Prendo la Parola in Questo Consesso Mondiale: A Multi-Genre 20th Century Corpus in the Political Domain

Sara Tonelli[†], **Rachele Sprugnoli**[‡], **Giovanni Moretti**[†‡]
[†]Fondazione Bruno Kessler, Trento, Italy
[‡]Università Cattolica, Milano, Italy
{satonelli,moretti}@fbk.eu
rachele.sprugnoli@unicatt.it

Abstract

English. In this paper we present a multi-genre corpus spanning 50 years of European history. It contains a comprehensive collection of Alcide De Gasperi's public documents, 2,762 in total, written or transcribed between 1901 and 1954. The corpus comprises different types of texts, including newspaper articles, propaganda documents, official letters and parliamentary speeches. The corpus is freely available and includes several annotation layers, i.e. key-concepts, lemmas, PoS tags, person names and geo-referenced places, representing a high-quality 'silver' annotation. We believe that this resource can foster research in historical corpus analysis, stylometry and computational social science, among others.[1]

1 Introduction

In recent years, political scientists and history scholars have started to exploit the availability of digital material to enrich their research, taking advantage of freely accessible online archives and easy-to-use tools for text processing and data extraction. Active communities have been created around topics such as the study of Parliamentary corpora (see the ParlaCLARIN[2] and ParlaFormat workshops[3]), the analysis of political manifestos[4] and of Presidential speeches.[5] Despite the importance of this research field, copyright and availability in machine-readable format still represent major issues, especially in those countries where no or only limited public initiatives have been undertaken to support the distribution of this kind of documents. For example, while in the US the Federal Digital System grants access to public Presidential documents through APIs and bulk-data repositories, in Italy an effort along this line has started only recently with the support of the Archive of the President of the Republic[6], but has not delivered substantial results so far.

This work represents a first attempt to deal with this lack of data, since we present and make available a large corpus of Italian public documents in the political domain. In particular, we release a comprehensive collection of Alcide De Gasperi's public documents issued between 1901 and 1954, which had been previously published in four volumes by Il Mulino (De Gasperi, 2006; De Gasperi, 2008a; De Gasperi, 2008b; De Gasperi, 2009) but were not machine-readable. Our repository contains all documents in three formats: txt, XML and tab-separated. Raw text files contain only the body of the documents, and may be straightforwardly used to extract embeddings or topics. XML files include metadata that cover not only the title, the date and the place of publication, but also key-concepts automatically extracted from each text and genre labels manually assigned by domain experts. Furthermore, the release includes silver annotation for lemma, part of speech, person names and place names with associated coordinates in a CoNLL-like format. All files and the corresponding descriptions can be downloaded at `https://dh.fbk.eu/technologies/corpus-de-gasperi` (with CC BY-NC-SA license). The corpus can also be navigated using the ALCIDE platform (Moretti et al., 2016) at this link: `http://alcidedigitale.fbk.eu/`.

[2] `https://www.clarin.eu/ParlaCLARIN`

[3] `https://www.clarin.eu/event/2019/parlaformat-workshop`

[4] `https://manifesto-project.wzb.eu/`

[5] `https://www.presidency.ucsb.edu/documents`

[6] `https://archivio.quirinale.it/aspr/`

2 Related Work

The political domain has been studied in computational linguistics from various perspectives. Annotated corpora have been created to analyse rhetoric and metaphors in political communication (Cardie and Wilkerson, 2008; Ahrens et al., 2018), study the impact of speeches on the audience (Guerini et al., 2013; Thomas et al., 2006) and understand the relationship between ideology and linguistic complexity (Schoonvelde et al., 2019). Resources have also been developed to train and test automatic systems for several types of NLP tasks, such as persuasiveness prediction (Strapparava et al., 2010), sentiment and emotion analysis (Young and Soroka, 2012; Rheault et al., 2016), text classification (Yu et al., 2008), topic-based agreement detection (Menini et al., 2017) and recognition of ideological positions (Hirst et al., 2010).

Many research activities have recently dealt with the digitisation and release of corpora containing historical political texts. For example, the corpus of speeches given in the British Parliament from 1803 to 2005 (i.e. the Hansard Corpus) has been automatically tagged using the Historical Thesaurus Semantic Tagger (Piao et al., 2014; Wattam et al., 2014) and then a part of it has been semantically enriched with information about speakers and topics (Nanni et al., 2019). In addition, the Canadian Parliamentary Debates (1901-present) have been standardised, enriched and distributed within the "Digging into Linked Parliamentary Data" project (Beelen et al., 2017). The period from 1947 to 2017 is instead covered by a dataset of Dutch and Danish party congress speeches (Schumacher et al., 2019).

As for Italian, to the best of our knowledge, the only available comprehensive study of the language of Italian politicians is the one by Bolasco (2015). He analyses the parliamentary proceedings of the Italian Chamber of Deputies in the period 1953-2008 using the TalTac2 software[7], thus providing a lexical and statistical analysis. Another project related to our work is "Voci della Grande Guerra" whose online platform allows to explore a corpus of documents related to the first World War including samples of parliamentary proceedings and political speeches (Lenci et al., 2016). Similarly to what we present in this paper, such documents have been automatically annotated and then partially revised by hand (De Felice et al., 2018). Compared with these two last works, our corpus is broader, having a multi-layered semantic analysis, and completely available for download in different formats, thus open to further analysis by the research community.

3 Corpus Description

Our corpus contains the complete collection of public documents by Alcide De Gasperi, the first Prime Minister of the Italian Republic and one of the founding fathers of the European Union. It includes 2,762 documents published between 1901 and 1954, for a total of around 3,000,000 tokens. The corpus is released as raw text, as XML with a minimal set of meta-data and associated key-concepts, and as CoNLL-like format, with additional information that have been fully or semi-automatically annotated (see Section 4). Texts, date and place of publication were automatically generated starting from the PDF files used to issue the volumes edited by Il Mulino. Each document of the collection was classified manually by a group of history scholars on the basis of a two-layered hierarchy that takes into consideration whether the text was originally released in an oral or written form, and its specific genre. It is important to note that different text genres correspond to different roles covered by De Gasperi during his life: e.g. daily press when he worked as a journalist for newspapers in Trentino, speeches in institutional venues when he was a Member of the Italian Parliament.

History scholars identified also four time spans to which each document can be assigned, that characterise different periods in De Gasperi's life. These correspond to the four volumes of the printed edition and are used to split the corpus into different periods based on the date of publication:

Vol. I : De Gasperi was a journalist and a students' leader. He was active mainly in Trento and in the Austrian Parliament (1901 – 1918).

Vol. II : De Gasperi founded Partito Popolare, became Parliament member in Rome and then left the Italian political life for several years after opposing the Fascist regime, working at the Vatican library and as a publicist (1919 – 1942).

Vol. III : De Gasperi founded the Christian-Democratic Party, became Prime Minister

[7] http://www.taltac.it/

Document	Type	Number
Written documents	Monographs / Prefaces	4
	Daily press	963
	Magazines	228
	Official documents	433
Speeches	Electoral / propaganda	473
	Party conferences	188
	Institutional venues	419
Not specified	Not specified	54

Table 1: Genre labels with corresponding statistics.

and was Italian delegate at the World War II peace conference (1943 – 1948).

Vol. IV : After Christian Democracy led by De Gasperi won the first general elections of the Italian Republic, he launched a plan of reforms to reconstruct Italy including social housing, labor policy and unemployment insurance (1949 – 1954).

4 Annotated Information

The annotations included in the release are:

- Lemma and PoS: the corpus has been lemmatised and PoS-tagged using the TextPro suite (Pianta et al., 2008). The module for the lemmatization is a rule-based system, whereas the part-of-speech annotation is statistical and has been trained on the EVALITA 2007 dataset (Tamburini, 2007) following the EAGLES tagset (Monachini, 1996).

- Person and place names: named entities have been tagged using the NER module included in TextPro and trained on the I-CAB corpus (Magnini et al., 2006). Geopolitical entities (GPEs) have also been geo-referenced using Nominatim[8] (Clemens, 2015). The number of person and place names per volume is provided in Table 2.

After running the automatic modules, the output was uploaded in the ALCIDE platform (Moretti et al., 2016) and, through its navigation interface, we identified annotations that were systematically wrongly tagged, and fixed them manually. An evaluation of the automatic annotation is reported in Section 5.

In addition to the annotations previously mentioned, each document is assigned to a set of key-concepts, that is a weighted list of n-grams representing the most important concepts of a text, automatically extracted using KD (Moretti et al., 2015).

5 Annotation Evaluation

We evaluated the quality of the automatic annotation produced by TextPro modules on a subset of our corpus. Indeed, since these modules were developed to perform best on contemporary texts, and typically trained on news, it is important to assess to what extent they can be reliably used on Italian documents of the XX Century in the political domain. To this end we manually annotated a gold standard made of documents written by De Gasperi between 1906 and 1911 for a total of 8,872 tokens. We chose texts belonging to the first period of De Gasperi's life because they are the oldest in the corpus and therefore the most linguistically different from the texts used for training the modules. Results of the evaluation are compared with the ones obtained by TextPro on contemporary texts.

5.1 Lemmatization

Table 3 shows TextPro accuracy obtained on our gold standard compared with the ones reported in Aprosio and Moretti (2018) and calculated on the Universal Dependencies (UD) test set for Italian (Bosco et al., 2013). The drop of 0.7 points in accuracy is mainly due to some repeated anomalies of the module in the lemmatization of definite and indefinite articles (which are lemmatized using the labels "det" and "indet", instead of singular masculine forms "il" and "uno") and to the non-recognition of truncated words, such as "far", "bel", "andar", "vuol", not common in contemporary texts. Other sources of errors are the presence of obsolete terms, e.g. "libello", "soziale", "donde", and the use of preterite (*passato remoto*, e.g. "andò", "apparve"), a grammatical tense not very frequent in contemporary news. Most of previously mentioned anomalies have been fixed through a set of rules applied after data processing: after this correction, accuracy has risen to 0.97.

5.2 PoS Tagging

The presence of obsolete words, truncated forms and preterite verbs leads to errors also in the PoS tagger of TextPro. However, for this module the impact is less evident than for lemmatization: as

[8] https://nominatim.openstreetmap.org/

VOL I		VOL II		VOL III		VOL IV	
PER	**GPE**	**PER**	**GPE**	**PER**	**GPE**	**PER**	**GPE**
4,126	6,168	2,890	2,956	3,018	4,324	5,701	6,308
Gesù Cristo	Trento	Gesù Cristo	Italia	Palmiro Togliatti	Italia	Pietro Nenni	Italia
Augusto Avancini	Alto Adige	Mussolini	Roma	Pietro Nenni	Trieste	Palmiro Togliatti	Europa
Karl Lueger	Trentino	Leone XIII	Germania	Marshall	Russia	Tito	Trieste

Table 2: Occurrences of PER and GPE per volume, with three top-frequent entities for each category.

	UD Test Set	De Gasperi Corpus
	Accuracy	Accuracy
Lemma	0.96	0.89

Table 3: Comparison of lemmatization performance on the Italian UD test set and on our gold standard.

reported in Table 4, on De Gasperi's documents the performance drop is only 0.1 points accuracy with respect to the results obtained on the UD test set. Table 5 gives details on the number and distribution of errors per grammatical category. Categories registering the higher quantity of mistaken tags are nouns, proper nouns, verbs and adjectives. Most mistakes concerning nouns are due to words capitalised to show formal respect towards highest representatives of the State or of the Church (e.g. "Vescovo") and German common nouns that all have the initial capital letter.

	UD Test Set	De Gasperi Corpus
	Accuracy	Accuracy
PoS	0.96	0.95

Table 4: Comparison of PoS tagging performance on the Italian UD test set and on our gold standard.

Grammatical Category	#errors	%errors
Adjectives	62	15.54
Adverbs	24	6.02
Conjunctions	6	1.50
Demonstrative Adjectives	8	2.01
Prepositions	10	2.51
Pronouns	12	3.01
Relative Pronouns	1	0.25
Articles	11	2.76
Nouns	94	23.56
Proper Nouns	91	22.81
Verbs	73	18.30
Acronyms	6	1.50
Foreign Terms	1	0.25

Table 5: PoS-tagging errors per category.

5.3 Persons and GPEs

In Table 6 the performance of automatic recognition of persons (PER) and geo-political entities (GPE) in De Gasperi's documents is compared with the scores TextPro obtained in the EVALITA 2007 campaign (Speranza, 2007), when trained and tested on a newswire corpus. The tool shows a drop in performance on our gold standard only in the recognition of persons' names (-0.16 F1 points), whereas place names seem to be more stable (+0.1 F1 points). In both categories, precision has decreased more than recall: to improve it, we manually checked the named entities detected by the automatic module in the whole corpus removing the wrong ones. We also verified the latitude and the longitude retrieved with Nominatim for all the GPEs assigning new correct coordinates to about 6% of them. Errors were mainly related to places that no longer exist or that have changed names after the death of De Gasperi, (e.g. "Prussia", "Congo Belga") and to little villages in the Trentino area (e.g. "Oltresarca", "Termon").

	EVALITA 2007 test set			De Gasperi corpus		
	P	R	F1	P	R	F1
PER	0.92	0.93	0.92	0.70	0.82	0.76
GPE	0.85	0.86	0.85	0.82	0.90	0.86

Table 6: Comparison of NER performance on news and on our gold standard.

6 Use Cases

The corpus has been used to perform a number of pilot studies, which have confirmed the potential of this kind of resource and could represent a starting point for further developments (Sprugnoli et al., 2016). Three of these studies are described in this Section.

A first analysis has been carried out with the goal of studying De Gasperi's rhetoric strategy through his use of verb tenses, considered as an important marker of temporality (Sprugnoli et al., 2018). This study is based on the paradigm proposed by Chilton (2004), who includes time among the three axes of the political discourse together with space and modality.

We run the morphological analyzer included in TINT NLP Suite to recognise the tenses of all verbs of the corpus. We then merge them into present, past and future tense and compare the distribution of the three classes across the four volumes. We observe that there is an evident difference between the use of verb tenses before and after 1943. Indeed, in the first two volumes past tenses are more frequently used, with a highly statistically significant difference with respect to volumes III and IV ($p < 0.001$ using Wilcoxon signed-rank test). On the other hand, after 1943 De Gasperi uses more present and future tense, again with high statistical significance. This can be explained by the fact that the last volumes contain many press reports describing the programmatic commitment of Christian Democracy as well as letters and telegrams sent by De Gasperi as Minister of Foreign Affairs, where the development of prospective collaborations is proposed. The last volume discusses also the reforms to be adopted for the reconstruction of the newly born Italian Republic and those about the forthcoming creation of a European Community. In general, after 1943 we observe a shift of focus from past events to the contemporary and future dimension.

A second analysis related to temporality deals with cited persons, which were linked to a Dbpedia entry using the Wiki Machine (Palmero Aprosio and Giuliano, 2016). Through this link, each person is associated with a *dbo:birthDate* and *dbo:deathDate* and then to a Past or Present label, again using the document date as a reference. Persons are considered part of the past if the referent was dead before the document publication time. Using the classification algorithm described in (Palmero Aprosio et al., 2017) we further assign a semantic category to each mention. A comparative analysis shows that contemporary persons are generally more cited than past ones, but also that the category of persons mentioned in the document changes significantly across the volumes: while in Volume I cited persons include politicians but also religious figures and artists, this range of figures decreases over time, with almost exclusively political figures mentioned in Volume IV. As an example, we report in Fig. 1 and Fig. 2 the top-cited persons in Vol. I and IV respectively: while in the early documents Beethoven, Dante and Nietzsche are highly cited, persons mentioned in the late documents include exclusively politicians and religious figures, all from present time or recent past. With reference to the previously cited dimensions in Chilton (2004), this shift should be seen in the light of De Gasperi's effort after 1943 to justify past and present policy, using mentioned persons to build a national ideology.

A third analysis focused on how temporal information is expressed in De Gasperi's documents (Speranza and Sprugnoli, 2018). To explore this aspect we manually annotated ten newspaper articles, published in 1914 and related to the outbreak of the Great War, following the It-TimeML guidelines (Caselli et al., 2011). This resource has been used in the EVENTI task organized within EVALITA 2014 (Caselli et al., 2014) and is freely available online. The average number of annotated events and temporal relations in the documents written by De Gasperi is higher than in contemporary newspaper articles annotated following the same guidelines, whereas the density of temporal expressions is comparable. Other differences concern the type of events, temporal expressions and temporal relations present in the historical texts. For example, De Gasperi frequently uses events expressing personal opinions about the topics covered in the articles. The high presence of speculations influences the temporal structure of the texts: in many cases events are not ordered chronologically but presented as simultaneous with respect to the time of writing. Moreover, temporal expressions are mainly non-specific or fuzzy: a characteristic that is less evident in other corpora of contemporary texts, and that may be related to the more speculative nature of political texts.

7 Conclusions

In this paper we present the release of the corpus of Alcide De Gasperi's public writings, including 2,762 documents and around 3 million tokens. We make available raw texts, XML files having a small set of metadata and key-concepts and CoNLL-like files with lemma, PoS, PER, GPE annotation together with the coordinates of place names. Based on an evaluation performed on all four annotation layers, we show that their quality is good, although annotation was performed automatically and only partially revised.

This is the first freely available corpus of this kind, and we hope that it can be used to foster research in political science, corpus linguistics and history, as well as to develop and test NLP systems using data that are different from widely used contemporary news.

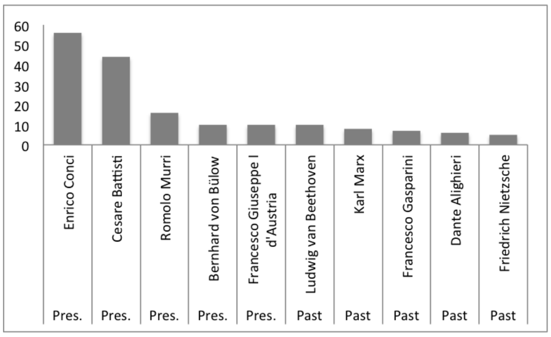

Figure 1: Past and present persons mentioned in Vol. 1.

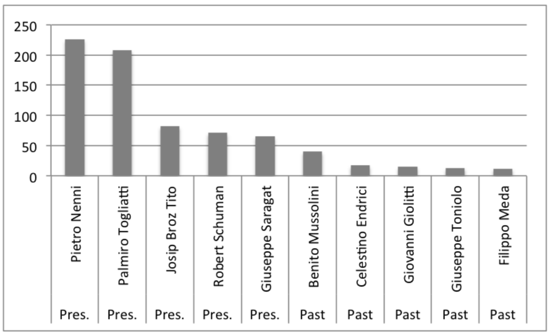

Figure 2: Past and present persons mentioned in Vol. 4.

Acknowledgments

We thank the colleagues from the Italian-German Historical Institute at Fondazione Bruno Kessler for their help in annotating De Gasperis corpus, and Edizioni Il Mulino, for giving access to the corpus and allowing its release. The project has been partially supported by Fondazione Cassa di Risparmio di Trento e Rovereto and Fondazione Cassa di Risparmio delle Province Lombarde.

References

Kathleen Ahrens, Huiheng Zeng, and Shun-han Rebekah Wong. 2018. Using a Corpus of English and Chinese Political Speeches for Metaphor Analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).*

Alessio Palmero Aprosio and Giovanni Moretti. 2018. Tint 2.0: an All-inclusive Suite for NLP in Italian. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018.*

Kaspar Beelen, Timothy Alberdingk Thijm, Christopher Cochrane, Kees Halvemaan, Graeme Hirst, Michael Kimmins, Sander Lijbrink, Maarten Marx, Nona Naderi, Ludovic Rheault, et al. 2017. Digitization of the Canadian parliamentary debates. *Canadian Journal of Political Science/Revue canadienne de science politique*, 50(3):849–864.

Sergio Bolasco, 2015. *Sulla costruzione di un corpus per l'analisi automatica del linguaggio parlamentare dei leader*, chapter 5. Camera dei Deputati.

Cristina Bosco, Montemagni Simonetta, and Simi Maria. 2013. Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In *7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69. The Association for Computational Linguistics.

Claire Cardie and John Wilkerson. 2008. Text Annotation for Political Science Research. *Journal of Information Technology & Politics*, 5(1):1–6.

Tommaso Caselli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. 2011. Annotating events, temporal expressions and relations in Italian: the It-TimeML experience for the Ita-TimeBank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 143–151. Association for Computational Linguistics.

Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. 2014. EVENTI EValuation of Events and Temporal INformation at Evalita 2014. In *Proceedings of the Fourth International Workshop EVALITA 2014*, pages 27–34.

Paul Chilton. 2004. *Analysing political discourse: Theory and practice*. Routledge.

Konstantin Clemens. 2015. Geocoding with openstreetmap data. *GEOProcessing 2015*, page 10.

Irene De Felice, Felice DellOrletta, Giulia Venturi, Alessandro Lenci, and Simonetta Montemagni. 2018. Italian in the Trenches: Linguistic Annotation and Analysis of Texts of the Great War. In *Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 160–164. Accademia University Press.

Alcide De Gasperi. 2006. Alcide De Gasperi nel Trentino asburgico. In *Scritti e discorsi politici di Alcide De Gasperi*, volume 1. Il Mulino.

Alcide De Gasperi. 2008a. Alcide De Gasperi dal Partito popolare italiano all'esilio interno 1919-1942. In *Scritti e discorsi politici di Alcide De Gasperi*, volume 2. Il Mulino.

Alcide De Gasperi. 2008b. Alcide De Gasperi e la fondazione della Democrazia cristiana, 1943-1948. In *Scritti e discorsi politici di Alcide De Gasperi*, volume 3. Il Mulino.

Alcide De Gasperi. 2009. Alcide de Gasperi e la stabilizzazione della Repubblica 1948-1954. In *Scritti* *e discorsi politici di Alcide De Gasperi*, volume 4. Il Mulino.

Marco Guerini, Danilo Giampiccolo, Giovanni Moretti, Rachele Sprugnoli, and Carlo Strapparava. 2013. The new release of CORPS: A corpus of political speeches annotated with audience reactions. In *Multimodal Communication in Political Speech. Shaping Minds and Social Action*, pages 86–98. Springer.

Graeme Hirst, Yaroslav Riabinin, and Jory Graham. 2010. Party status as a confound in the automatic classification of political speech by ideology. In *Proceedings of the 10th International Conference on Statistical Analysis of Textual Data (JADT 2010)*, pages 731–742.

Alessandro Lenci, Nicola Labanca, Claudio Marazzini, and Simonetta Montemagni. 2016. Voci della Grande Guerra An Annotated Corpus of Italian Texts on World War I. *Italian Journal of Computational Linguistics*, pages 101–108.

Bernardo Magnini, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. 2006. I-CAB: the Italian Content Annotation Bank. In *LREC*, pages 963–968.

Stefano Menini, Federico Nanni, Simone Paolo Ponzetto, and Sara Tonelli. 2017. Topic-based agreement and disagreement in US electoral manifestos. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2938–2944.

Monica Monachini. 1996. ELM-it: EAGLES specifications for Italian morphosyntax lexicon specification and classification guidelines. Technical report, Centre National de la Recherche Scientifique Paris, France.

Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. 2015. Digging in the Dirt: Extracting Keyphrases from Texts with KD. In *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*.

Giovanni Moretti, Rachele Sprugnoli, Stefano Menini, and Sara Tonelli. 2016. ALCIDE: Extracting and visualising content from large document collections to support Humanities studies. *Knowledge-Based Systems*, 111:100–112.

Federico Nanni, Stefano Menini, Sara Tonelli, and Simone Paolo Ponzetto. 2019. Semantifying the UK Hansard (1918-2018). In *Proceedings of JCDL19*.

Alessio Palmero Aprosio and Claudio Giuliano. 2016. The Wiki Machine: an open source software for entity linking and enrichment. *ArXiv e-prints*, September.

Alessio Palmero Aprosio, Sara Tonelli, Stefano Menini, and Giovanni Moretti. 2017. Using Semantic Linking to Understand Persons' Networks Extracted from Text. *Front. Digital Humanities*, 2017.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The TextPro Tool Suite. In *Proceedings of Language Resources and Evaluation Conference*, pages 2603–2607, Marrakech, Morocco.

Scott Piao, Fraser Dallachy, Alistair Baron, Paul Rayson, and Marc Alexander. 2014. Developing the Historical Thesaurus Semantic Tagger. In *The Digital Humanities Congress 2014*.

Ludovic Rheault, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. 2016. Measuring emotion in parliamentary debates with automated textual analysis. *PloS one*, 11(12):e0168843.

Martijn Schoonvelde, Anna Brosius, Gijs Schumacher, and Bert N Bakker. 2019. Liberals lecture, conservatives communicate: Analyzing complexity and ideology in 381,609 political speeches. *PloS one*, 14(2):e0208450.

Gijs Schumacher, Daniel Hansen, Mariken ACG van der Velden, and Sander Kunst. 2019. A new dataset of Dutch and Danish party congress speeches. *Research & Politics*, 6(2):2053168019838352.

Manuela Speranza and Rachele Sprugnoli. 2018. Annotation of Temporal Information on Historical Texts: a Small Corpus for a Big Challenge. *Formal Representation and the Digital Humanities*, page 203.

Manuela Speranza. 2007. EVALITA 2007: The Named Entity Recognition Task. In *Proceedings of the EVALITA 2007 Workshop on Evaluation of NLP Tools for Italian*, pages 66–68, Rome, Italy.

Rachele Sprugnoli, Giovanni Moretti, Sara Tonelli, and Stefano Menini. 2016. Fifty years of european history through the lens of computational linguistics: the de gasperi project. *IJCol-Italian journal of computational linguistics*, 2(2):89–100.

Rachele Sprugnoli, Giovanni Moretti, and Sara Tonelli. 2018. Temporal Dimension in Alcide De Gasperi: Past, Presentand Future in Historical Political Discourse. In *AIUCD 2018 - Book of Abstracts*, pages 77–80.

Carlo Strapparava, Marco Guerini, and Oliviero Stock. 2010. Predicting Persuasiveness in Political Discourses. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1342–1345.

Fabio Tamburini. 2007. Evalita 2007: The Part-of-Speech Tagging Task. *Intelligenza artificiale*, 4(2):57–73.

Matt Thomas, Bo Pang, and Lillian Lee. 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.

Stephen Wattam, Paul Rayson, Marc Alexander, and Jean Anderson. 2014. Experiences with Parallelisation of an Existing NLP Pipeline: Tagging Hansard. In *LREC*, pages 4093–4096.

Lori Young and Stuart Soroka. 2012. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.

Bei Yu, Stefan Kaufmann, and Daniel Diermeier. 2008. Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5(1):33–48.