	64.82	63.83	76.60
Excuse	18.00	17.00	17.27	17.00	6.60
Justification	59.80	62.30	64.75	63.05	67.50
Concession	48.00	37.74	24.30	13.01	10.80
			BOW+Negation+VerbNet		
F_1	53.22	54.77	56.01	53.05	55.29
Accuracy	54.22	56.11	58.84	56.74	62.01
Denial	63.60	64.73	65.60	63.45	75.95
Excuse	17.80	14.97	17.27	13.63	3.64
Justification	56.40	60.17	65.63	63.78	67.20
Concession	39.80	36.32	27.56	16.39	10.68

Table 6: Classification of reputation defence strategies using the extended training data with observed word pairs. The performance of classification of each strategy is reported in terms of average F_1. *t* is the threshold used for accepting the candidate labels.

	Original	t > .90	t > .80	t > .70
Train	379	453	486	573
		BOW		
F_1	51.32	48.52	47.63	47.51
Accuracy	53.35	49.99	49.15	48.94
Denial	62.40	56.94	54.73	57.54
Excuse	13.60	13.60	11.64	17.00
Justification	55.60	53.44	53.76	52.53
Concession	36.40	34.83	33.60	29.35
		BOW+Negation+VerbNet+Similarity+Senti.+Disc.		
F_1	56.92	49.00	49.10	49.53
Accuracy	57.59	50.84	50.62	51.26
Denial	65.00	56.60	56.01	56.61
Excuse	18.00	13.60	9.40	12.53
Justification	59.80	54.00	54.15	54.65
Concession	48.00	38.60	39.73	40.17
		BOW+Negation+VerbNet		
F_1	53.22	49.81	48.55	48.18
Accuracy	54.22	51.25	49.90	49.36
Denial	63.60	58.10	57.24	57.79
Excuse	17.80	18.10	18.10	27.42
Justification	56.40	54.32	53.30	51.66
Concession	39.80	36.94	34.58	30.89

Table 7: Classification of reputation defence strategies using the extended training data with patterns. The performance of classification of each strategy is reported in terms of average F_1. *t* is the threshold used for accepting the candidate labels.

chine model with a linear kernel[4] and the features proposed by Naderi and Hirst, including answer bag-of-words representations (weighted using *tf-idf*) of the answers, VerbNet verb classes, positive and negative sentiments, and negations in the answers, as well as discourse relations and similarity measure between the question and answer. We extracted the sentiments using OpinionFinder (Wilson et al., 2005) and discourse relations using End-to-End PDTB-Styled Discourse Parser (Lin et al., 2014). We further used the word2vec embeddings (Mikolov et al., 2013) for computing the similarity between the questions and answers.

Table 6 shows the results of the classification with the extended data using the observed word pairs approach. We used various threshold values (*t*) for accepting the candidates for the extension of the training data (train). Since in each fold the size of the extended data varies, we report the average size of the training sets of all folds. The baseline is the original dataset without any added data (the column specified as *original* in Table 6). The average F_1 measure of each reputation defence strategy is also presented. As shown in the table, by adding the automatically assigned labels to the training set, the performance of the classification of the *denial* and *justification* strategies improves; however, the data extension does not improve the classification of the *excuse* and *concession* strategies. Examining the extended data, we find that most of the added instances are *denial* and *justification* instances, and only a few pairs of questions and answers are annotated with the *excuse* and *concession* strategies. The reputation defence dataset consists of the total of only 36 *excuse* and 95 *concession* annotations; thus it is expected that the extended dataset includes very few of these strategies. Using the automatically added labels, the average F_1 measure of *denial* and *justification* reaches about 75% and 67%, respectively.

When we added the discourse relation and sentiment features, we did not observe any improvement in classification for the extended data. This can be due to having noise in the automatically assigned labels, and also the noisy nature of discourse relations and sentiment annotations.

Table 7 presents the results of the classification with the extended data using pattern acquisition approach. Extending the data using this approach does not result in a high-quality dataset and the performance of the classification drops very quickly. To improve the quality of the labels, we further examined whether removing the patterns that appeared in all the other strategies help. For example, for denial, we removed the patterns that appeared in non-denial examples. After removing the patterns that were shared between different strategies, we computed the scores introduced in Section3.3; however, we did not observe any improvements. Reputation defence strategies do not apply to all question and answer pairs (see the example in Table 8), and although we removed the few question and answer pairs annotated with *none* from the seed examples, we might be able to find these cases using a threshold value for accepting the candidate labels.

[4]LibSVM implementation (Pedregosa et al., 2011).

Q. Mr. Speaker, a week after the latest escalation in the conflict in Bosnia, when 370 peacekeepers, including 55 Canadians, were taken hostage by Serbian forces, there has been a flurry of statements and meetings which failed to produce any concrete results leading to the release of the hostages. This morning, the International Red Cross said that the Bosnian Serbs told them they would release the hostages unconditionally, either today or tomorrow. Could the Deputy Prime Minister confirm the statement by the Red Cross that the Bosnian Serbs will release the 370 peacekeepers who are being kept hostage sometime during the next few hours, although Bosnian Serb leader Radovan Karadzic said yesterday that no hostages could be released without guarantees that all air strikes would be suspended?
A. Mr. Speaker, we received communications mentioning that a few hostages might be released today, but at 11.13 a.m., we were unable to confirm whether that was the case.

Table 8: An example of an answer where *none* of the strategies apply; 1995-06-02, Gilles Duceppe (Q) and Sheila Copps (A).

5. Conclusion

We presented two approaches to automatically induce a corpus of reputation defence strategies. We considered pattern-based representation of reputation defence strategies and the observed pairs of words from the cross-product of questions and answers. We evaluated the generated data using the two proposed approaches against crowd annotation, and also assessed its contribution in the classification task. The observed word pairs approach resulted in a higher quality dataset. We found that the extended dataset using the observed word pairs contributes positively to the performance of the classifier, even though it contains noisy and weak labels.

Acknowledgements

This research is financially supported by the Natural Sciences and Engineering Research Council of Canada. We thank the anonymous reviewers for their suggestions.

Bibliographical References

Aristotle, Translated by Kennedy, G. (2007). *On Rhetoric: A Theory of Civic Discourse*. Oxford University Press.

Beelen, K., Thijm, T. A., Cochrane, C., Halvemaan, K., Hirst, G., Kimmins, M., Lijbrink, S., Marx, M., Naderi, N., Rheault, L., Polyanovsky, R., and Whyte, T. (2017). Digitization of the Canadian parliamentary debates. *Canadian Journal of Political Science*, 50(3):849–864.

Benoit, W. L. and Henson, J. R. (2009). President Bush's image repair discourse on Hurricane Katrina. *Public Relations Review*, 35(1):40–46.

Benoit, W. L. (1995). *Accounts, Excuses, and Apologies: A Theory of Image Restoration Strategies*. State University of New York Press, Albany.

Biran, O. and McKeown, K. (2013). Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 69–73, Sofia, Bulgaria. Association for Computational Linguistics.

Biran, O. and Rambow, O. (2011). Identifying justifications in written dialogs. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, pages 162–168, Washington, DC, USA. IEEE Computer Society.

Brinson, S. L. and Benoit, W. L. (1999). The tarnished star: Restoring Texaco's damaged public image. *Management Communication Quarterly*, 12(4):483–510.

Burns, J. P. and Bruner, M. S. (2000). Revisiting the theory of image restoration strategies. *Communication Quarterly*, 48(1):27–39.

Coombs, W. T. and Holladay, S. J. (2008). Comparing apology to equivalent crisis response strategies: Clarifying apology's role and value in crisis communication. *Public Relations Review*, 34(3):252–257.

Downey, S. D. (1993). The evolution of the rhetorical genre of apologia. *Western Journal of Communication*, 57(1):42–64.

Habernal, I., Eckle-Kohler, J., and Gurevych, I. (2014). Argumentation mining on the web from information seeking perspective. In Elena Cabrio, et al., editors, *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 26–39. CEUR-WS.

Harlow, W. F., Brantley, B. C., and Harlow, R. M. (2011). BP initial image repair strategies after the Deepwater Horizon spill. *Public Relations Review*, 37(1):80–83.

Hirst, G., Riabinin, Y., Graham, J., Boizot-Roche, M., and Morris, C., (2014). *From Text to Political Positions: Text analysis across disciplines*, chapter Text to ideology or text to party status?, pages 61–79. John Benjamins Publishing Company, Amsterdam.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2014). A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Lyon, L. and Cameron, G. T. (2004). A relational approach examining the interplay of prior reputation and immediate response to a crisis. *Journal of Public Relations Research*, 16(3):213–241.

Marcu, D. and Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368–375, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Naderi, N. and Hirst, G. (2017). Recognizing reputation defence strategies in critical political exchanges. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 527–535, Varna, Bulgaria.

Nguyen, H. and Litman, D. (2016). Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1127–1137, Berlin, Germany. Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peldszus, A. (2014). Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics.

Pitler, E., Louis, A., and Nenkova, A. (2009). Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 683–691, Suntec, Singapore. Association for Computational Linguistics.

Rheault, L., Beelen, K., Cochrane, C., and Hirst, G. (2016). Measuring emotion in parliamentary debates with automated textual analysis. *PloS one*, 11(12):e0168843.

Sheldon, C. A. and Sallot, L. M. (2008). Image repair in politics: Testing effects of communication strategy and performance history in a faux pas. *Journal of Public Relations Research*, 21(1):25–50.

Stab, C. and Gurevych, I. (2014). Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Tsur, O., Davidov, D., and Rappoport, A. (2010). ICWSM - A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In William W. Cohen et al., editors, *International AAAI Conference on Weblogs and Social Media*. The AAAI Press.

Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., Hirst, G., and Stein, B. (2017). Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Wang, L. and Cardie, C. (2014). Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 97–106, Baltimore, Maryland. Association for Computational Linguistics.

Whyte, T. (2014). Some honourable members: A quantitative analysis of parliamentary decorum in canada and the united kingdom. In *23rd World Congress of Political Science*, Montreal, Canada.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhang, E. and Benoit, W. L. (2009). Former Minister Zhang's discourse on SARS: Government's image restoration or destruction? *Public relations review*, 35(3):240–246.

Zhang, J., Spirling, A., and Danescu-Niculescu-Mizil, C. (2017). Asking too much? The rhetorical role of questions in political discourse. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1572, Copenhagen, Denmark. Association for Computational Linguistics.

Exploring the Political Agenda of the Greek Parliament Plenary Sessions

Dimitris Gkoumas, Maria Pontiki, Konstantina Papanikolaou, Haris Papageorgiou
Institute for Language and Speech Processing, Athena Research and Innovation Center
Artemidos 6 & Epidavrou, 15125, Athens, Greece
{dgkoumas, mpontiki, konspap, xaris}@ilsp.gr

Abstract

In this paper we present primary content analysis results for the Greek Parliament data in the context of Natural Language Processing and Text Mining approaches. The raw minutes of the Greek Parliament plenary sessions of the last 26 years are processed and transformed into a structured and machine readable format, and then clustered based on the analysis of their content using topic modelling techniques. Inspired by and following the work of Greene and Gross (2017) for the European Parliament, we employ a two-layer methodology for applying topic modelling in a Non-negative Matrix Factorization framework to a timestamped corpus of political speeches in order to explore dynamic topics. The results are visualized in various ways (by topic, by time) providing at the same time information about the contribution of each Parliament Member, political party and region (constituency) to each topic, and by extent, the ability to explore how the political and policy agenda has been shaped and evolved in Greece over time.

Keywords: Greek Parliament Data, Dynamic Topic Modelling

1. Introduction

Parliament plenary sessions depict the peak of the legislative work done by policy makers (members of the Parliament, government officials) and thus, their content constitutes a valuable source for the exploration of the way in which political and policy agendas are shaped and evolve over time (e.g. how problems and issues are defined, constructed, and placed on the political and policy agenda), as well as of the way in which the individual Parliament Members and political groups react and act over time (e.g. voting behavior). The recent work of Greene and Gross (2017) that analyses the content of the European Parliament plenary sessions using a dynamic topic modelling approach, indicates that the detection of latent themes in a timestamped corpus of political (legislative) speeches can provide insights of the way in which the political agenda reacts to exogenous events (e.g. the emergence of the Eurocrisis) and evolves over time. Such analysis can also supply information about the Parliament members' reactions to different stimuli. Different types of topic modelling approaches have been used in the political science literature also tracing the political attention of individual politicians over time based on the themes they speak about (Quinn et al., 2010), or capturing the political priorities expressed in Congressional press releases (Grimmer, 2010).

In this context, we present a work focusing on the Greek Parliamentary (GrParl) data. Inspired by and following the work of Greene and Gross (2017), we adopt their two-layer strategy for applying topic modelling in a Non-negative Matrix Factorization (NMF) framework to explore dynamic topics in the GrParl plenary sessions. The contribution of our work is two-fold: 1) Transformation (digitization/normalization/processing) of the GrParl plenary sessions minutes into a structured and machine readable format (Section 2) making them for the first time available for Natural Language Processing (NLP) and Text Mining approaches. 2) A platform that enables an insightful navigation across the GrParl plenaries based on the results of the topic modelling analysis (Section 3); the speeches can be explored by dynamic topics and by time. Information about the contribution of each GrParl member, political party and region (constituency) to each topic is also provided enabling a more perceptive monitoring of the political and policy agenda in Greece over time for all stakeholders (e.g. journalists, political & social scientists, policy makers).

2. Data Collection

The Greek Parliament provides the proceedings of all the plenary sessions during the time period from 1989 to 2015. The data was in different formats (e.g. txt/doc/docx/pdf/jpg). With the exception of the image format files covering the time period 1994-1996, the data was transformed into a structured and machine readable xml version, and organized based on its timestamp. In particular, the collection consists of 4356 plenary sessions containing in total 1.063.546 unique speeches.

2.1 Data Processing

The data was processed using our in-house Athena R.C./ILSP suite of NLP tools for the Greek language (Papageorgiou et al., 2002; Prokopidis et al., 2011). In particular, the text of each plenary session was enriched with the following types of information:

- Segment annotation: the start/end point of each speech.
- Speaker annotation: the name of each speaker and his/her political affiliation (the name of the party he/she represents in the Greek Parliament).
- POS tagging: the Part-of-Speech (e.g. adjective, noun) of each word in the text.
- Temporal annotation: the date of each speech.

In a next phase, the data was organized by the date of the speeches.

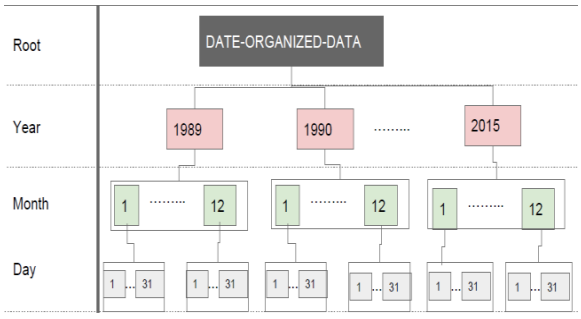


Figure 1: GrParl Data Organization.

2.2 Data Organization

Dividing temporal data into time windows of fixed duration when applying clustering techniques is often suggested in the literature (Sulo et al., 2010). After several experiments, we decided to set one month windows, and to have as many speeches as possible per window. In particular, all speeches made by Parliament members in the same month belong to the same window (see Figure 1). Thus, we ended up having 283 windows for the time period from 1989 to 2015. Hence, we were able to discover topics that were relevant to each month of each year as well as to monitor their dynamics over time.

3. Topic Modelling

Topic modelling is a widely used data analysis technique that provides an effective way to obtain insights in large collections of unlabeled data; topic models are used for inferring low-dimensional representations that uncover the latent semantic structures of different types of data; textual (Mcauliffe and Blei 2008), image (Wang, Blei, and Li 2009), or audio (Hoffman, Blei, and Cook 2009) data, among others. The state of the art topic modelling approaches are often based on Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Non Negative Matrix Factorization (NMF) (Greene and Gross 2017; Kuang, Choo, and Park 2015). NMF algorithms provide semantically meaningful output that is easily interpretable in clustering applications (Kuang, Choo, and Park 2015).

In particular, the TF-IDF weighting scheme that the NMF is using enables the calculation of the importance of a word to each text in a collection of documents based on weighted term-frequency values, and thus, the generation of various but semantically coherent topics that are less likely to be represented by the same high-frequency terms. Such an attribute is important for models applied in the context of political speech in parliaments, as it provides the ability to "*differentiate between broad procedural topics relating to the day-to-day running of plenary and more focused discussions on specific policy issues*" (Greene and Gross 2017).

3.1 Dynamic NMF Method

Following the work of Greene and Cross (2017), we employed a Dynamic NMF approach for the GrParl data. In Dynamic NMF, the idea is that after identifying a certain number of topics in each 1-month-window, the ones which are semantically similar are grouped together as "dynamic topics", rather than identifying topics from the beginning. In a first phase, the NMF model runs on the time windows into which the data is divided and generates topics for each window. In a second phase, the NMF algorithm is applied again, with the number of topics decided after applying the Topic Coherence via Word2Vec (TC-W2V) measure proposed by O' Callaghan et al. (2015); TC-W2V evaluates the relatedness of a set of top terms describing a topic by computing a set of vector representations (word embeddings) for all of the terms in a large corpus using the Word2Vec tool introduced by Mikolov et al. (2013). For more details about the NMF method consult Greene and Cross (2017).

The input for the topic models was the preprocessed and organized data described in Section 2. In particular, each window corresponds to the monthly sessions and consists of the speeches of all Parliament members. During the construction of the document-term matrix the stop words were filtered out (using a relative list) and only the content words were kept (adjectives, nouns, adverbs and verbs). During the TF-IDF calculation rare terms that appeared in less than 10 documents were removed. Given that the TF-IDF scheme calculates the frequency of each term in each document, and also in the whole collection of the documents, the initially generated topics were somewhat noisy. In order to deal with this, units with length less than three lines were discarded and only the nouns and the adjectives were kept. In this way, we ended up with 283 comprehensive and coherent window-based topics and 90 dynamic topics. To detect the number of dynamic topics *k*, we used the TC-W2V coherence measure as in (Greene and Cross, 2017) using the GrParl speeches as a training corpus for the extraction of the word embeddings.

3.2 Presentation of the Results

The extracted topics are listed in a table, where for each dynamic topic a user can see its most descriptive terms, the number of parties and the regions (constituencies) associated with it, and its frequency (i.e. the number of the time windows in which it appears). For example, as illustrated[1] below in Figure 2, D87 is a migration related topic described by the terms: "immigrant" (μετανάστης), "nationality" (ιθαγένεια), "foreigner" (αλλοδαπός), "immigration" (μετανάστευση), "refugee" (πρόσφυγας), "asylum" (άσυλο), "illegal immigrant" (λαθρομετανάστης), "hellenism" (ελληνισμός), "society" (κοινωνία), "homeland" (πατρίδα), "reception" (υποδοχή), "language" (γλώσσα), "emigrant" (απόδημος), "integration" (ένταξη), "identity" (ταυτότητα), "criminality" (εγκληματικότητα), "victim" (θύμα), "government" (πολιτεία).

ID	Terms	#Parties	#Regions	Frequency
D87	μεταναστης, ιθαγενεια, μεταναστευτικος, αλλοδαπος, μεταναστευση, προσφυγας, ασυλο, λαθρομεταναστης, ελληνισμος, κοινωνια, πατριδα, υποδοχη, γλωσσα, αποδημος, ενταξη, ξενος, ταυτοτητα, εγκληματικοτητα, θυμα, πολιτεια,	18	58	61

Figure 2: Snapshot of the table presenting the dynamic topics.

[1] The results are available at the following link: http://194.177.192.82/presentation/index.html

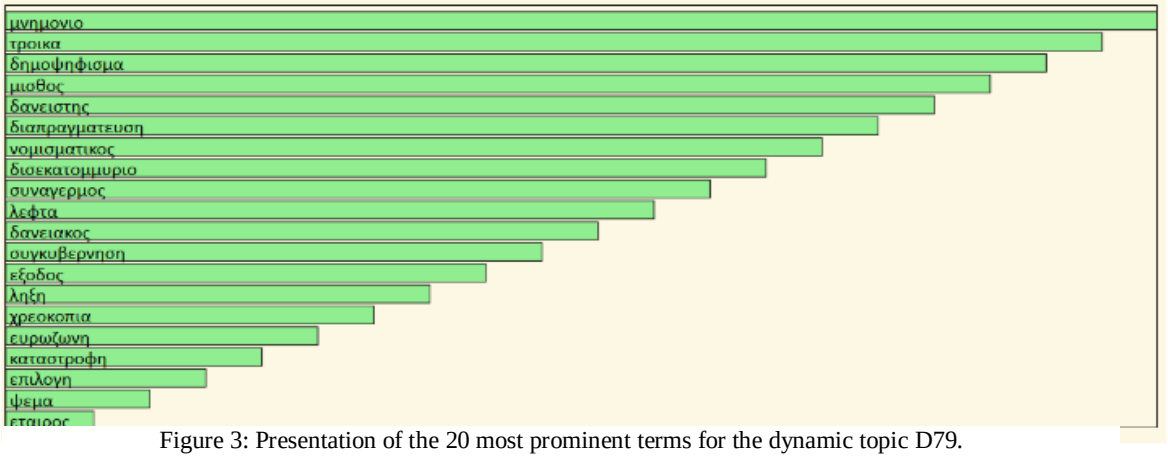


Figure 3: Presentation of the 20 most prominent terms for the dynamic topic D79.

By clicking on a row, users are able to study in detail the corresponding dynamic topic. In particular, the following types of information are provided:

- The 20 most prominent terms of each dynamic-topic (Figure 3). The words are presented in ascending order according to their importance, with the first one being the most descriptive term for each dynamic topic. For example, as illustrated above in Figure 3, the top term for the topic D79 is the word "memorandum" (μνημόνιο) and it is highly associated with the words "Troika" (Τρόικα), "referendum" (δημοψήφισμα), "salary" (μισθός), "creditor" (δανειστής), "negotiation" (διαπραγμάτευση), "monetary" (νομισματικός), etc.
- The most important monthly topics that are related to each dynamic topic. The window topics are listed in a table similar to the one in Figure 2, and presented in ascending order, with the first being the most important one. For each window topic, a user can see its timestamp, the most descriptive terms, the number of the politicians, parties and regions (constituencies) associated with it, as well as the number of the corresponding segments (speeches). By clicking on a table's row more information is available about the corresponding window topic.
- Information about the speakers that are related to each dynamic topic. In detail, users can see the speakers' names, the party they belong to, the region they are elected to represent and the date of each speech. By clicking on a table's row users can read or download the corresponding speech.
- A chart demostrating the contribution of each party to each dynamic topic (Figure 4).
- A chart demostrating the contribution of each region to each dynamic topic.

Users have also the option to explore the whole corpus of the GrParl plenary sessions based on the time windows that it has been divided to (283 in total) for the purposes of our analysis. As illustrated in Figure 5, for each time window we present the number of the topics, the terms, the parties, the politicians and the segments it is associated to. Again, by clicking on a table 's row, users are able to explore the corresponding window in more detail. Finally, users can use specific keywords as search terms to discover related topics.

4. Discussion

In this paper we present primary topic modelling results for the GrParl data. Following the two-layer NMF

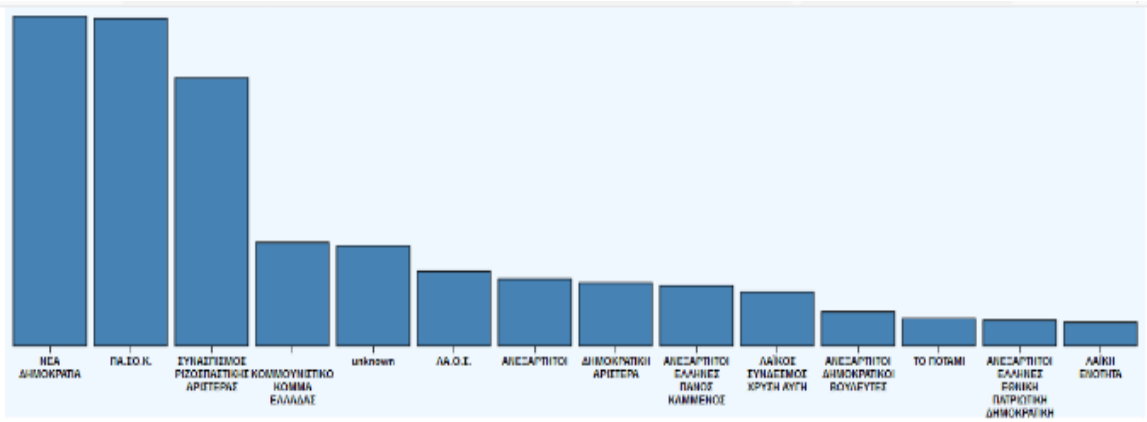


Figure 4: Chart demostrating the contribution of each political party to the dynamic topic D79.

Show 50 entries Search:

Time Window	#Topics	#Terms	#Parties	#Politicians	#Segments
2008_12	19	304	5	291	1216
2009_12	10	189	8	286	1369
2010_12	10	178	8	285	1187
2008_03	14	229	5	276	1654
2013_12	17	261	9	276	1444
2010_05	14	239	8	273	1523
2015_12	25	371	8	273	1327

Figure 5: Time-window based presentation of the GrParl corpus.

methodology proposed by Greene and Gross (2017) for identifying dynamic topics in large political speech corpora over time, we analysed the minutes of the GrParl plenary sessions during the time period from 1989 to 2015. More than one million speeches made by Parliament members during the last 26 years in Greece have been processed and analysed, and can be explored by topic and by time. Information about the contribution of each GrParl member, political party and region to each topic is also provided enabling a more insightful navigation across the GrParl plenaries. The next step of our work is the qualitative analysis of the extracted dynamic topics and the assignment of a descriptive label to each topic based on the most descriptive words appearing in speeches related to it.

In the context of an interdisciplinary work with political scientists, we plan to explore the evolution of selected dynamic topics of interest over time focusing on specific case studies, to examine their connection to significant events (e.g. refugee crisis, Eurocrisis), and to compare them with the corresponding ones in the European Parliament and, if possible, in other European countries. Future work also involves applying other types of content analysis techniques, for example Quotations Extraction and Sentiment Analysis, in order to enrich the original data with more insights for anyone interested to scrutinize the policy agenda as well as the policy makers' reactions and actions in Greece over time (e.g. journalists, political scientists, policy makers, stakeholders).

5. Acknowledgements

We acknowledge support of this work by the projects "Computational Science and Technologies: Data, Content and Interaction" (MIS 5002437) and "APOLLONIS, Greek Infrastructure for Digital Arts and Humanities, Language Research and Innovation" (MIS 5002738), which are implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund). Part of the work reported here was made possible by using the CLARIN infrastructure**.**

6. Bibliographical References

Blei, D. M., Ng, A.Y., Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022.

Greene, D. and Cross, J.P. (2017). Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis*.

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. Political Analysis 18(1), 1–35.

Hoffman, M., Cook, P., and Blei, D. (2009). Bayesian spectral matching: Turning Young MC into MC Hammer via MCMC sampling. In: International Computer Music Conference.

Kuang, D., Choo, J. and Park, H. (2015). Nonnegative matrix factorization for interactive topic modeling and document clustering. In: Partitional Clustering Algorithms, pp. 215–243.

Mcauliffe, J.D., and Blei, D.M. (2008). Supervised topic models. In: Advances in neural information processing systems, pp 121–128.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. CoRR abs/1301.3781.

O'Callaghan, D., D. Greene, J. Carthy, and P. Cunningham (2015). An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications (ESWA) 42(13), 5645–5657.

Papageorgiou, H., Prokopidis, P., Demiros, I., Giouli, V., Konstantinidis, A., and Piperidis S. (2002). Multi–level XML–based Corpus Annotation. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Spain, pp. 1723–1728.

Prokopidis, P., Georgantopoulos, B., and Papageorgiou, H. (2011). A suite of NLP tools for Greek. In: Proceedings of the 10th International Conference of Greek Linguistics, Komotini, Greece, pp. 373–383.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev (2010). How to analyze political attention with minimal assumptions and costs. American J. Political Science 54(1), 209–228.

Sulo, R., T. Berger-Wolf, and R. Grossman (2010). Meaningful selection of temporal resolution for dynamic networks. In Proc. 8th Workshop on Mining and Learning with Graphs, pp. 127–136. ACM.

Wang, C., Blei., D., and Fei-Fei, L. (2009). Simultaneous image classification and annotation. In: Computer Vision and Pattern Recognition.

Findings from the Hackathon on Understanding Euroscepticism Through the Lens of Textual Data

Federico Nanni[a,*], Goran Glavaš[a], Simone Paolo Ponzetto[a], Sara Tonelli[b], Nicolò Conti[c], Ahmet Aker[d,e], Alessio Palmero Aprosio[b], Arnim Bleier[f], Benedetta Carlotti[g], Theresa Gessler[h], Tim Henrichsen[i], Dirk Hovy[j], Christian Kahmann[k], Mladen Karan[l], Akitaka Matsuo[m], Stefano Menini[b], Dong Nguyen[n,o], Andreas Niekler[k], Lisa Posch[f], Federico Vegetti[p], Zeerak Waseem[d], Tanya Whyte[q], Nikoleta Yordanova[a]

[a]University of Mannheim, [b]Fondazione Bruno Kessler, [c]Unitelma Sapienza, [d]University of Sheffield, [e]University Duisburg-Essen, [f]GESIS Leibniz Institute for the Social Sciences, [g]Scuola Normale Superiore, [h]European University Institute, [i]Scuola Superiore Sant'Anna, [j]Bocconi University, [k]Liepzig University, [l]University of Zagreb, [m]London School of Economics, [n]Alan Turing Institute, [o]University of Edinburgh, [p]Central European University, [q]University of Toronto

Abstract

We present an overview and the results of a shared-task hackathon that took place as part of a research seminar bringing together a variety of experts and young researchers from the fields of political science, natural language processing and computational social science. The task looked at ways to develop novel methods for political text scaling to better quantify political party positions on European integration and Euroscepticism from the transcript of speeches of three legislations of the European Parliament.

Keywords: Political Text Scaling, Euroscepticism, Text as Data, Hackathon, Computational Social Science

1. Introduction

The unprecedented availability of large amounts of records of digital materials presents tremendous opportunities for political scientists, sociologists, historians, as well as any researchers focused on studying the present times (Grimmer and Stewart, 2013; Graham et al., 2016). However, traditionally trained social scientists and humanities scholars often lack the methodological expertise to examine, manage, and extract information from these large and noisy datasets of primary sources. On the other side of the spectrum, data scientists and natural language processing (NLP) researchers, who work with these resources on a daily basis, usually do not have the background knowledge to identify and address relevant research questions adopting these materials, particularly when it comes to extremely complex phenomena like the impact of the financial crisis in different socio-economical strata, the perception of the migrant crisis across countries and, especially for this paper, the growth of skepticism towards the European Union (EU).

The Goal. For these reasons, we decided to organize a three-day interactive seminar (a 'hackathon') that took place during the first half of December 2017 at Villa Vigoni,[1] with the aim of bringing together PhD and Postdoc researchers from different disciplines and guide them to work together on a series of new datasets. In this context, researchers had the possibility of sharing methodologies, discussing research questions, and cooperating in small interdisciplinary groups.

The focus of the hackathon was to develop new text-scaling algorithms for better understanding how Eurosceptic opinions emerge in institutional debates.

Outline. In the remainder of the paper, we first offer an overview of quantitative approaches for measuring Euroscepticism. Next, we present related work on the adoption of NLP approaches in political science research (in particular for text scaling) and on the benefits of hackathons for enhancing interdisciplinary research. We then describe the datasets adopted, the gold standard, and the addressed task. We share all resources used during the hackathon with the research community to support further work on the topic. The different approaches developed during the hackathons are briefly described before presenting the quantitative evaluation. Finally, the paper is wrapped up with a conclusion.

2. Background: Measuring Euroscepticism

During the last decade, a widespread opposition toward the EU strongly consolidated in several European countries. This phenomenon brought to the rise in consensus of parties critical to the EU from both the left and the right side of the political spectrum (Halikiopoulou et al., 2012). An example of such progressions is given by the results of the last European Parliament (EP) elections, where so-called Eurosceptic parties won 74 seats at the expenses of their mainstream counterparts when compared to previous EP elections in 2009.

Euroscepticism. This complex socio-political phenomenon has generally been labeled using the media-driven concept of Euroscepticism.[2] This term and its academic

*Regarding the order of the authors: first the five organizers, then the participants (alphabetical order).

[1] http://villavigoni.it

[2] Euroscepticism was first used by *The Times* in 1985 (hyphenated version of the term). It then spread to the political and academic environment becoming a real subfield of European studies 'a cottage industry of "Euroscepticism studies"' (Mudde, 2012).

study evolved hand in hand with the development of the EU itself. Initially, the study of Euroscepticism was fragmented and limited to countries where the phenomenon was present (Usherwood and Startin, 2013), while, since the 90s, alongside some events crucial to the evolution of the EU (e.g., the signing of the Single European act, the Maastricht Treaty), Euroscepticism has been extensively studied from two main perspectives: mass Euroscepticism and party-based Euroscepticism. The first one deals with voters' attitudes towards the EU, while the second one focuses on political parties' stances on the EU and European integration. In spite of the growing importance of Euroscepticism, scholars still struggle to provide a uniquely valid definition of the concept (Usherwood, 2016).

Issues with the Definition. Euroscepticism was firstly defined by Taggart as a "contingent and conditional opposition to the EU integration as well as a total and unconditional opposition to it" (Taggart, 1998). Since then, and after the dichotomy distinction between 'hard' and 'soft' Euroscepticism was coined (Taggart and Szczerbiak, 2002), a vivid dialogue between scholars in the field emerged to find the best one-size-fits-all definition of the concept (Flood, 2002; Kopeckỳ and Mudde, 2002; Conti, 2003; Rovny, 2004). In parallel with the absence of a precise definition of Euroscepticism, five main problems connected to this concept need to be further stressed. Firstly, the term itself may lead to conceptual confusion since it is composed by a prefix 'Euro' used as a proxy for the EU, a central component 'sceptic' which refers to the contraposition to the pro-EU "religious orthodoxy" (Cotta, 2016), and the suffix '-ism', which is generally used to identify ideologies, even if Euroscepticism cannot be considered as an ideology per-se but as a component of other ideologies (Flood, 2002; Vasilopoulou, 2009). Secondly, the term Euroscepticism is clearly negatively constructed (Crespy and Verschueren, 2009) and can thus be misused in political competitions to disparage political challengers both in an inter-party and in an intra-party perspective (Cotta, 2016). Thirdly, Euroscepticism's negative construction implies the recognition of a positive pro-European side that is in turn not well specified, i.e., it is sometimes difficult to draw clear boundaries between which party is or is not Eurosceptic. For example, is a party asking to reform the EU to be considered as Eurosceptic? If this is the case, how can we classify parties rejecting the EU? Fourthly, Euroscepticism generally identifies the EU and the European integration as a monolithic unit without distinguishing between what the EU is (the complex of EU institutions ruling member states, united under a single European community) and what the EU does (the outputs of the EU decision-making process in various policy fields). Lastly, as mentioned above, criticism towards the EU evolved hand in hand with the EU itself, therefore Euroscepticism has changed diachronically and cross-nationally. All the problems connected to the concept of Euroscepticism have led to more recent studies, arguing that it would be better to talk about Euroscepticisms using the plural form (Usherwood, 2016) or to reconceptualise it using the more neutral concept of 'political opposition' (Carlotti, 2017). Besides the above-mentioned problems, Euroscepticism is still widely used to understand both voters' and parties' positioning to the EU.

Traditional Approaches. Various sources of data have been used to estimate the position of political parties on Euroscepticism. Firstly, public opinion surveys, such as the European Election Study, allow measuring voters' perceptions of party positions via issue scales (Adams et al., 2014). Usually, such surveys are conducted periodically with each survey wave sampling new respondents, thus prohibiting a longitudinal analysis of changes in individual perceptions about party positioning. Second, party manifestos for national (Conti, 2003) and European elections (Schmitt et al., 2007) have served to estimate parties' stated preferences. However, as party manifestos are drafted for the purpose of elections, naturally they only offer a snapshot of parties' preferences every four to five years. Third, voting advice applications, such as EU Profiler for the European Parliament elections, offer data on political parties' self-positioning on various issue scales at election time. These data is even scarcer and do not offer more than a glimpse into parties' EU stances either. Fourth, to capture parties' revealed positions on European integration, scholars have relied on expert surveys, such as Chapel Hill Expert Survey (Polk et al., 2017), and surveys of members of parliament (Whitaker et al., 2017), which are conducted once every couple of years. A fifth common measure of parties' EU positions is based on parliamentary roll call votes (Hix et al., 2007). While roll call votes offer fine variation over time, they have been criticized for suffering from the selection bias (Carrubba et al., 2006; Yordanova and Mühlböck, 2015). Vote choice may also not reveal true preferences because it is constrained by party disciplining and the institutional rules (Hug, 2016), as well as strategic behavior on the party of legislators (Mühlböck and Yordanova, 2017).

While all these different approaches have already offered solid insights into the phenomenon, each of these solutions runs the risk of capturing only a few aspects of the overall perception of Continental society towards the European Union, and especially the reasons behind the growth of a widespread opposition to its politics and role. For this reason, the social science practice of survey research has always tried to move beyond these limitations (De Vreese, 2007). However, every social scientist knows the difficulties that survey research brings, both in terms of the time needed to conduct an extensive study and the accuracy of the final results, especially when it comes to analyzing extremely complex topics.

NLP-based Approaches. This brings us to the newest trend in estimating party stances with the use of textual data, i.e., *political-text scaling* (Grimmer and Stewart, 2013), such as from party speeches, press releases, parliamentary questions, etc (Wilde et al., 2014). The major advantages of this methodology are the abundance of such data and the recent developments of NLP approaches precisely tailored for supporting such applications (Glavaš et al., 2016; Glavaš et al., 2017a; Menini and Tonelli, 2016; Menini et al., 2017; Nanni et al., 2016; Zirn et al., 2016, *inter alia*). More substantively, it allows generating time-varying estimates of parties' EU positions. As with any other data, though, researchers have to carefully consider

the generation process behind textual data and its implications for the study at hand. For instance, when it comes to parliamentary speeches, parties may strategically decide whom to allow to speak so as to appear unified to the public (Proksch and Slapin, 2015). Also, different ideological dimensions seem to underlie voting behavior and speeches (Proksch and Slapin, 2010). Understanding party and institutional constraints of giving speeches as well as legislators' motivations to speak, is thus essential in judging what speech can tell us about political preferences.

3. Related Work

In this section we briefly present an overview of previous studies on political text scaling and the advantages of organizing shared tasks and hackathons in order to build new bridges between interdisciplinary communities.

Political Text Scaling. The goal of political scaling is to order political entities, such as political parties and politicians, according to the position they expressed in textual content. The type of orientation could be ideological (i.e., left vs. right) as well as policy-specific (regarding economics or welfare). Documents such as parties' election manifestos or transcripts of speeches are commonly used as the data underpinning this type of analysis (Grimmer and Stewart, 2013). Although the idea of estimating ideological beliefs is not new (Abelson and Carroll, 1965), nevertheless the first models able to estimate these beliefs from texts have only appeared in the last fifteen years (Laver and Garry, 2000; Laver et al., 2003; Slapin and Proksch, 2008; Proksch and Slapin, 2010). The seminal works by Laver and Garry (2000) and Laver et al. (2003) are widely considered the starting points of this field of research. These supervised approaches rely on predefined dictionaries of words or reference documents for establishing the position of unlabeled texts. In order to avoid the manual annotation effort (and the biases that this could add to the model), Slapin and Proksch (2008) proposed Wordfish, an unsupervised scaling model which has become the *de facto* standard method for unsupervised political text scaling. This approach models document positions and contributions of individual words to those positions as latent variables of the Poisson naïve Bayes generative model, i.e., they assume that words are drawn independently from a Poisson distribution. They estimate the positions by maximizing the log-likelihood objective in which word variables interact with document variables.

While this previous methodological research has been conducted by the political science community, in recent years works on political text scaling have also been presented by NLP groups (e.g., Nanni et al. (2016)). Among them, in particular Glavaš et al. (2017b), has proposed a new text scaling approach that leverages semantic representations of text, making it suitable both for mono- and cross-lingual political text scaling. The authors of this paper have shown that the semantically-informed scaling models better predict the party positions than Wordfish in two different political dimensions and that the models exhibit no drop in performance in the cross-lingual setting compared to monolingual one.

Gold Standard for Scaling. Generally, expert surveys are regarded as one of the most popular survey-based approaches for the estimation of parties positioning on several issues and as gold-standard for measuring the quality of text scaling algorithms. The rationale behind them is that experts in the field (e.g., political scientists) evaluate parties positioning on several issues on the basis of their domain knowledge. The resulting parties positioning is given by the aggregation of experts' judgments using measures of central tendency (e.g., the mean). Nonetheless, as various experts in the field suggest, the use of the Chapel Hill expert survey, as every expert survey, shows both advantages and drawbacks; this section briefly overviews them. The first problem connected to expert surveys is that it is not clear 'what' experts actually evaluate (Budge, 2000) since they are generally asked questions with 'minimal instructions' (Gemenis, 2015). In other words, experts are asked to provide judgments without having 'reference points', consequently making such judgments interpersonally and cross-nationally incomparable. Steenbergen and Marks (2007) demonstrate that such inter-expert disagreement correlates with certain parties' characteristics like their size and ideological background. However, according to the proponents of expert surveys, such an inter-experts disagreement may be solved through statistical aggregation, since the errors 'will cancel out' (Steenbergen and Marks, 2007). However, such an error component is not only a function of parties' characteristics, but also of experts' personal characteristics such as their knowledge or ideological background. This last consideration is connected to the second main problem of expert surveys: there is a great variance in the criteria used by experts to make their judgments. According to Curini (2010), "the estimation of parties positioning on the basis of survey data (broadly speaking) is not always consistent since respondents tend to place parties they like closer to where they perceive themselves to be, and to place those parties they dislike farther away then the actual position would warrant, thus producing a bias known as rationalization or projection" (see also Granberg and Brown (1992).[3] Since expert surveys aim is to estimate parties' positioning and not to infer the attributes of the experts' population on the basis of a set of actual respondents, relying on them may affect the validity of the obtained results (Curini, 2010).

Besides these problematic aspects, expert surveys are able to provide information in a common, standardized format across a wide range of countries. They are generally regarded as having weight and legitimacy, since they reflect the judgments of experts who are presumably well informed about the topic. Lastly, expert surveys are easily compared to other forms of analyses like the content analysis of parties manifestos or the observation of legislative behavior (through the use of roll call votes), which are in turn not free from biases either. Despite the potential drawbacks of the Chapel Hill expert survey, we relied on it to have a quick and easy way to position parties along the pro-against

[3] More specifically assimilation effects realize themselves when respondents shorten the distance between themselves and the party they favor while widening the distance between themselves and the parties they do not support.

European integration dimension.

Hackathons in DH and CSS. In the last decade, the NLP community has been involved in the organization of several activities aiming to bridge the gap between the field, the digital humanities and the computational social sciences. From workshops[4] and shared-tasks,[5] all the way through seminars,[6] summer schools,[7] and tutorials,[8] large efforts have been made to present and working together on new datasets, tools, and platforms in order to address relevant research questions, following a "more hack, less yack" attitude (Nowviskie, 2014). Among these efforts, some interdisciplinary hackathons similar to ours have been organized in the recent years: the Archives Unleashed[9] series organized five times since 2016, brought together digital archivists, humanities scholars, and computer scientists interested in the use of web archives (such as the Internet Archive) for studying the recent past. Other similar projects have been focused on biodiversity,[10] the 2016 US Elections,[11] and Tibetan studies (Almogi et al., 2016).

Inspired by these previous projects, in the hackathon we organized at Villa Vigoni, we decided to combine this highly interdisciplinary setting with a shared-task focused on developing new algorithms for text scaling.

4. The Hackathon

At the beginning of December 2017, 18 researchers (mainly PhD students and postdocs) with a background in political science, computational social science, or natural language processing took part in the hackathon. Upon arrival, the participants have been divided by the organizers in five interdisciplinary teams, named after national European football teams that did not manage to qualify to the final stage of 2018 World Cup. Then, the participants received an overview of the hackathon's shared-task, which they had 48 hours to address. The task was to develop new text-scaling algorithms tailored for identifying Eurosceptic opinions in institutional debates. Following, the organizers introduced the datasets and evaluation framework, as presented next.[12]

Parliamentary Text Collection. Given the focus on institutional debates, the organizers first crawled and provided to the participants all individual speeches of all European Parliament representatives, in all languages available (i.e., in the original language of the speech and all manual translations to other languages, if existing) from the official website of the European Parliament.[13] The collected materials cover 4 legislations (5th to 8th) and almost 20 years of European politics (1999-2017), and include a large variety of topics, ranging from the advent of Euro, the enlargement of the Union to the economic and refugee crises, and the growth of Euroscepticism. The raw corpus consists of four subparts (one for each legislation period), with one XML document per representative aggregating all speeches that each one delivered over the course of the legislation period (see Fig. 1). Besides the speeches (content and date for each one), for each representative we also obtained the information on the national party and European party group.

Leg. period	# parties	Min. len	Avg. len
5th (1999–2003)	25	14.5K	127.7K
6th (2004–2008)	30	13.9K	96.4K
7th (2009–2013)	24	54.9K	467.0K

Table 1: Per-legislation term statistics of the European Parliament dataset used in the hackaton.

New Benchmark Dataset. For the purpose of the hackathon, we considered only the speeches made or manually translated into English. We concatenated all speeches of all representatives of the same party into a single party-level document. Following previous works (Proksch and Slapin, 2010; Glavaš et al., 2017b), we selected the parties from the five largest European countries: Germany, France, United Kingdom, Italy, and Spain. Finally, we discarded the parties for which the aggregate texts over the whole legislation period ended up being shorter than 10.000 tokens. We decided to use only the data from completed legislation periods, which is why we discarded the ongoing eighth period. In Table 1 we provide some details on the final datasets produced for each legislation period – the number of parties along with the smallest and average party-text length (in number of tokens).

Gold Standard. As gold standard party positions we consider the European integration dimension from the Chapel Hill expert survey (years 2002, 2006, 2010). The Chapel Hill expert survey estimates national parties positioning on a variety of policy issues, European integration included. It is conducted every four years (in the occasion of EP elections) since 1999. The number of included countries increased through time, moving from 14 Western European Countries in 1999 to 31 countries in 2014, thus including those EU member states entering the EU during the various EU enlargement steps. The last wave of the Chapel Hill expert survey includes 268 parties from 31 countries.

Task Formulation. Given a series of documents, each one representing the concatenation of all speeches of the candidates of a European party, develop an algorithm able to place them into a single-dimensional space between 0 and 1, where 0 represents a strongly Eurosceptic position and 1 strongly in favour of European integration. To do so, any external resource could be used (i.e., information from a knowledge base such as DBpedia (Auer et al., 2007)), excluding information regarding the political position of the party to be scaled (e.g., the Italian Movimento 5 Stelle as being an Eurosceptic party). This output had to be derived solely from the textual content of the document.

[4] https://sites.google.com/site/nlpandcss/nlpcss-at-emnlp-2016
[5] https://sharedtasksinthedh.github.io/
[6] https://cds.nyu.edu/text-data-speaker-series/
[7] http://essexsummerschool.com/
[8] http://topicmodels.west.uni-koblenz.de/
[9] http://archivesunleashed.org/
[10] https://www.idigbio.org/content/citscribe-hackathon
[11] https://brown.columbia.edu/election-hackathon
[12] We make all the collections used during the hackathon available at: https://federiconanni.com/hack-vigoni/
[13] http://www.europarl.europa.eu

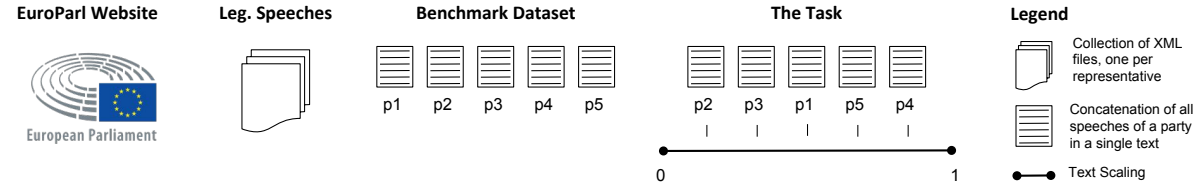


Figure 1: Graphical representation of the creation of the shared task.

5. Overview of the Proposed Approaches

All teams started tackling the scaling problem by manually inspecting the data to observe their structure. Thanks to this, they realized that the majority of the available data were not dealing with issues related to the EU and the European integration project (and therefore to determine if a party is pro or against the EU) but rather with technical aspects belonging to supranational decision-making process (e.g., discussion related to a specific policy issue). To keep only the relevant information, they adopted different filtering strategies. Next, the scaling step has been also approached in different ways. We report below a short overview of the different systems proposed and developed during the hackathon; we also encourage all the participants to continue collaborating on these initial ideas and to present the final results as independent research papers.

Team Greece (Aker, Carlotti, Matsuo, Niekler): To keep only the relevant information, this team used a dictionary for filtering out document irrelevant for the European integration. They identified EU resources (list of terms) available online discussing issues related to the EU and the European integration project, and constructed a dictionary containing only uni-grams and bi-grams. The entries of the dictionary were used in the filtering process. Each speech is regarded as one instance, which consists of multiple sentences. This filtering works at the sentence level. They used the dictionary entries to filter out any sentence within each speech that does not share any entry in the dictionary.
Using the trimmed instances (containing sentences related to the problem) they perform standard bag-of-words feature extraction (with uni-grams and bi-grams) along with feature selection. For feature selection they disregarded any word that occurred in more than 75% of the instances as well as in only 1% of the instances. Furthermore, they used chi-square test to remove further insignificant words leading to a feature vector containing 1500 words.
For each instance they extracted a feature vector containing those significant 1500 terms. The feature values are simple word counts. They used a linear SVM regression model, where the outcome is the true score, with hyper-parameter tuning. The model is capable to score each instance between 1 (pro EU) and 0 (non pro EU). As a number of speech instances are coming from the affiliates of one party, there are a number of predictive scores for each party. They use the median as the final predictive score. For comparison purposes, they repeated the experiments without the filtering process, i.e., feature vectors were extracted without removing any sentence. However, they applied the same feature selection as performed with the dictionary filtering case. They refer to this last experiment as "without dictionary" and the former experiment as "with dictionary". Against their intuition, the obtained results show that the inclusion of all datasets and sentences performs better on the task than filtering the sentences. This needs further investigation in order to improve and adapt the dictionary to the task.[14]

Team Ireland (Bleier, Menini, Waseem, Yordanova): The team used Will Lowe's package Jfreq[15] to pre-process the documents: they lowercased, removed numbers and currencies and stemmed all remaining words. A tailored list of stopwords, created considering the specificities of the corpus, was also adopted. Then they used the R implementation of Wordfish (Slapin and Proksch, 2008), from the Austin package, to scale the documents and they tested different word-filtering approaches to improve the results.

Team Italy (Gessler, Hovy, Karan): The team filtered first the speeches based on a list of manually selected keywords, then used paragraph2vec (Le and Mikolov, 2014) on all the speeches (from both scored and unscored parties) to learn distributed party (and word) representations. Then, the resulting matrix of the representation for known parties, together with their respective scores, was used as input to a canonical correlation analysis (CCA) (Hardoon et al., 2004). This step tries to find the first component that explains the variation in the party representations with respect to the observed scores. The fitted CCA model was then applied to the matrix of representations for new parties and the resulting one-dimensional vector was used as score prediction. These scores reach correlation values of up to 0.73 (Spearman) with the gold standard scores.

Team Netherlands (Aprosio, Henrichsen, Nguyen): The team explored a supervised learning approach. The data was segmented into individual speeches. A Linear Regression model was trained based on the labeled data to estimate a score for the individual speeches, and the final score was computed by taking the mean of these scores. The data included speeches covering a wide range of topics. However, speeches about the enlargement were considered the most relevant to a party's position regarding European integration. Therefore, for the final predictions, only those speeches were included that were about the enlargement based on one of the following words: 'enlargement', 'integration', 'accession', 'extension'. To overcome the small amount of labelled data, Ridge Regularization was used

[14] Code is available here: https://github.com/eisioriginal/eu_scepticism_regression

[15] http://conjugateprior.org/software/jfreq/

to prevent overfitting (alpha=1.5) and a small amount of noise was added to the labels. Scikit-learn (Pedregosa et al., 2011) was used to train the models, and the hyper-parameters were set using cross-validation on the training set. Only words were kept that appeared in at least 10 speeches and words appearing in more than 10% of the speeches were discarded. Both unigrams and bigrams were used. The results on the validation data were 0.573 (Spearman) and 0.733 (Pearson). The submitted runs included a model trained on both the training and validation data, and a model trained on only the training data.

Team Wales (Kahmann, Posch, Vegetti, Whyte): The approach of the team was based on Party Manifestos data, containing sentences classified (by experts) into different policy categories. The manifestos of UK parties were used because they were the only ones written in English. Some standard pre-processing was applied (lowercasing, removing numbers and stopwords, stemming). The policy categories of interest are European Community/Union (+/−) and National Way of Life (+/−). Based on these sentences a Naive Bayes classifier with three classes was used: (1) not related, (2) pro EU and (3) contra EU. After training this classifier on the manifesto data, it was applied to the test data. Before that, the speeches were split into single sentences and pre-processed. The classifier yielded three values for every sentence of every party. The three values indicate the posterior probability of a sentence belonging to one of the three categories. In order to get a single value for every party, they first excluded all sentences under a certain probability threshold (0.25, 0.5, 0.75) in the first category (not related). Having done this, they calculated a ratio score for every party computed as follows: $\log(\frac{\sum EU_{pro}}{\sum EU_{contra}})$. In the last step they normalized these party scores to the range [0,1].

6. Evaluation

We provided the datasets comprising the 5th and 7th legislation periods as development datasets to the participants. We kept the 6th legislation period (aggregate party texts and gold party positions from Chapel Hill Expert Survey) for final evaluation.

Evaluation Metrics. We use three evaluation metrics for comparing model-produced positions with the gold-standard positions:

- *Pairwise accuracy (PA)* is the percentage of pairs of parties for which the gold scores for the two parties on European integration are in the same relative order as the predicted scores for these two parties. In other words, prediction for a pair of parties A and B is considered correct if party A is more eurosceptic (pro-European) than party B both according to the gold standard and predicted position;
- *Spearman correlation (r_S)* between the set of gold party positions on European integration and those predicted by participants' systems;
- *Pearson correlation (r_P)* between the gold and predicted sets of party positions.

Team/Model	PA (%)	r_S (%)	r_P (%)
Random	51.3	2.6	6.2
WordFish (baseline)	61.8	34.5	29.5
Team Ireland	57.6	17.5	29.4
Team Wales	60.4	28.9	28.2
Team Netherlands	66.8	46.6	59.3
Team Greece	68.5	54.2	64.5
Team Italy	**70.3**	**54.3**	**72.8**

Table 2: Official hackathon results – scaling performance achieved by the best submitted run of each team.

Pairwise accuracy and Spearman correlation capture only the correctness of the ranking of the parties. In contrast, Pearson correlation also takes into consideration the extent to which automated scaling reflects the gold distances between party positions. Put differently, a system that produces the position scores that generate the same party ranking (i.e., the same order of parties from most eurosceptic to most pro-European) as the gold scores will have the perfect PA and r_S, but it will only have perfect r_P if it predicts exactly the same position scores as in the gold standard for all parties. Before evaluating the systems, we linearly scaled both the gold standard scores and system-produced scores to the $[0, 1]$ range.

Results. In Table 2 we show the performance achieved by the best submitted run of each team of participants on the dataset compiled from the 6th legislative period of the European Parliament. Along with the performances of the best runs from all teams, we show the performance of the WordFish model (Proksch and Slapin, 2010), the *de facto* standard model for text scaling in political science. As a sanity check, we also evaluated a baseline that randomly generates party positions.

All teams outperformed the random baseline by a wide margin. Three teams also outperformed the standard scaling algorithm WordFish, with the best-performing approach (Team Italy) outperformed WordFish by 10% in terms of pairwise accuracy, and 20 and 40 points in terms of Spearman and Pearson correlation, respectively.

7. Discussion

In addition to the quantitative outcome of the task, the hackathon made possible that scholars from very different backgrounds, research topics, and methodologies spent two days working together sharing ideas and approaches, each of them excited to think out-of-their-own disciplinary box. While it is generally not so easy to establish such communication channels across disciplines, given the different methodological approaches, research focuses, and even vocabulary (e.g., the meaning of the verb "to code" in computer science and political science), the participants have been incredibly willing to establish a common ground, for cooperating and addressing the task presented to them.

There is still much work that, as organizers of such events, we can do to improve this type of collaborative shared-task, from offering easier-to-digest presentations on the theoretical foundations of the political-science topic under study

to establishing methodological debates accessible to the entire audience. Nevertheless, we hope that the collaborations that will bloom thanks to this hackathon will facilitate the communication across research fields and support the future of interdisciplinary research between NLP and political science.

8. Conclusions

In this paper we presented an overview and the results of the first shared-task hackathon organized on the topic of scaling transcripts of speeches from the European parliament regarding Euroscepticism. The hackathon brought together 23 researchers (5 organizers and 18 participants) from 15 institutions with a large variety of backgrounds, from political science to computational social science and natural language processing, which worked together in five small teams for 48 hours in order to develop new approaches for the task. The output of the hackathon has been incredible: in two days these teams developed methods capable of outperforming the most established scaling algorithm in the field, WordFish, by a large margin. This highlights the immense potential of interdisciplinary collaborations and the usefulness of shared-task hackathons for bridging different research communities.

Acknowledgments

We thank the Deutsche Forschungsgemeinschaft (DFG) for generously supporting the research seminar and the hackathon. This work was also supported by the SFB 884 on the Political Economy of Reforms at the University of Mannheim (project C4). We thank the staff of the "Villa Vigoni Centro Italo-Tedesco per l'Eccelenza Europea" for their immense hospitality.

9. References

Abelson, R. P. and Carroll, J. D. (1965). Computer simulation of individual belief systems. *The American Behavioral Scientist (pre-1986)*, 8(9):1–24.

Adams, J., Ezrow, L., and Somer-Topcu, Z. (2014). Do voters respond to party manifestos or to a wider information environment? an analysis of mass-elite linkages on european integration. *American Journal of Political Science*, 58(4):967–978.

Almogi, O., Dankin, L., Dershowitz, N., and Wolf, L. (2016). A hackathon for classical Tibetan. *arXiv preprint arXiv:1609.08389*.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Budge, I. (2000). Expert judgements of party policy positions: Uses and limitations in political research. *European Journal of Political Research*, 37(1):103–113.

Carlotti, B. (2017). The odd couple: analyzing united kingdom independence party (UKIP) and italian five stars movement's (FSM's) european union (EU)-opposition in the european parliament (EP). *Italian Political Science Review/Rivista Italiana di Scienza Politica*, pages 1–24.

Carrubba, C. J., Gabel, M., Murrah, L., Clough, R., Montgomery, E., and Rebecca, S. (2006). Off the Record: Unrecorded Legislative Votes, Selection Bias and Roll-Call Vote Analysis. *British Journal of Political Science*, 36(4):691–704.

Conti, N. (2003). *Party attitudes to European integration: a longitudinal analysis of the Italian case*. Sussex European Institute Brighton.

Cotta, M. (2016). Un concetto ancora adeguato? l'euroscetticismo dopo le elezioni europee del 2014. *Contro l'Europa? I diversi Scetticismi verso l'Integrazione Europea*, pages 233–248.

Crespy, A. and Verschueren, N. (2009). From euroscepticism to resistance to european integration: an interdisciplinary perspective. *Perspectives on European politics and society*, 10(3):377–393.

Curini, L. (2010). Experts' political preferences and their impact on ideological bias: An unfolding analysis based on a Benoit-Laver expert survey. *Party Politics*, 16(3):299–321.

De Vreese, C. (2007). A spiral of euroscepticism: The media's fault? *Acta Politica*, 42(2-3):271–286.

Flood, C. (2002). Euroscepticism: A problematic concept (illustrated with particular reference to France). In *32nd annual UACES conference, Belfast*, pages 2–4.

Gemenis, K. (2015). An iterative expert survey approach for estimating parties' policy positions. *Quality & Quantity*, 49(6):2291–2306.

Glavaš, G., Nanni, F., and Ponzetto, S. P. (2016). Unsupervised text segmentation using semantic relatedness graphs. In **SEM*, pages 125–130.

Glavaš, G., Nanni, F., and Ponzetto, S. P. (2017a). Cross-lingual classification of topics in political texts. In *NLP+CSS Workshop*, pages 42–46.

Glavaš, G., Nanni, F., and Ponzetto, S. P. (2017b). Unsupervised cross-lingual scaling of political texts. In *EACL*, pages 688–693.

Graham, S., Milligan, I., and Weingart, S. (2016). *Exploring big historical data: The historian's macroscope*. World Scientific.

Granberg, D. and Brown, T. A. (1992). The perception of ideological distance. *Western Political Quarterly*, 45(3):727–750.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Halikiopoulou, D., Nanou, K., and Vasilopoulou, S. (2012). The paradox of nationalism: The common denominator of radical right and radical left e euroscepticism. *European Journal of Political Research*, 51(4):504–539.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Hix, S., Noury, A. G., and Roland, G. (2007). *Democratic Politics in the European Parliament*. Cambridge University Press, New York.

Hug, S. (2016). Party pressure in the european parliament. *European Union Politics*, 17(2):201–218.

Kopeckỳ, P. and Mudde, C. (2002). The two sides of euroscepticism: party positions on European integration in East Central Europe. *European Union Politics*, 3(3):297–326.

Laver, M. and Garry, J. (2000). Estimating policy positions from political texts. *American Journal of Political Science*, 44 (3):619–634.

Laver, M., Benoit, K., and Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97 (2):311–331.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196.

Menini, S. and Tonelli, S. (2016). Agreement and disagreement: Comparison of points of view in the political domain. In *COLING*, pages 2461–2470.

Menini, S., Nanni, F., Ponzetto, S. P., and Tonelli, S. (2017). Topic-based agreement and disagreement in us electoral manifestos. In *EMNLP*, pages 2928–2934.

Mudde, C. (2012). The comparative study of party-based euroscepticism: the sussex versus the north carolina school. *East European Politics*, 28(2):193–202.

Mühlböck, M. and Yordanova, N. (2017). When legislators choose not to decide: Abstentions in the european parliament. *European Union Politics*, 18(2):323–336.

Nanni, F., Zirn, C., Glavaš, G., Eichorst, J., and Ponzetto, S. P. (2016). TopFish: topic-based analysis of political position in US electoral campaigns. In *PolText*.

Nowviskie, B. (2014). On the origin of "hack" and "yack". *Journal of Digital Humanities*, 3(2):3–2.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Polk, J., Rovny, J., Bakker, R., Edwards, E., Hooghe, L., Jolly, S., Koedam, J., Kostelka, F., Marks, G., Schumacher, G., Steenbergen, M., Vachudova, M., and Zilovic, M. (2017). Explaining the salience of anti-elitism and reducing political corruption for political parties in europe with the 2014 chapel hill expert survey data. *Research & Politics*, Online first:1–9.

Proksch, S.-O. and Slapin, J. B. (2010). Position taking in european parliament speeches. *British Journal of Political Science*, 40(3):587–611.

Proksch, S.-O. and Slapin, J. (2015). *The Politics of Parliamentary Debate: Parties, Rebels, and Representation*. Cambridge University Press, Cambridge.

Rovny, J. (2004). Conceptualising party-based euroscepticism: Magnitude and motivations. *Collegium: news from the College of Europe= nouvelles du Collège d'Europe*, (29):31–48.

Schmitt, H., Braun, D., Popa, S. A., Mikhaylov, S., and Dwinger, F. (2007). *European Parliament Election Study 2014, Euromanifesto Study*. GESIS Data Archive.

Slapin, J. B. and Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

Steenbergen, M. R. and Marks, G. (2007). Evaluating expert judgments. *European Journal of Political Research*, 46(3):347–366.

Taggart, P. and Szczerbiak, A. (2002). *The party politics of Euroscepticism in EU member and candidate states*. Sussex European Institute Brighton.

Taggart, P. (1998). A touchstone of dissent: Euroscepticism in contemporary western european party systems. *European Journal of Political Research*, 33(3):363–388.

Usherwood, S. and Startin, N. (2013). Euroscepticism as a persistent phenomenon. *JCMS: Journal of Common Market Studies*, 51(1):1–16.

Usherwood, S. (2016). 2 modelling transnational and pan-european euroscepticism. *Euroscepticism as a Transnational and Pan-European Phenomenon: The Emergence of a New Sphere of Opposition*, page 14.

Vasilopoulou, S. (2009). Varieties of euroscepticism: the case of the european extreme right. *Journal of Contemporary European Research*, 5(1):3–23.

Whitaker, R., Hix, S., and Zapryanova, G. (2017). Understanding members of the european parliament: Four waves of the European parliament research group MEP survey. *European Political Science Review*, 18(3):491–506.

Wilde, P., Michailidou, A., and Trenz, H. (2014). Converging on euroscepticism: Online polity contestation during european parliament elections. *European Journal of Political Research*, 53(4):766–783.

Yordanova, N. and Mühlböck, M. (2015). Tracing the bias in roll call votes: Party group cohesion in the european parliament. *European Political Science Review*, 7(3):373–399.

Zirn, C., Glavaš, G., Nanni, F., Eichorts, J., and Stuckenschmidt, H. (2016). Classifying topics and detecting topic shifts in political manifestos. In *PolText*.

A Pilot Gender Study of the Danish Parliament Corpus

Dorte Haltrup Hansen, Costanza Navarretta, Lene Offersgaard
University of Copenhagen, Centre for Language Technology
Njalsgade 136, DK-2300 Copenhagen, Denmark
dorteh@hum.ku.dk, costanza@hum.ku.dk, leneo@hum.ku.dk

Abstract

This paper describes a pilot analysis of gender differences in the revised transcripts of speeches from the sittings in the Danish Parliament in the period from 2009 to 2017. Information about the number and duration of the speeches, the gender, age, party, and role in the party was automatically extracted from the transcripts and from other data on the Danish Parliament web site. The analysis shows statistically significant differences in the number and duration of the speeches by male and female MPs, and we also found differences in speech frequency with respect to the age of the MPs. Our analysis confirms previous studies on parliamentary data in other countries showing that the role of the MPs in their party influences their participation in the debates. Furthermore, we found that female ministers were speaking more in the period with a female prime minister than they did under a male prime minister. In the future, we will determine the statistical significance of the various parameters we have analysed in this paper and automatically extract linguistic information which can further determine differences between male and female MPs of different age and from different parties.

Keywords: gender, parliament, CLARIN

1. Introduction

This paper presents a pilot analysis of gender differences in the revised transcripts of speeches from the Danish Parliament over a time span of eight years. The active participation of women in politics is historically relatively new, and women are still underrepresented in Parliament. Furthermore, researchers have found gender specific differences in e.g. the subjects addressed by female and male members of Parliament (MPs) and their speech frequency. This is also the case in Sweden (Bäck et al., 2014) which is a country culturally near to Denmark.

Our study follows this line of research in that we aim to investigate whether it is possible to determine gender differences in the Danish parliamentary speeches by analysing the speeches with respect to their duration as well as the number of the speakers, their age, party and role in the party.

The paper is organized as follows. First, we account for relevant studies that indicate that there are differences in the frequency and subject of female and male MPs. Then, we describe the Danish data and we present a preliminary analysis of the data and a discussion of some of the gender differences that can be seen in them. Finally, we conclude and shortly describe future work.

2. Gender Studies in Political Speeches

The analysis of gender differences in political speeches has involved many aspects, not only the intrinsic characteristics of the political speeches comprising lexical, morphological, syntactic, semantic, pragmatic and rhetorical aspects, but also factors including culture, society, identity definition by the politicians and the media.

Paxton et al. (2007) report different studies on the political participation of women in different countries showing that the percentage of female parliament members is generally low, but it varies by country, being highest in Scandinavia (38% in 2005) and lowest in the Middle East (8% in 2005).

Sivrić and Jurčić (2014) analyse six political speeches in which six Croatian and US male and female politicians address a large audience. They find no differences between the two gender's discourse at the syntactic and semantic level, while they notice that the three female politicians have a tendency to use more implicit meaning and more contrasting constructions than the male politicians. They conclude that some of the language characteristics traditionally attributed to male and female discourse can be found in the political speeches produced by both genders and it depends on how politicians choose to show a particular identity. For example, Hilary Clinton wants to appear strong and thus her discourse has characteristics which usually are described as masculine.

Bäck et al. (2014) investigate the role of gender in the number of speeches held by Swedish politicians in the Parliament. They find that female politicians talk less than male politicians. Furthermore, they investigate other factors than speech frequency such as the domain of expertise, the position in the party of the politicians and their personal background. The data show that women talk as much as men in cases of what they define as "softer" policy areas for which they also often are the responsible in the party. Softer areas comprise culture, education, health and social issues, while harder technical areas comprise energy, finance, macro economy, foreign affairs, national security and transport.

The statistical analysis of gender differences in various types of discourse has shown that there are small, but consistent differences in language use (Newman et al., 2008). The features were found with the text analysis program, Linguistic Inquiry and Word Count, LIWC (Pennbaker et al., 2001) which analyses texts according to 74 linguistic categories. The only consistent effects of gender through the various text types were long words, swear words, articles, pronouns, and social words. Moreover, numerous features are often used in automatic gender identification in different types of data such as parts of the general language British National Corpus (Koppel et al., 2002), scientific articles (Vogel and Jurafskyi, 2012) transcribed telephone conversations (Boulis and Ostendorf, 2005) and computer-mediated written communication, e.g. (Corney et al., 2002; Ljubešić

et al., 2017). Gender identification has been part of the tasks under the Author profiling events organized by PAN[1] where many of the efforts deal with social media data, such as blogs and tweets (Martinc et al., 2017).

The automatic identification of the gender, age (young/old) and political affiliation (left or right wing) of politicians from the edited transcriptions of speeches held by politicians in the Swedish parliament between 2003 and 2010 is addressed by Dahllöf (2012). He extracted the initial 200 words of the speeches and selected the excerpts so that for each task he had at least 77 speakers per group and each speaker had at least 21 contributions. He used features from information extraction and balanced training and test data. Words characterizing each class were used as binary features according to information gain theory. Thus no linguistic analysis of the speeches was performed. The classifier was a support vector machine and ten-fold cross-validation was the evaluation method. Accuracy for gender identification was higher for right-wing politicians than for left-wing ones and was in the range 72.8-80.1%. The fact that short excerpts were used as data restricted the number of possible features to be used for classification.

3. Description of the data

The Danish Parliament Corpus 2009 – 2017 consists of Hansards (transcripts of parliamentary speeches) from the sittings in the Chamber of the Danish Parliament, *Folketinget*. The first editions of these official reports are published immediately after the meetings as pdf-files on the website of the Parliament[2] where it is also possible to see video recordings from the meetings. The final proofread editions are published up till one year after the first ones.

The corpus consists of xml files dumped for us from the Parliament's database by the IT department at the Danish Parliament. Each file contains speeches from one parliamentary year, running from October to June. The xml files do not refer to any xml scheme.

Year	Chairman	Male	Female	*Total*
2009-2010	189,749	3,836,276	2,579,952	*6,416,228*
2010-2011	136,830	2,939,173	1,783,022	*4,722,195*
2011-2012	125,314	3,330,297	1,492,207	*4,822,504*
2012-2013	127,073	3,119,049	1,596,624	*4,715,673*
2013-2014	118,266	2,577,144	1,414,419	*3,991,563*
2014-2015	132,718	2,640,560	1,368,437	*4,008,997*
2015-2016	191,275	3,696,753	1,606,241	*5,302,994*
2016-2017	147,393	2,964,842	1,553,551	*4,518,393*
Total	1,168,618	25,104,0944	13,394,4533	***39,667,1655***

Table 1: The total number of words[3] in the corpus

The files are marked for meetings, speeches, name of speaker, party of speakers and timing of the speeches. A speech is here defined as a single intervention by a MP; it can be a question as well as a longer debate contribution. The date of birth/age and the gender of the speakers for this study have been extracted from additional sources on the web site of the Parliament[4]. The consistent mark-up makes the data a very rich source for sociological and linguistic analysis.

3.1 License and Accessibility

The Danish Parliament Corpus 2009-2017 follows the license for Open Data[5] stating:

The Danish Parliament grants a world-wide, free, non-exclusive and otherwise unrestricted right of use of the data in the Danish Parliament's open data catalogue. The data can be freely:

- copied, distributed and published,
- adapted and combined with other material,
- exploited commercially and non-commercially.

Following the copyright act the speeches can be distributed without the consent of the speaker but only in a way where the author/speaker of each text/speech is clearly stated. Furthermore, the Danish Parliament must be acknowledged as the source. To our understanding this correlates to CLARIN PUB BY or CC-BY.

It is the aim to share the corpus in the CLARIN community as soon as a new repository system is implemented in CLARIN.dk. The version of the corpus used for this study includes meetings until May 4th, 2017, and the reports for the latest parliamentary year have not been published as the final edition. The reports of all other meetings are the final editions. The first version of the Danish Parliament Corpus 2009-2017 will be shared at CLARIN.dk in the same format we received it from the Parliament. Enriched versions of the corpus will be shared later on. For reproduction of the results in this study, the current version of the data will be available upon request to the authors.

3.2 Corpus language characteristics

According to the *Office of the Folketing Hansard* the reports are verbatim (exact transcripts of the speeches), but slightly edited following the guidelines:

- The spoken language is adapted into a colloquial and syntactically coherent written language with a liberal approach to what is deemed correct language.
- The editing is done carefully to ensure that the intentions of the speaker are clear.
- Factual errors and slips of the tongue are corrected.
- The appropriate formal requirement rules are observed.

In this process punctuation marks are added, and smaller corrections are made to make the speeches compliant to Danish syntax for written language, e.g. pauses and hesitations are omitted. Therefore stylistic analysis and investigations which include factors such as "sentence" length must take into account that spoken language has been artificially converted to written language.

[1] http://pan.webis.de/
[2] http://www.ft.dk/
[3] The corpus is not tokenized, therefore a word is here defined as a sequence of characters delimited by white space.
[4] http://www.ft.dk/da/medlemmerhttp://www.ft.dk/da/medlemmer
[5] http://www.ft.dk/~/media/sites/ft/pdf/dokumenter/aabne-data/conditions_for_use_-of_the_danish_parliaments-_open_data.ashx?la=da

The Standing Orders of the Danish Parliament state rules for the speeches, which the Speaker (the chairman of the Parliament) enforces during the debates. According to the Standing Orders, the MPs in the Danish Parliament are not allowed to applause or express disapproval during the debates, and when debating or asking questions the MPs must be addressed "hr."(Mr.) or "fru" (Ms.) and their full names, while the ministers must be addressed using the minister title. Furthermore, the informal pronoun "du" (you) should not be used. These rules impact the language, which becomes more polite and respectful than the language used in e.g. interviews or TV debates. One could say that the language is solemn and very formal compared to spontaneous speech.

4. A pilot gender study in the Danish Parliament Corpus

The aim of the present pilot study is to provide an overview of the distributional figures in relation to male and female members of the Danish Parliament. We want to test if there are non-linguistic parameters that might be useful in determining gender differences and in automatic gender classification.

The corpus used in the study is an extract of The Danish Parliament Corpus 2009 – 2017. All comments from the Speaker (chairman or -woman) are omitted in this study, since these will affect the general gender figures.

4.1 Corpus figures in a gender perspective

The figures shown below reflect the fact that we are looking for clues to characterize the speeches of female and male MPs, focusing on how the female MPs are represented in relation to their male colleagues.

Election	**Male**	**Female**	***Total***	**% Female**
Election 2007	113	66	*179*	36.9
Election 2011	111	68	*179*	38.0
Election 2015	113	66	*179*	36.9
Avg.	**112**	**67**	***179***	**37.24**

Table 2: Election results from the period covered by the corpus[6]

In compliance with the observations of Paxton et al. (2007), the number of elected female members is 37-38% and quite stable over the years.

Speeches	**Male**	**Female**	***Total***	**% Female**
Speeches	119,441	62,751	*182,192*	34.4
Speaking time, hours	2484.59	1324.78	*3809.37*	34.8
Words	25,104,094	13,394,453	*38,498,547*	34.8

Table 3: The total number of speeches' time and words in the corpus

We also investigated whether there is a statistically significant difference in the time the female and male MPs speak in the corpus (the MPs who actually speak) and the difference is significant (two-tailed unpaired t-test, df =1768, t = 2.4194 and p < 0.0156. Applying Welsh version of t-test, df = 1524, t = 2.486 and p = 0.13).

Table 4 compares the number of male and female MPs in the Danish Parliament in the corpus with respect to their age.

Age / MPs	**Male**	**Female**	***Total***	**% Female**
20-29	18	23	*41*	56.1
30-39	62	57	*119*	47.9
40-49	86	64	*150*	42.7
50-59	87	37	*124*	29.8
60-69	64	21	*85*	24.7
70-79	12	6	*18*	33.3
Total	**329**	**208**	***537***	38.7

Table 4: Members grouped by age

Table 4 shows that young women are very strongly represented in the Danish parliament in the period 2009 - 2017, and that women over 50, on the other hand, are very poorly represented.[7]

In Table 5 the number of speeches in the different age groups are given.

Age /speech	**Male**	**Female**	***Total***	**% Female**
20-29	4,839	5279	*10,118*	52.2
30-39	27,813	24,666	*52,479*	47.0
40-49	37,923	20,755	*58,678*	35.4
50-59	26,771	7,536	*34,307*	22.0
60-69	19,804	3,497	*23,301*	15.0
70-79	2,291	1,018	*3,309*	30.8
Total	**119,441**	**62,751**	***182,192***	34.4

Table 5: Speeches grouped by age

The table shows that women in general held fewer speeches than it would be expected from their number in the parliament. This is especially the case for female MPs in the age group 50-69.

In Table 6, the distribution of male and female MPs with respect to their party is given. Left-wing parties are marked with red, while right-wing parties are marked with blue. The central party *The Social Liberal Party* (marked with the letter RV) is in white.

MPs /party:	**Male**	**Female**	***Total***	**% Female**
EL	19	11	*30*	36.7
SF	21	24	*45*	53.3
ALT	6	4	*10*	40.0
S	87	50	*137*	36.5
Left-wing				**40.1**
RV	3	5	*8*	62.5
KF	4	2	*6*	33.3
V	89	38	*127*	29.9
DF	51	31	*82*	37.8
LA	19	8	*27*	29.6
Right-wing				**32.6**

Table 6: MPs in right and left-wing parties

[6] Extracted from https://www.dst.dk/ and http://www.ft.dk

[7] The % difference in Table 1 and 3 are caused by members stopping, taking leave of absence and new taking over. The substitutes are not necessary of the same gender.

The proportion of female MPs in the left-wing parties is higher than in the right-wing parties, and especially in *The Socialist People's Party* (SF) [8].

In Table 7, we show the number of speeches held by the MPs of the different parties.

The table indicates that the left-wing female MPs give fewer speeches than the female MPs from the right-wing-parties compared to their seats in Parliament. The women who speak less frequently compared to their seats in Parliament are those who belong to the most left-wing party, *The Red-Green Alliance* (EL) who have 36.7% seats but only give 27.7% speeches compared to their male colleagues. The female MPs of *The Social Liberal Party* (RV) give 54.5% speeches vs. 62.5% seats and the right-wing party *The Liberal Party* (V) give 27.2% vs. 33.3 % seats.

Speeches/ party:	**Male**	**Female**	***Total***	**% Female**
EL	15,872	5,916	*21,788*	27.2
SF	7,344	7,967	*15,311*	52.0
ALT	2,276	1,508	*3,784*	39.9
S	24,684	13,537	*38,221*	35.4
Left-wing				**36.6**
RV	5,986	7,180	*13,166*	54.5
V	26,949	11,968	*38,917*	27.2
KF	8,528	4,495	*13,023*	52.0
DF	17,839	6,504	*24,343*	39.9
LA	8,738	2,802	*11,540*	35.4
Right-wing				**33.8**

Table 7: Speeches held by MP in right and left-wing parties

The final observation leads us to investigate the relationship between the number of speeches and the role of the MPs.

In the Danish Parliament most of the bills are introduced by ministers but to some extent also by spokespersons from parties outside the government. When a bill is debated, the proposer - typically the minister - gives the first speech, followed by the spokespersons from the other parties who give their opinion or ask questions to the proposer. Therefore, the role as minister or spokesperson gives more opportunities to speak.

In Table 8 the number of male and female ministers from the various parties is shown.

MPs/ ministers:	**Male**	**Female**	***Total***	**% Female**
SF	4	2	*6*	33.3
S	12	7	*19*	36.8
Left-wing				**35,1**
RV	6	4	*10*	40
V	14	10	*24*	41.7
KF	5	6	*11*	54.55
LA	3	3	*6*	50
Right-wing				**48.8**

Table 8: Ministers from the different parties

[8] See all parties: http://www.thedanishparliament.dk/Members/Members_in_party_groups.aspx

Compared to the number of elected female MPs, relatively few left-wing female MPs have been ministers. The left – wing won one election while the right-wing won two elections in the period covered by the corpus (2007 and 2015).

In Table 9, we show the number of speeches held by female ministers belonging to the various parties. Although the right-wing parties have the biggest number and percentage of female ministers, these ministers do not speak as many times as their female colleagues from the left-wing parties. Especially the female MPs from *The Social Democratic Party* (S) speak much more than expected. This might be related to the period in which there was a female prime minister, which will be investigated in the following section.

Speeches/ ministers:	**Male**	**Female**	***Total***	**% Female**
SF	1,148	345	*1,493*	23.1
S	5,249	5,202	*10,451*	49.8
Left-wing				**36.5**
RV	1,707	1,002	*2,709*	37,0
V	8,444	5,065	*13,509*	37.5
KF	2,214	1,553	*3,767*	41.2
LA	777	488	*1,265*	38.6
Right-wing				**36.4**

Table 9: Speeches by the ministers

4.2 Prime ministers, ministers and speeches

We have further analysed two election periods: 2011-2015 with a female prime minister and 2015-2017 with a male prime minister. Although the first period is longer than the second, we can compare the percentage distribution of ministers and the duration of their speeches.

Table 10 contains the number of ministers and of MPs in the entire corpus as well as the hours they have spoken and the number of speeches they have held. The same data for the two election periods with a female and a male prime minister, respectively, are given in Tables 11 and 12.

Entire corpus (2009-2017)	**Male**	**Female**	***Total***	**% Female**
Ministers	44	32	*76*	42.1
MPs elected	112	67	*179*	37.4
Ministers, hours	460.05	317.58	*777.63*	40.8
Other MPs, hours	2024.54	1007.20	*3031.74*	33.2
Speeches, ministers	18,762	13,167	*31,929*	41.3
Speeches, other MPs	100,679	49,584	*150,263*	33.0

Table 10: Number of ministers, speaking time, and number of speeches in the entire corpus

Female prime minister (2011-2015)	**Male**	**Female**	**Total**	**% Female**
Ministers	23	14	*37*	37.8
MPs elected	111	68	*179*	38.0
Ministers, hours	195.76	158.47	*354.23*	44.7
Other MPs, hours	944.23	414.28	*1358.51*	30.5
Speeches, ministers	8,104	6,592	*14,696*	44.9
Speeches, other MPs	45,558	20,407	*65,965*	30.9

Table 11: Number of ministers, speaking time, and number of speeches under a female prime minister

Male prime minister (2015-2017)	**Male**	**Female**	**Total**	**% Female**
Ministers	16	10	*26*	38.5
MPs elected	113	66	*179*	36.9
Ministers, hours	128.39	55.21	*183.60*	30.1
Other MPs, hours	535.39	271.69	*807.28*	33.7
Speeches, ministers	5,942	2,658	*8,600*	30.9
Speeches, other MPs	29,138	13,802	*42,940*	32.1

Table 12: Number of ministers, speaking time, and number of speeches under a male prime minister

The comparison of Tables 11 and 12 shows a clear tendency: although the percentage of female ministers under the male prime minister was higher, the female ministers under the female prime minister spoke much more and for a longer time.

As there has only been one female prime minister in Denmark, we only have data from one election period, but the comparison of data from that election period with data from the current election period shows that the female ministers under a female prime minister speak more than the female ministers under a male prime minister.

In the next section we will consider the speaking time of all MPs, focusing on the members of the standing committees.

4.3 Committee members and subject areas

Not only the ministers play a central role in the parliament, so do the members of the standing committees who are spokespersons for the parties they represent.

In general, 62.7% of the male MPs and 67.1% of the female MPs are spokesmen in one or more committees in the current election period. Since the information about this role is not present in the corpus, it is very difficult to extract the exact distribution of speaking time for each committee policy area. It is however possible to extract the total amount of speaking time for the members of the committees.

Table 13 shows the amount of speaking time for male and female ministers, spokesmen and non-spokesmen MPs in the election period 2015-2017.

Speaking time (2015 – 2017)	**Male**	**Female**	**Total**	**% Female**
Ministers, hours	128.39	55.21	183.60	30.1
Spokesmen, hours	456.69	226.15	*682.84*	33.1
Other MPs, hours	78.70	45.54	*124.24*	36.7

Table 13: The total amount of speaking time for all MPs in the period 2015 – 2017

Table 13 shows that the majority of the debates are done by the spokespersons, followed by the ministers. The regular MPs do not have much speaking time. This follows to a great extent the Standing Orders of the Danish Parliament which states that the time allotted to speakers in general debates is a maximum of 20 minutes for ministers, a maximum of 10 minutes for spokesmen and a maximum of 5 minutes for other members.

The Danish parliament has 25 standing committees. These have been summed up to the 13 groups listed below, where e.g. *Foreign affairs* also comprises development aid, *Agriculture* comprises food, fishing and agriculture and *Social* comprises health, children, impaired and elderly people.

Since the corpus does not contain information about the spokesman role, we had to retrieve this information from the parliament website. For the same reason, we only have access to information about the present sitting and therefore only take into account the period 2015 -2017. In Table 14, the number of male and female spokesmen in the various policy areas is given. The table is sorted based on the fraction of female speakers.

Policy areas	**Male**	**Female**	**Total**	**% Female**
Foreign affairs	25	4	29	13.8
Domestic affairs	16	4	20	20.0
Labour and industry	12	3	15	20.0
Economy	13	4	17	23.5
Infrastructure	25	8	33	24.2
Defence	9	3	12	25.0
Food and agriculture	14	5	19	26.3
Church and culture	10	6	16	37.5
Environment	10	7	17	41.2
Law	7	5	12	41.7
Immigration	7	8	15	53.3
Social	23	28	51	54.9
Education	8	12	20	60.0

Table 14: The percentage of female spokespersons of the standing committees in the period 2015 - 2017

The data in the table clearly show that women are strongly represented in the "softer" subject areas, like *immigration*, *social* and *education*, whereas the more technical areas are dominated by men. The observations support the findings of Bäck et al. (2014).

5. Conclusion

In this paper, we have presented a first study of the participation of female and male members of the Danish Parliament in the period 2009-2017 by looking at their number, their age, their party, and their role comparing these factors to the frequency and duration of the various politicians' speeches. More specifically, our study shows that the number of female MPs under 29 is larger than the number of male MPs from the same age group and that in general women speak less frequently and for a shorter time than male MPs in proportion to their seats in Parliament. The difference in speaking time between female and male MPs is statistically significant. The data also show that women belonging to a left-wing party speak less frequently than women from the right-left party compared to their seats in Parliament. The data also indicate that ministers and spokesmen speak more frequently than simple MPs and that female ministers

under a male prime minister give fewer speeches than female ministers under a female prime minister even though their percentages in the two periods are similar.

We also found that there were relatively more male spokesmen than female ones in the period covered by the corpus. The Danish data seem to confirm the findings of Bäck et al. (2014) in the Swedish data that female MPs often spoke about "softer" political areas for which they were responsible, while male MPs spoke about "harder" subjects.

In the future, we will combine the various features we have looked at in this study, and calculate their statistical significance. We have started to do an automatic extraction of the subject areas addressed by the MPs in order to automatically determine i.a. differences in the speeches with respect to the gender, party, role and age of the speaker. Moreover, we are looking at linguistic features that can be used to determine the gender, age and role of MPs automatically in line with the work of e.g Pennbaker et al. (2001) and Dahllöf (2012). Furthermore, since both audio- and video-recordings are available for the most recent Parliament debates, we will also address these multimodal data (speech and gestures) in the future.

6. Acknowledgements

The work reported here has received funding through CLARIN.DK, which is supported by the grant "INFR 2011: Digitalt Humaniora Laboratorium (DigHumLab)" from the Danish Ministry of Higher Education and Science.
We would like to thank Hanne Fersøe for her comments, Anders Gilbro Nielsen from the IT department at the Danish Parliament, Folketinget, for being very helpful in providing the data to us, and Anne Jensen who pointed out this opportunity for us, and with whom we were looking much forward to doing this work. R.I.P. Anne.

7. References

Bäck, H., Debus, M. and Müller, J. (2014). Who Takes the Parliamentary Floor? The Role of Gender in Speech-making in the Swedish Riksdag. *Political Research Quarterly* 67(3), pp. 504–18.

Boulis, C. and Ostendorf, M. A. (2005). Quantitative Analysis of Lexical Differences Between Genders in Telephone Conversations. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* , 2005 New Brunswick, NJ, Association for Computational Linguistics, pp. 435-442.

Corney, M., de Vel, O., Anderson, A. and Mohay G. (2002). Gender-preferential Text Mining of E-mail Discourse. In *Proceedings of the 18th Annual Computer Security Applications Conference*, Alamitos, IEEE Computer Society, pp. 282-289.

Dahllöf, M. (2012). Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches - A comparative study of classifiability. *Literary and Linguistic Computing*, Volume 27, Issue 2, 1 June 2012, pp. 139-153.

Koppel, M., Argamon, S. and Shimoni, A. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17(4), 401–412.

Ljubešić, N., Fišer, D. and Erjavec, T. (2017). Language-independent Gender Prediction on Twitter. In *Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science*, pp. 1–6.

Martinc, M., Škrjanec, I., Zupan, K. and Pollak, S. (2017). Author Profiling - Gender and Language Variety Prediction—*Notebook for PAN at CLEF 2017*. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, pp. 11-14.

Newman, M. L., Groom, C. J., Handelman, L. D. and Pennebaker J. W. (2008). Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes*. 45 (3), pp. 211-236.

Paxton, P., Kunovich, S. and Hughes, M. M. (2007). Gender in Politics. In *Annual Review of Sociology*, Vol. 33, pp. 263–284.

Pennebaker, J., Francis, M. E. and Booth, R. J. (2001). Linguistic Inquiry and Word Count (LIWC): LIWC2001.

Sivrić, M. and Jurčić, D. (2014). Gender Differences in Political Discourse. In *Journal of Foreign Language Teaching and Applied Linguistics*, pp. 173-185.

Vogel, A. and Jurafsky, D. (2012). He said, she said: Gender in the ACL anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pp. 33–41.

The Parliamentary Debates as a Resource for the Textometric Study of the French Political Discourse

Sascha Diwersy, Francesca Frontini, Giancarlo Luxardo
PRAXILING UMR 5267 Univ Paul Valéry Montpellier 3 & CNRS - Montpellier, France
name.surname@univ.montp3.fr

Abstract

This paper reports on a corpus collecting together the French parliamentary debates in plenary sitting. It outlines the design and data format of the samples and presents various usage scenarios related to their textometric use.

Keywords: political discourse, parliamentary corpora, metadata

1. Introduction

This contribution presents a corpus that contains the transcriptions of French parliamentary debates; we also discuss the possibilities of its exploration from a textometric perspective.

The paper is structured as follows. Section 2 outlines the rationale behind the creation of the corpus, its composition and the concepts involved in its design. Section 3 describes how the corpus can be processed to be implemented in a publishing platform. In section 4, we then introduce some key elements of a textometric methodology and illustrate the exploration of the corpus by giving brief sketches of corresponding usage scenarios. Section 5 provides a summary and discusses possible directions for future developments of the resource presented in this paper.

2. The TAPS-fr corpus

2.1. Rationale and composition of TAPS-fr

The *Assemblée Nationale* publishes on an open access basis[1] a number of datasets, and among them its debates in plenary sitting, dating back to 2013 and provided as a regularly updated data dump at `http://data.assemblee-nationale.fr/travaux-parlementaires/debats`.[2].

The textual data have been processed and assembled according to a methodology discussed in the following and called *Transcription and Annotation of Parliamentary Speech* (TAPS), with the aim of offering a large-scale resource to researchers working on French political discourse, especially from a data-driven perspective. The methodology is kept as generic as possible, in order to be reused for debates of additional parliaments, possibly in other languages.

We call the corpus described here TAPS-fr. It is primarily designed to provide a methodological support for investigations in the French tradition of textometry (French: *textométrie*), which integrates both searches based on full-text indexing and multivariate exploratory data analysis (Lebart et al., 1998). The open data publishing of the French parliamentary debates is part of a trend known as Open Government Data and described with the eight principles defined by the Sebastopol meeting held in 2007[3]. One of the challenges for the projects initiated in this trend is set by the fact that these open data, while published in large amounts and accessible with a relative ease of reuse, are "raw data": little is known about the conditions of their production (Plancq, 2016).

Legislature	Period	Nr of sessions	Nr of words
14	05/13-12/13	152	5,200 K
14	01/14-02/17	873	28,600 K
15	06/17-12/17	156	4,700 K
Total			38,500 K

Table 1: Composition of the TAPS-fr

We subdivided the TAPS-fr corpus into three subcorpora described by Table 1:

1. The first months (May 2013 - December 2013) represent a small subcorpus, which was not processed in depth so far (the source webpage states that the debates were fully transcribed only from October 2013).

2. The second subcorpus was the one mostly used for our experiments: it comprises the debates of the last months of the 14th legislature (January 2014 - February 2017).

3. A third corpus includes the debates of the 15th legislature up to the end of December 2017.

2.2. A machine-processable format geared to multiple needs

We distinguish four formats handled for the processing of TAPS-fr, all of them being XML-based:

1. The source format: it is the format used by the raw data, which is subdivided in three components (actors, bodies - *organes* - and sittings); the text is included in the sittings section and refers to actors (members of

[1] The "Licence ouverte / Open Licence" is a free licence created by the French governmental mission Etalab.

[2] A selection of parliamentary records from the *Assemblée Nationale* has already been collected and published in TEI format in (Truan, 2017) as part of a broader project on perceptions of the other in various European countries.

[3] https://opengovdata.org/

parliament, members of government, etc.) belonging to various bodies (e.g. parliamentary commissions).

2. The TAPS format: it is the result of a conversion applied to the source data, in order to extract the relevant metadata and annotation useful to our applications (see below).
3. The XML-TXM format is a customization of the TEI data model used by the TXM software as a pivot format in order to present semantic and editorial annotations.
4. The CWB format (defined by the IMS Open Corpus Workbench) encapsulates lexical and syntactic annotations and is used by a search engine, allowing text retrieval based on queries expressed in the CQP (Corpus Query Processor) syntax. This is a compound format with XML tags and token records appearing on separate lines (one surface form is associated to tab-delimited token-level annotations).

The descriptions of the TXM (for "textometry") platform (Lavrentiev et al., 2013) and of the CWB environment (Evert and Hardie, 2011) are beyond the scope of the present document. However, the following two sections describe how they have been implemented in this project.

The TAPS format basically relies on the concepts of metadata and annotations defined in the TXM environment, which distinguishes structural units and lexical units.

Metadata (the association of a variable and a set of modalities) are used to partition the corpus, to create subcorpora and to retrieve the text. They are defined on various structural units: XML elements, text segments, paragraphs, sentences, and possibly other units defined by the user. Metadata are therefore viewed as properties of structural units.

Each processed lexical unit has several properties, such as word form, lemma and part-of-speech (grammatical category).

The conversion from the source format to the TAPS format creates a number of files, each one associated to a parliamentary sitting. The metadata that is extracted from the source format can be associated either globally to each file or to a single speech (the intervention of a person in the debate). The format of the files complies with the TEI data model, so that the metadata associated to a file are described in the TEI header. For the single speech, the `<u>` element (utterance) was chosen[4]: this element is originally defined by the TEI guidelines for the transcription of oral corpora, it is extended by the definition of a number of attributes relevant to our application (e.g.: role in the debate, nomination in the parliamentary structures, nomination in the government, political affiliation...). The assignment of attributes at the utterance level and within the TEI headers implies some redundancy, however it provides an easier reuse for the text retrieval. Table 2 specifies the major structural units defined within the data model of the TAPS-fr corpus.

[4]This approach was also adopted by (Truan, 2017) and by the authors of the SlovParl corpus (Pančur et al., 2017), whereas earlier versions of the latter opted for the `<sp>` element defined by the TEI module for encoding performance texts (cf. `https://github.com/SIstory/SlovParl`).

Structural Unit	Associated Metadata (descriptors)	XML Element
sitting	date-time, year, parliamentary term	`<text>`
speech	speaker name, speaker role, parliamentary group, speech type (debate, interruption, vote explanation, etc.)[5]	`<u>` (utterance)
paralinguistic event	description	`<incident>`
sentence[6]	–	`<s>`

Table 2: Main structural units encoded in the TAPS-fr corpus

Starting from the TAPS format, the TXM environment performs several steps of conversion and generates files in the XML-TXM format as well as in the CWB format, which, in both cases, can include linguistic annotations added to the lexical units. While TXM's import modules allow for automatic morphosyntactic tagging and lemmatisation by means of TreeTagger[7] (Schmid, 1994), it is possible to pre-process the corpus data outside the platform by using other NLP toolkits. In our specific case, we chose the freely available processing pipeline Bonsai[8] (Candito et al., 2010b; Candito et al., 2010a) in order to add syntactic dependency annotations. The latter have been extended by several categories whose purpose is to optimize the processing of queries exploring the dependency relations annotated in the corpus. The categories in question are marked by an asterisk in Table 3, which outlines the overall data model of the word level annotations within the TAPS-fr corpus.

3. The publishing framework

The four formats described in section 2.2 enable to publish the TAPS-fr corpus in two different contexts: either within the TXM interface or in the TAPS format for the purpose of improving interoperability. Both options provide some conformance level with the TEI guidelines.

Unlike older software tools developed within the French community of "*analyse des données textuelles*" (textual data analysis), TXM was designed to support applications

[5]From a linguistic point of view, this descriptor, which is not included in the data model of the corpus provided by (Truan, 2017), is particularly important when it comes to differentiate effects of register variation ranging from highly formulaic to less formal speech (as in the case of e.g. interruptions). It should be noted that this metadata element can be easily retrieved from the raw data dump we used to build the TAPS-fr corpus, whereas it is only partially or not all available in the other source formats (HTML or pdf) used by the *Assemblée Nationale* to publish its minutes on-line.

[6]This unit is optional as it is only provided in the case of specific processing steps pertaining to linguistic annotation.

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

[8]`https://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html`

Lexical Property	Description
word	surface form or punctuation sign
lemma	lemma corresponding to the surface form
cpos	coarse grained part of speech (PoS)
pos	fine grained PoS (+ subcategorization)
feat	morphological features
deprel	syntactic function of the token in the dependency relation to its head
headword *	surface form of the syntactic head
headlemma *	lemma of the syntactic head
headcpos *	coarse grained PoS of the syntactic head
headpos *	fine grained PoS (+ subcategorization) of the syntactic head
headfeat *	morphological features of the syntactic head

Table 3: Linguistic annotations at the word level

for textual criticism. The ability to produce a critical edition of a historical source, typically a "synoptic edition". i.e. a formatted output presented alongside a facsimile (e.g. a manuscript), is provided by the use of the TEI guidelines.

The TXM conformance to TEI is implemented through the use of the TXM pivot format (XML-TXM[9]), which is basically derived from the need to generate a document to be rendered in a browser (with specific layout directives), while allowing the navigation in the text through the support of the TEI `<w>` element (which encapsulates morphosyntactic and other annotations of the words).

In the context of TAPS-fr, as a parliamentary corpus, the basic requirement is to provide a single edition allowing navigation in a browser and keeping the editorial annotations made by the transcribers. Possible extensions would be to provide multiple editions including: the translation to a different language (but this usage would be unlikely in the case of the *Assemblée Nationale*) or links to audio or video recording.

The TAPS-fr corpus is available from the textometry portal of the Praxiling laboratory [10], either directly browsable online in the TXM environment or as a downloadable resource (with the possibility to process it offline in the desktop version of the TXM software).

In order to comply with the TEI model, the TAPS sitting files (already mentioned above) contain a `TEIheader`: apart from the information related to the publication conditions (`<fileDesc>`), this header also describes the date of the sitting and the speakers involved (`<creation>` element in `<profileDesc>`).

[9]In addition to the generic XML format, TXM also integrates TEI with an import module, called TXM-XTZ (XML TEI Zero), which is able to interpret the semantics associated to a minimal set of TEI elements, through the application of XSL stylesheets.

[10]http://textometrie.univ-montp3.fr/

4. The analytical framework

In this section we will briefly illustrate the application of two standard methods in textometry - correspondence analysis (CA) and the identification of characteristic items by frequency specificities - to the TAPS-fr corpus. Correspondence analysis (cf. (Benzécri, 1973), (Lebart et al., 1998, 45ssq)) is a useful technique in providing a condensed view of divergences relating to samples (resulting from a partition in the corpus) and lexical items.

We illustrate this by means of a plot generated on the basis of a CA (Figure 1)[11] performed on the speeches in the second subcorpus (cf. section 2) using the political group of each speaker (excluding the sitting presidents and the members of government) as differentiating variable. It is possible to observe that the first (horizontal) axis opposes the right-wing groups (UMP-LR, UDI), which have negative coordinates, to the left-wing groups (SRC-SER, Écolo, RRDP, GDR), which are located on the positive side, whereas on the second (vertical) axis, the socialist group (SRC-SER), which forms the major part of the government majority during that period, stands in contrast to the group of left-wing opposition parties GDR.

An efficient way to single out the lexical (and grammatical) items implicated in the opposition of extralinguistic factors highlighted by CA is the computation of frequency specificities based on the hypergeometric distribution (Lafon, 1980), a lexico-statistical approach similar to the keyword analysis used in the British tradition of corpus linguistics (cf. amongst others (Rayson, 2003)). Figure 2 highlights some of the nouns that are more specific to the discourse of the right-wing parliamentary group *UMP-LR*. These include nouns referring to the nation (*Français* 'French') and other classic elements of conservative ideology both in social (e.g. the series *famille* 'family', *parent* 'parent', *enfant* 'child') and economic terms (e.g. nouns designating learned professions such as *médecin* 'doctor, physician', *notaire* 'notary, solicitor').

5. Conclusion and outlook

In this paper, we introduced TAPS-fr, a corpus of debates from the *Assemblée Nationale* by giving a brief sketch of the methodology underlying its creation, of its data model and of some application scenarios that illustrate the exploitation of this resource within the analytical framework of textomtry. The corpus, whose preliminary version is now accessible at `http://textometrie.univ-montp3.fr/`, will be published in its stable version in the Ortolang[12] CLARIN repository for long term preservation, under the CC BY-NC 4.0 license. As the TAPS-fr is meant to be a monitor corpus[13], it will continually be expanded on the basis of the regularly updated raw

[11]The CA plot was generated by means of the R packages FactoMineR (Husson et al., 2013) and explor (Barnier, 2017). We have chosen these packages instead of TXM's CA command because they allow for a more flexible manipulation of the graphical output. The axis descriptions indicated by the horizontal and vertical arrows have been added in a post-processing step.

[12]`https://www.ortolang.fr/`

[13]For the notion of monitor corpus see amongst others (McEnery and Hardie, 2011, 6sq.).

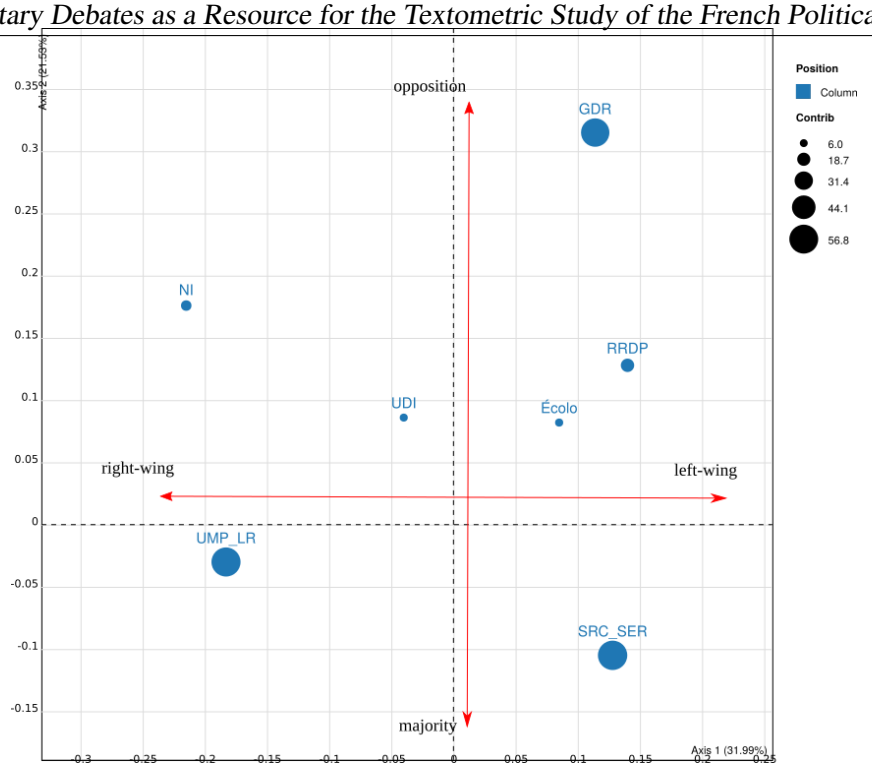


Figure 1: CA plot based on a partition by political group (word-rows are not displayed). Point sizes indicate contribution.

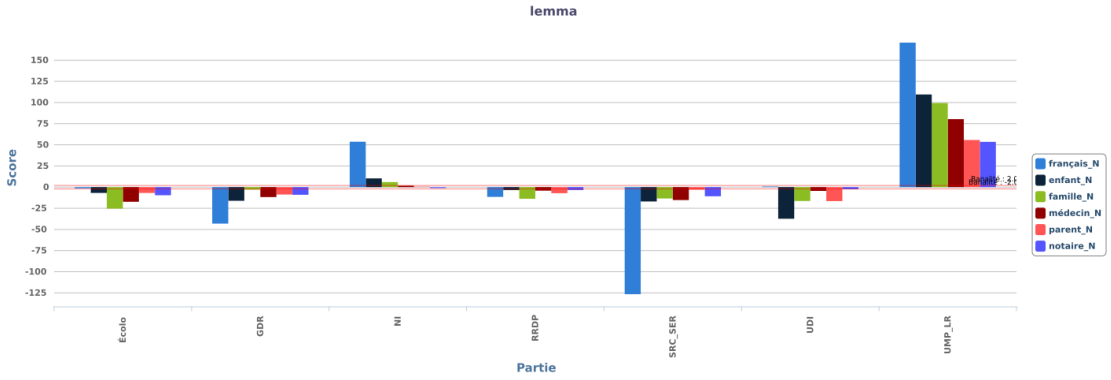


Figure 2: Most characteristic nouns specific to the discourse of the parliamentary group UMP-LR

dump provided by the *Assemblée Nationale*'s open data service, but we also intend to include successively samples of parliamentary sessions prior to 2013. In the latter case, this implies the necessity to adapt our current approach to the processing of material coming from various source formats (primarily HTML and XML with different schema specifications) with varying granularity of directly retrievable information, which might lead to slight revisions of the data model presented in this paper. The future stabilization of our methodology could lay ground not only to the continuous construction of a resource providing broad coverage of the debates at the French *Assemblée Nationale*, but also to the project of creating an extended textual base, which by integrating the plenary sessions of the *Sénat*, the second chamber of the French Parliament, would constitute a large scale corpus of institutional and political discourse in contemporary France at the national level.

6. Bibliographical References

Barnier, J., (2017). *explor: Interactive Interfaces for Results Exploration*. R package version 0.3.3.

Benzécri, J.-P. et collaborateurs. (1973). *L'analyse des données : L'analyse des correspondances*, volume 2. Dunod Paris.

Candito, M., Crabbé, B., and Denis, P. (2010a). Statistical French dependency parsing: treebank conversion

and first results. In *Seventh International Conference on Language Resources and Evaluation - LREC 2010*, pages 1840–1847, La Valletta, Malta, May. European Language Resources Association (ELRA).

Candito, M., Nivre, J., Denis, P., and Henestroza Anguiano, E. (2010b). Benchmarking of Statistical Dependency Parsers for French. In *23rd International Conference on Computational Linguistics - COLING 2010*, pages 108–116, Beijing, China, August. Coling 2010 Organizing Committee.

Evert, S. and Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference, Birmingham*, Birmingham, UK.

Husson, F., Josse, J., Lê, S., and Mazet, J. (2013). FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R. Technical report.

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1(1):127–165.

Lavrentiev, A., Heiden, S., and Decorde, M. (2013). Analyzing TEI encoded texts with the TXM platform. October.

Lebart, L., Salem, A., and Berry, L. (1998). *Exploring Textual Data*, volume 4 of *Text, Speech and Language Technology*. Springer Netherlands, Dordrecht.

McEnery, T. and Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Pančur, A.,
v Sorn, M., and Erjavec, T. (2017). Slovenian parliamentary corpus SlovParl 2.0. Slovenian language resource repository CLARIN.SI.

Plancq, C. (2016). Utiliser les données ouvertes (open data). Un exemple avec les débats en séance publique à l'Assemblée nationale.

Rayson, P. (2003). *Matrix:a statistical method and software tool for linguistic analysis through corpus comparison*. phd, Lancaster University, February.

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

Truan, N. (2017). Débats parlementaires sur l'europe à l'assemblée nationale (2002-2012). ORTOLANG (Open Resources and TOols for LANGuage) –https://hdl.handle.net/11403/fr-parl/v1.

Using Data Packages to Ship Annotated Corpora of Parliamentary Protocols: The *GermaParl* R Package

Andreas Blätte
University of Duisburg-Essen
andreas.blaette@uni-due.de

Abstract

This paper suggests to disseminate linguistically annotated and indexed versions of corpora of parliamentary debates as R data packages. The *GermaParl* Corpus of Parliamentary Protocols serves as an example to illustrate the advantages of this approach. Keeping data in data packages offers established mechanisms to version and document data, and for ensuring the reproducibility of the data. The package may include further annotation layers, and functionality to analytically exploit additional annotations. Finally, sharing packages via a CRAN-like repository is a user-friendly way to make data available.

Keywords: Corpus (Creation, Annotation, etc.), Document Classification, Digital Humanities

1. Introduction

Corpora of parliamentary protocols are available for an increasing number of countries, and are gaining a strong standing as language resources. There are many advantages of corpora of parliamentary protocols for substantial research and methodological innovation. A variety of disciplines may benefit from these data. The largely unproblematic licensing conditions for plenary proceedings provide a particularly important reason, why it is worth to invest time and energy into this family of corpora: Corpora of parliamentary protocols are sustainable because the raw data is in the public domain.

It is a common problem that copyright law and restrictive licenses impede research; consider for instance the difficulty to share corpora of newspaper articles and social media content for reproduction. Corpora of plenary protocols can be prepared, stored, analyzed, and shared without fear.

Corpora of parliamentary protocols will typically be large corpora. Because of their size, these corpora call for computational support for exploiting the analytical potential of the data. They can be used to develop and test all kinds of procedures and algorithms. But size and technical data formats give rise to another restriction that may make the data exclusionary. Even without legal restrictions, many researchers who might use corpora of parliamentary protocols very productively, but who do not have a strong technical background, will have great difficulties to handle corpora of considerable size.

Hosting the data centrally on a server and offering a web interface is a reasonable solution to serve users who work with standardized analytical procedures. CQPweb (Hardie, 2012), SketchEngine (Kilgarriff et al., 2004; Kilgarriff et al., 2014) and NoSketchEngine[1] are powerful, server-based systems that deservedly have attracted communities of users and developers. But these systems are too restrictive for users who intend to explore all kinds of algorithms, and that approach data with an experimental stance. Because of security reasons, administrators will usually want the technically more ambitious users to run their experiments on their own local machines. And of course, a server accessed by many users will often be not the best place to conduct computationally expensive experiments.

The solution suggested here is to share annotated and indexed corpora as data packages. The R data package *GermaParl* serves as the showcase. One line of code will be enough to install a linguistically annotated and indexed version of *GermaParl* on your personal machine (or server). Moreover, the data package includes a detailed documentation how the corpus has been prepared, a vignette how it can be used, and custom functionality to create thematically defined subcorpora for specific research purposes. The suggested approach can serve the aims offering open data, making the preparation of the data reproducible and transparent and minimizing barriers of entry. Therefore, offering corpora of parliamentary protocols as data packages is the suggestion of this paper.

2. Versions of *GermaParl*

The *GermaParl* Corpus – the naming of the corpus is inspired by the *DutchParl* corpus (Marx and Schuth, 2010) – is a corpus of parliamentary protocols of the German Bundestag. The data has been consolidated for the years 1996 to 2016. The corpus covers the period for which txt files are publicly available. The data is explained and documented elsewhere (Blätte, 2018; Blätte and Blessing, 2018). Here, we focus on the data types that can be offered and shared *after* the initial preparation of the data.

- *XML (TEI standardization):* The basic variant of the corpus is an XMLification of the raw data (txt and pdf documents) that follows the standards of the Text Encoding Initiative (TEI)[2]. The XML version of the corpus is available at a GitHub repository.[3] GitHub has many advantages such as an accessible display of data,

[1] `https://nlp.fi.muni.cz/trac/noske`.

[2] See `www.tei-c.org`.

[3] See `https://github.com/PolMine/GermaParlTEI`. Of course, git has been designed to support the development of code, but its logic makes it suited very well for versioning corpora. The strongest argument against keeping corpora at GitHub is the usual size limitation of repositories to 1 GB. Moving to GitLab, or a self-hosted GitLab server are viable alternatives to GitHub.

an option to download the data and a system for managing issues and to manage user feedback. However, the XML/TEI variant of the corpus is not the data format to actually work with. A set of further processing steps is necessary.

- *Linguistically annotated corpus:* In a manner that is common in Natural Language Processing (NLP), the XML/TEI variant of the corpus is passed through a pipeline for linguistic annotation, using standard tools. The NLP tool currently used is *Stanford Core NLP*[4] The output of *Stanford Core NLP* (JSON, in this case) is transformed into a verticalized data format that can be imported into the *Corpus Worbench* (CWB)[5].
- *CWB indexed version:* Indexing and query engines are crucial to make corpora a useful resource for research. The *Corpus Workbench* (CWB) is one of the older systems. But it is still a powerful, mature system that is well-maintained. Due to the flexibility and power of the *Corpus Query Processor* (CQP) and its uncompromising open source orientation, it keeps being a good choice as an indexing and query engine for scientific purposes. The linguistically and structurally annotated data that has been prepared in step two is imported into the CWB, achieving a considerable data compression and a data format that researchers can work with efficiently and productively at the same time.

Technically advanced and experienced users may work efficiently with the XML/TEI version of the corpus. Students from the social sciences, early stage researchers, and newcomers to the *eHumanities* or the *computational social sciences* will find it difficult to make productive use of that kind of XML. Thus, the best way to offer the data, is to grant access to the CWB indexed variant of the corpus. That could be done with a server-based system (such as CQPweb), thus restricting the flexibility of users. Another approach is to grant access to the (zipped or tarred) data at some kind of online storage, so that users can download the corpus and install it themselves locally. The approach suggested here is to wrap the data in an R data package that may include a fully developed documentation and specialized functionality. This can be hosted without a lot of effort at a CRAN-style repository. Conventional R mechanisms make downloading and installing the package minimally demanding, if the package is prepared appropriately. The next section explains how this is implemented in the *GermaParl* data package.

3. Hosting and installing the *GermaParl* R Data Package

The size of the *GermaParl* data package (almost 1 GB) exceeds the size limitations for packages by the *Comprehensive R Archive Network* (CRAN) by far.[6] Of course, from the perspective of users, it would be ideal to be able to install an established text resource from CRAN. But offering an alternative is neither difficult for those offering data, nor difficult to handle from the perspective of users. Because it is easy, it is fairly common to host and administer a 'private" CRAN-style repositories. The package *miniCRAN* supports setting up a private CRAN-like repository in an enterprise setting[7], the package *drat* offers functionality to insert packages into a repository, and is specialized on using (or abusing) GitHub Pages to host a CRAN-like repository[8].

All that is necessary to host a CRAN-like repository is to mimick the directory structure of CRAN, and to register any new package that you put in the repository, so that some metadata is written to a file called 'PACKAGES'. The aforementioned packages *miniCRAN* and *drat* support that, but it can also be done 'manually'. The directory structure of CRAN-like repositories is designed to host the source tarballs of packages as well as binary versions for macOS and Windows. As data packages with corpus data will usually not include code that needs compilation, it is just necessary to put the package into the *src* directory of the CRAN-like repository.

Residing in the folder src/contrib of the PolMine repository at `http://polmine.sowi.uni-due.de/packages`, the *GermaParl* package can be installed in an R session using the *install.packages* function.

```
install.packages(
  "GermaParl",
  repos = "http://polmine.sowi.uni-due.de/packages"
)
```

The package includes a configuration mechanism that will adjust paths in the so-called registry files describing the annotation of a CWB indexed corpus, so that they point correctly to the binary data files in the package. An even simpler installation mechanism is provided by the R package *polmineR*.[9]

```
library(polmineR) # load polmineR package
install.corpus("GermaParl") # install the corpus
```

This is all it takes to have the corpus installed, and to be ready to perform analyses. The following lines of code are examples for initial checks and basic analyses.

```
library(polmineR)
use("GermaParl") # activate GermaParl
corpus() # to see that the corpus is present
size("GERMAPARL") # get the size of GermaParl
kwic("GERMAPARL, query = "Corpus") # concordances
```

[4] See `https://stanfordnlp.github.io/CoreNLP/`. The command-line version of Stanford Core NLP does not work robustly with the structural XML annotation of the corpus. We iterate through the text nodes of the XML documents using an R package 'ctk' (*corpus toolkit*) that offers bindings for Stanford Core NLP, see `https://github.com/PolMine/ctk`.

[5] See `http://cwb.sourceforge.net/`

[6] According to the *CRAN Repository Policy*, packages should usually not exceed 5 MB, see `https://cran.r-project.org/web/packages/policies.html`.

[7] See https://CRAN.R-project.org/package=miniCRAN

[8] See `https://cran.r-project.org/package=drat`.

[9] The *polmineR* package can be installed from CRAN, see `https://CRAN.R-project.org/package=polmineR`. The most recent version of the package is available at GitHub (see `https://www.github.com/PolMine/polmineR`). See the README at CRAN for installation instructions.

A small example may demonstrate that users can indeed proceed quickly to substantial research once *polmineR* – a specialized package to work with CWB indexed corpora using R – and *GermaParl* are installed.[10] Let us assume that you are interested in the adjectives preceding mentions of the European Union ("Europäische Union" in the German corpus). Figure 1 displays a screenshot of an RStudio session (RStudio is the IDE we would recommend for working with R) to do this little exercise. After loading the *polmineR* package, the *GermaParl* corpus is activated by calling *use*. Then, the query (Q) is defined to find matches for the combination of an adjective and the EU. The result of calling the 'count'-method, a data.table is stored as variable Y, and the columns interesting for us (match, count and share) are viewed.

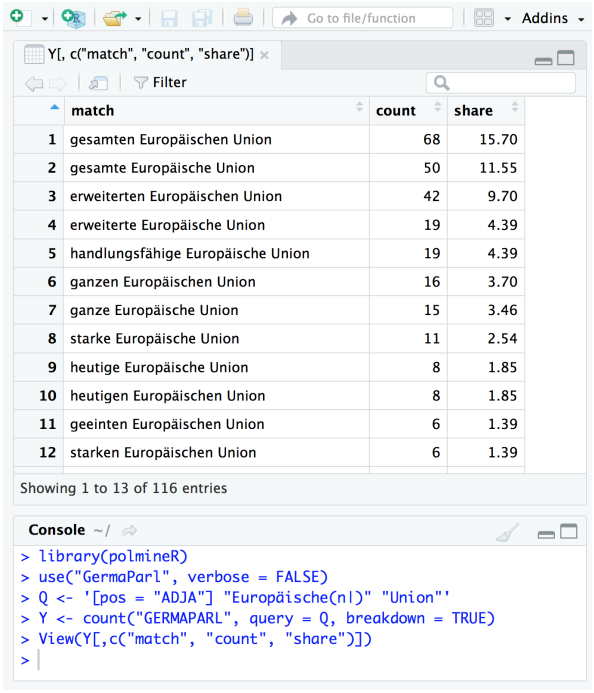


Figure 1: Adjectives before 'European Union' in *GermaParl* (RStudio session).

In a second, related example, we assume we are interested in references to an *enlarged* European Union, this is one of the occurring combinations we saw in the previous exercise. The CQP query we use is '[lemma = "erweitert"] ("EU" — "Europäische(n—)" "Union")'. So we use the linguistic annotation of the corpus and start with the lemma "erweitert" (enlarged). We then allow for the alternatives "European Union" ("Europäische Union"), and its abbreviation as "EU". Five commands lead to the barplot in figure 2: We load *polmineR*, activate *GermaParl*, define the query (variable Q), retrieve the hits using the hits-method, turn the object into a data.table, and produce the barplot. As you may learn from the progress bar in figure 2, it did not even take a second to find the matches for query Q in the 100 million token corpus. Speed is one the advantages of sharing an indexed and compressed corpus.

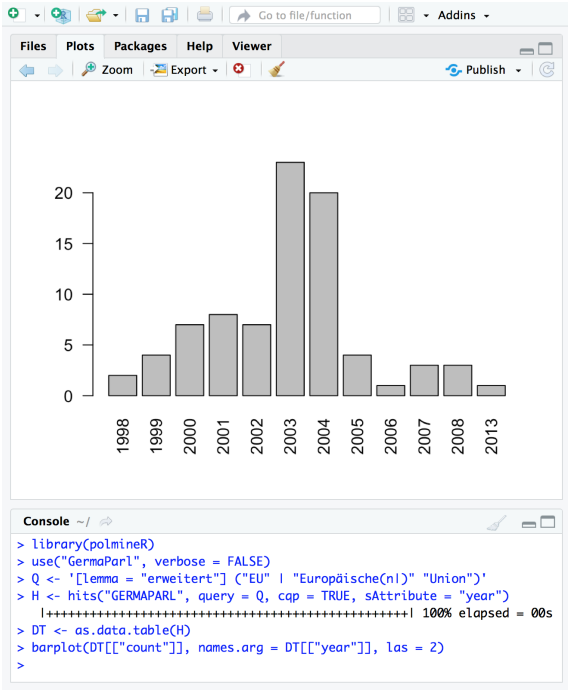


Figure 2: Enlarged 'European Union' (by year).

4. Accessing Data Documentation

One of the big advantages of wrapping corpora into a data package is that it can be accompanied by additional information and extensive documentation. Apart from the standard info file provided for by the CWB, a short document that provides some basic information on a corpus, the *GermaParl* package at the present stage includes two documents called 'vignettes' in the R jargon. The first vignette (with the title 'GermaParl') offers a general introduction to the corpus. It explains corpus preparation, and the structural and linguistic annotation that the corpus has. The second vignette ('MakingOfGermaParl') documents the workflow from XML/TEI documents to the CWB indexed version. These documents are easily accessible from R and can be inspected as follows.

```
Corpus$new("GERMAPARL")$showInfo() # show info file
browseVignettes(package = "GermaParl")
vignette("GermaParl", package = "GermaParl")
vignette("MakingOfGermaParl", package = "GermaParl")
```

From the point of view of reproducible research, the vignette 'MakingOfGermaParl' deserves particular attention. It is a complete documentation of the steps that take the XML corpus from the XML/TEI variant through the NLP pipe to the import into the CWB. It is generated from an *Rmarkdown* document. Following an advice of Hadley Wickham, it is part of the *GermaParl* git repository, i.e. it is stored in the folder data-raw (Wickham, 2015). Upon

[10] I should like to thank the anonymous reviewers for suggesting to include this kind of example.

executing the code in the document and generating the html document from the original Rmarkdown, the CWB indexed corpus is being prepared. Thus, the corpus data included in the package is perfectly reproducible.

Different versions of the 'Making Of'-document result in the different versions of the package. The raw data package is under version control, i.e. it is kept in a git repository. A technical difficulty is that the binary files of the indexed and compressed CWB corpus are large. Using Git LFS (for Large File Storage) is the appropriate solution for this scenario. Because of the size of this git repository, it is hosted at a private GitLab server of the PolMine Project. In a manner known from GitHub, GitLab offers an issue tracker that is very useful to manage issues, user feedback and feature requests. The changes that the corpus has seen are documented in an accessible manner with the file NEWS.md that is included in the package.[11].

5. How to Put Data in a Data Package

Hadley Wickham has written an excellent book on developing R packages (Wickham, 2015).[12] Developing an R data package is usually much easier than writing a package with complex code. To wrap a CWB indexed corpus into a package, we have chosen to put the binary files of the corpus into a package subdirectory inst/extdata/cwb/NAME-OF-THE-CORPUS, and the registry file describing the corpus into a directory inst/extdata/cwb/registry.

The only tricky part is to infuse a configuration mechanism into the package that will set correctly the paths pointing to the data directory with the binary files and the info file. Our best practice is to use an R template script called 'setpaths.R' in a subdirectory 'tools' that is called from the package configure script (for Linux and macOS), or configure.win script (on Windows) respectively.[13]

6. Features and Extra Functionality

Corpora of parliamentary protocols cover all kinds of issues across time. They are multi-purpose corpora. The multiple audiences these corpora target justify why it makes sense to invest resources in developing and maintaining these corpora. But large multi-purpose corpora engender the wish to create thematically defined subcorpora. Having the corpus wrapped into an R data package offers a convenient way to supplement the data with specialized functions to address issues such as this one.

A classification of speeches or agenda items based on a theoretically justified typology of issues and respective training data would be ideal. At the present stage, an additional annotation layer derived from optimized topic models is added to the *GermaParl* corpus. A standard topic modelling approach (Latent Dirichlet Allocation, LDA) has been used, taking as documents parliamentary speeches (not agenda items, for instance).[14] Following what is emerging as good practice, a set of topic models with varying numbers of topics has been trained. A set of parameters has been used to estimate the quality of the models, using the R package *ldatuning*.[15] According to rules of thumb to optimize the number of topics, around 250 topics is a good choice for a topic model for *GermaParl* (see figure 1).

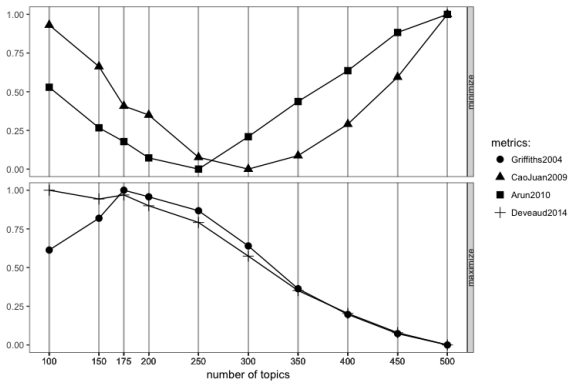


Figure 3: Visualisation of topicmodel optimization exercise.

The integer values of the five topics most prevalent in a speech according to the optimized topic model (k = 250) have been added as a structural attribute to the corpus. Once a user has has identified the topics relevant for his research based on the lists of tokens associated with topics, it is easy to formulate a query on the structural attributes to create a thematic subcorpus.[16]

Distributing a corpus in an R data package does not only offer a coherent way to distribute additional annotations, and the documentation of it. Functionality for additional analytical techniques specific to the data that is disseminated can be included in the package.

7. License and Attribution

The license chosen for the data package is a CLARIN PUB+BY+NC+SA license. The CLARIN licenses[17] are modeled on the Creative Commons licenses. The

[11] A nice side effect of the data package is that it is easy to generate a website from the different documentation files included in a package. The package *pkgdown* offers a handy mechanism to do this, see `https://github.com/r-lib/pkgdown`. Generating a website to promote and document the data is possible with minimal cost, for *GermaParl*, see `http://polmine.sowi.uni-due.de/docs/GermaParl/`

[12] See also `http://r-pkgs.had.co.nz/`.

[13] The script is included in the *ctk* package, see `https://github.com/PolMine/ctk/tree/master/inst/R`.

[14] Identifying speeches is not a trivial question, as parliamentary speeches are interrupted by interjections frequently. The *polmineR* package includes a function with a heuristic to identify speeches.

[15] See `https://CRAN.R-project.org/package=ldatuning`.

[16] This feature is currently only available in the development version of the GermaParl package, but it will be available in an upcoming release. A full documentation of the topic modelling exercise will be provided in an additional vignette. A future version of the *GermaParl* package will also include the functionality to generate classifications based on manually created training data. The data has already been prepared by trained coders in a CLARIN-funded project 'Plenarprotokolle als öffentliche Sprachressource der Demokratie' in 2015/16.

[17] See `https://www.clarin.eu/content/license-categories`.

CLARIN license is derived from the CC Attribution-NonCommercial-ShareAlike 3.0 Unported License.[18] Thus, the elements of the license mean:

- *PUB*: The language resource can be distributed publicly.
- *BY*: Attribution – You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- *NC*: NonCommercial – You may not use the material for commercial purposes.
- *SA*: ShareAlike – If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

Excluding a commercial usage of the corpus is a tactical issue. The restrictive licenses of most commercial publishing houses create considerable burdens for academics to use newspapers in their research. If corpora of newspaper articles are prepared as a corpus, they usually cannot be shared for reproduction and further research. So for the sake of (unfulfilled) reciprocity, it seems well justified if academics who invest energy in preparing corpora of parliamentary protocols preclude commercial users from using their data freely – until they can use commercial data freely, too.

Asking users to attribute the data they are using should be a standard in academic practice, but it is not necessarily how the digital world behaves. However, the R community has always had very strong ties to academia, and packages are meant to be quoted. Packages include statements of authorship and can include binding suggestions how they should be cited.[19] Using an R data package to disseminate corpus data implies a solution how authorship can be attributed. Offering data in a way that is quotable is an incentive to share data.

8. Perspectives

The *GermaParl* corpus of parliamentary protocols is made available as XML (TEI standardization) at a GitHub repository. Yet many users cannot be expected to be sufficiently acquainted with the NLP techniques necessary to turn the XML into a linguistically annotated corpus without hassle. This paper suggests that offering a linguistically annotated and indexed version of the corpus wrapped in an R data package may be a neat way to disseminate the data. It lowers barriers of entry for academic users that are not full-fledged computational linguists. What is more, the R data package *GermaParl* is intended to suggest a way how corpus preparation can be documented and made transparent. Making progress towards reproducible research is the ultimate aim.

To be sure, the suggestion is accompanied by the idea that *GermaParl* might become part of a larger family of corpora of plenary protocols that will be available as R data packages via CRAN-like repositories. For future research, it might indeed be very productive to have shared ideas how we maintain and share our corpora. The basic corpus preparation – attaining XML – is time-consuming and may absorb considerable attention. But there is a set of relevant questions and best practices beyond the XML stage of data preparation. How do we share the results of supervised, or unsupervised learning, for instance? There will be many further questions that the emerging availability of corpora of parliamentary protocols will engender. It will be good to have some common ideas how we maintain, document and share our data – for the sake of being somewhat more cumulative in our research endeavours.

9. Bibliographical References

Blätte, A. and Blessing, A. (2018). The GermaParl Corpus of Parliamentary Protocols. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2018, 7-12 May 2018, Miyazaki, Japan*.

Blätte, A. (2018). Zum Verwechseln ähnlich? Eine Klassifikationsanalyse parlamentarischen Diskursverhaltens auf Basis des PolMine-Plenarprotokollkorpus. In Joachim Behnke, et al., editors, *Computational Social Science. Die Analyse von Big Data*. Nomos, Baden-Baden.

Hardie, A. (2012). CQPweb combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380 – 409.

Kilgarriff, A., Rychl, P., Smr, P., and Tugwell, D. (2004). The Sketch Engine. *Information Technology*.

Kilgarriff, A., Baisa, V., Buta, J., Jakubek, M., Kovv, V., Michelfeit, J., Rychl, P., and Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, pages 7–36.

Marx, M. and Schuth, A. (2010). DutchParl. The parliamentary documents in dutch. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.

Wickham, H. (2015). *R packages. Organize, test, document and share your code*. O'Reilly, Sebastopol, CA.

[18] See `https://creativecommons.org/licenses/by-nc-sa/3.0/` for further explanations.

[19] On CITATION files see `https://cran.r-project.org/doc/manuals/r-release/R-exts.html`.