IJCoL

Italian Journal
of Computational Linguistics

Rivista Italiana
di Linguistica Computazionale

Volume 2, Number 2
december 2016

Special Issue:
Digital Humanities and Computational Linguistics

Guest Editors:
John Nerbonne, Sara Tonelli

Accademia
university
press

editorial board

editorial office

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

direttore responsabile
Michele Arnese

isbn 978-88-99982-26-3

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_2_2

IJCoL

Volume 2, Number 2
december 2016

Special Issue:
Digital Humanities and Computational Linguistics

Guest Editors:
John Nerbonne, Sara Tonelli

CONTENTS

Fifty years of European history through the Lens of Computational Linguistics: the De Gasperi Project

Rachele Sprugnoli[**]
Fondazione Bruno Kessler and
Università di Trento

Giovanni Moretti[*]
Fondazione Bruno Kessler

Sara Tonelli[*]
Fondazione Bruno Kessler

Stefano Menini[**]
Fondazione Bruno Kessler and
Università di Trento

In 2013 a collaboration was started between the Digital Humanities research unit and the Italian-German Historical institute at Fondazione Bruno Kessler, whose goal was to develop tools and strategies to give new insight into the public documents written by Alcide De Gasperi. Through the analysis of textual occurrences, semantic structures, and temporal patterns, the project, ending in 2017, has investigated the formation and evolution of De Gasperi's political action and the rhetorical discourse that accompanied its developments. Research of this kind has contributed to better understanding De Gasperi's thoughts, the links between his language and his political views and culture, and the instruments of his communication. On the other hand, it allows the design and development of automated research tools for the political domain. This is a distinctly interdisciplinary research project, in which historical inquiry interacts with methods of linguistic analysis and with information and communication technologies.

1. Introduction

Political communication in the last twenty years has been more and more characterised by distinctive styles, which signal the kind of contact a politician estabilishes with the audience. Silvio Berlusconi's jokes, the colorful language of the North League for the Independence of Padania and *vaffa*/'f***-off' mantra by so-called Five Star Movement are clearly not just a matter of words, but shape and define the audience a politician is talking to. Even if less evident, the same phenomena could also be observed with politicians of the last century, for example Alcide De Gasperi, Palmiro Togliatti, Pietro Nenni, Aldo Moro, Enrico Berlinguer.

The centrality of word in political history is nothing new: the word is the basic instrument of communication, the space in which politics is action. Through language, we build consensus, define parties with their values, and let ideologies take shape (Wodak 1989). This leads to a series of questions, which researchers from different disciplines have tried to address (Chilton 2004; Howarth, Norval, and Stavrakakis 2000):

[*] Digital Humanities Group - Via Sommarive 18, Povo, Italy. E-mail: {moretti,satonelli}@fbk.eu
[**] Digital Humanities Group and Università di Trento - Via Sommarive, Povo, Italy. E-mail: {sprugnoli,menini}@fbk.eu

what are the rules of political language? How does communication between politicians and society work? To what extent are politicians' words influenced by citizens? And more generally, what are the strategies by which we build consensus?

In recent years, studies in this area have benefited from new methods and techniques offered by Digital Humanities research. This includes, for example, the possibility to process large amounts of data, analyse them from different perspectives and display such analyses following data visualisation principles. The inter-disciplinary project on De Gasperi's public documents, which is currently ongoing at Fondazione Bruno Kessler (FBK) in Trento, is part of this trend. The project is in fact a collaboration between the Italian-German Historical Institute and the Digital Humanities research unit at FBK. Its goal is to give new insight into De Gasperi's communication strategy with the help of innovative tools for text analysis. This project represents one of the few attempts to overcome the gap between lexical and rhetorical studies in the political domain, and to our knowledge is the first one that covers the whole public life of an Italian politician, also thanks to the availability of the complete collection of De Gasperi's public documents. Furthermore, most works on Italian political discourse have focused on politicians from the Second Republic (i.e. after 1994), and the few studies related to De Gasperi (Desideri 1984; Vinciguerra 2016) have approached his discourse with traditional methods, without the help of computational tools, considering only documents from a specific time period. In this project, we apply for the first time close and distant reading (Moretti 2013) to the study of Italian political communication. We believe that this multi-faceted analysis of De Gasperi's public documents will give new insight into the main phases of Italian recent history.

The paper is structured as follows: in Section 2 we provide an overview of projects dealing with political communication in the Digital Humanities area. In Section 3 the De Gasperi project and the creation of the corpus are briefly presented. In Section 4 the first project phase is detailed, with a discussion about the performance of NLP modules and a preliminary set of corpus-based findings. Then, we describe in Section 5 the ongoing work and the plans for future research inside the project. Finally, we draw some conclusions and comment the project findings in Section 6.

2. Related Works

In the field of computational linguistics, political texts are the focus of many studies aimed at shedding lights on the peculiarities of political communication and rhetoric (Cardie and Wilkerson 2008). Annotated corpora, for example (Guerini et al. 2013; Thomas, Pang, and Lee 2006), have been created to predict persuasiveness and thus the impact of speeches on the audience (Strapparava, Guerini, and Stock 2010) and to develop opinion mining systems (Balahur, Kozareva, and Montoyo 2009). The literature also reports works on the automatic recognition of ideological positions in political texts (Hirst, Riabinin, and Graham 2010), classification of texts by parties (Yu, Kaufmann, and Diermeier 2008) and sentiment analysis of political communication (Young and Soroka 2012). Lately, attention has been given to the analysis of big data (Sudhahar, Veltri, and Cristianini 2015) and historical documents (Rule, Cointet, and Bearman 2015).

The historical dimension is crucial also in recent Digital Humanities projects. For example "Political Language in the Middle Ages" investigates how political words and concepts change in medieval Latin texts by employing a computer-based corpus-linguistic approach (Cimino, Geelhaar, and Schwandt 2015). Semantic analysis is instead central in the SAMUELS (Semantic Annotation and Mark-Up for Enhancing Lexical Searches) project, in which the Hansard corpus, containing the speeches given in the British Parliament from 1803 to 2005, has been automatically tagged using the Historical Thesaurus Semantic Tagger (Piao et al. 2014; Wattam et al. 2014). Keyword extraction, topic modelling and readability analyses are combined with data visualization to analyze argumentation in English political negotiations in the VisArgue project (Gold et al. 2015).

As for Italian, computational linguistics approaches have been applied mostly to newspaper articles and social media texts in order to analyse how political issues are portrayed outside institutional forums (Stranisci et al. 2015; Delmonte, Gîfu, and Tripodi 2013). 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 (1953-2008) from a statistical and lexical point of view using the TalTac2 software[1]. While this kind of processing is also performed in the De Gasperi project, our goal is broader, in that we aim at performing a multi-layered semantic analysis of De Gasperi's corpus, thus enabling a higher-level interpretation of the temporal and discourse dimension in the politician's documents.

3. Background to the project and corpus creation

The analysis of De Gasperi's public documents has been the first collaboration between the ICT and the History Center at Fondazione Bruno Kessler. In the first phase, from 2013 to 2015, it was mainly an internal project devoted to the creation of a software infrastructure to perform corpus-based analyses of lage document collections in the political domain. The De Gasperi corpus was used as a testbed to design text analysis tools and visualisations in collaboration with history scholars. The second phase, started at the end of 2015, was jointly funded by Fondazione Cassa di Risparmio di Trento e Rovereto and Fondazione Cariplo, and will last till 2017, with the goal to investigate De Gasperi's rhetorical strategies and in particular his use of the past, present and future dimension with different types of audience.

Our project is built around 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. This corpus comprises 2,762 documents (around 3,000,000 tokens) published between 1901 and 1954. Starting from the PDF files used to issue the 4 volumes edited by Il Mulino (De Gasperi 2006, 2008a, 2008b, 2009), we created a corpus of XML files containing the content of each document together with a set of metadata, i.e. title, date and place of publication. Given that different types of political documents are included in the corpus (Cortelazzo and Paccagnella 1981), a history scholar defined two tag hierarchies: one concerns the different public roles played by De Gasperi during his career, while the other includes the types of documents in the corpus (e.g. written or oral). The two hierarchies were defined with the goal to analyse whether De Gasperi changed his communication strategy in different roles and contexts, and what was the impact of different audiences on the content of the documents.

A screenshot of the two taxonomies is displayed in Fig. 1. The documents in the corpus were tagged with one or more labels from each taxonomy. This was done semi-automatically with the help of some rules that, looking at the source, title and date of the document, guessed which role De Gasperi was holding at the time and under which circumstances the document was issued. The labels were then manually checked. Table 1 shows the number of documents tagged in the corpus with a document type label

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

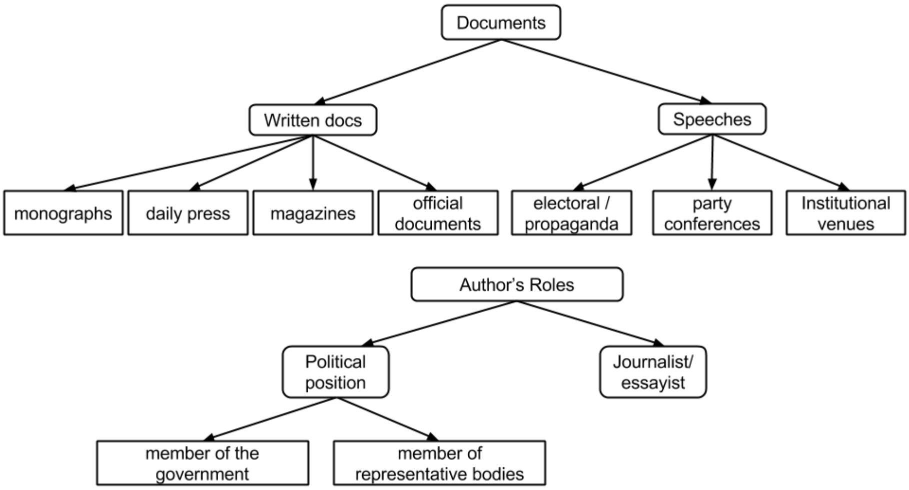

Figure 1
Taxonomies of documents and author's roles.

Table 1
Distribution of labelled documents in the corpus according to the taxonomy in Fig. 1.

DOCUMENT	TYPE	#
Written documents	monographs	2
	daily press	955
	magazines	196
	official documents	436
Speeches	electoral/propaganda	486
	party conferences	186
	institutional venues	421

ROLE	TYPE	#
Political position	government	998
	repr. bodies	161
Journalist/ essayist		1238

(left) and a role type label (right). Around 97% of the corpus is tagged with a document type label, and 86% with a role type. The documents without a tag do not fall under any of the defined categories. Some labels are very likely to appear together, for instance the *daily press* label from the *Written Docs* taxonomy and the *Journalist/Essayist* label as author's role.

4. First project phase

The first part of the project was devoted to the development of tools enabling history scholars to perform corpus-based analyses of De Gasperi's documents, without a specific topic in mind. We rather aimed at making available a range set of NLP functionalities applied to the political domain. The outcome of this effort is the ALCIDE platform (Moretti et al. 2016), which includes among others string-based search, co-occurrence analysis, persons' and place identification and disambiguation, persons' network extraction, keyword analysis.

Table 2
Comparison of NER performance on news and on a subset of De Gasperi corpus

	News			**De Gasperi corpus**		
	P	R	F1	P	R	F1
PER	0.92	0.93	0.92	0.70	0.82	0.76
ORG	0.69	0.60	0.64	0.23	0.39	0.29
LOC	0.78	0.69	0.73	0.50	0.50	0.50
GPE	0.85	0.86	0.85	0.82	0.90	0.86
TOTAL	0.83	0.80	0.82	0.62	0.76	0.69

Table 3
Comparison of PoS tagging performance on news and on a subset of De Gasperi corpus

	News	**De Gasperi Corpus**
	Accuracy	Accuracy
PoS	0.96	0.95

4.1 Evaluation of NLP modules

A first challenge faced during the project was the need to assess the performance of NLP tools on our corpus, since such tools are typically trained on contemporary news and may be unsuitable to process De Gasperi's language. We therefore created two benchmarks to evaluate the performance of the the Named Entity Recognizer (NER) and the PoS-tagger in the TextPro suite (Pianta, Girardi, and Zanoli 2008), which was used to analyse the corpus. For NER, we selected a subset of documents written between 1906 and 1911 (around 9,000 tokens) and we manually annotated persons (PER), organizations (ORG), locations (LOC) and geo-political entities (GPE). Then, we used this gold annotations to evaluate the performance of TextPro NER, which was originally trained on contemporary newspaper stories. Results are reported in Table 2. We compare them with the performance of the tool scored in the EVALITA 2007 campaign (Speranza 2007), when trained and evaluated on a newswire corpus. As expected, the tool shows a drop in performance on De Gasperi's documents, which is rather limited only on GPEs. We noted that the main source of error was the missing names in the gazetteer used by the NER. Geographical names seem to be less affected by this problem because they tend to remain more stable across domains and in different time periods, and they are more likely to be found also in the newswire training data. After a first evaluation, the missing NEs were added to the tool 'white list' so that the analyses obtained after a second run would have a better quality and could be used more reliably by history scholars.

A second evaluation involved the PoS tagger of TextPro, i.e. TagPro, which is used as a basis for keyword extraction and for advanced co-occurrence search. TagPro is based on a supervised approach taking into account a rich set of linguistic features such as prefix, suffix, orthographic and morphological information of the word to be tagged and of the previous and the following one. For the evaluation we manually assigned PoS tags to the same documents used for NER evaluation and we calculated the accuracy of TagPro, which was originally trained and tested on contemporary news yielding 0.96 accuracy (Zanoli and Pianta 2009). As shown in Table 3, on De Gasperi's documents the performance drop is only 1 percentage point in terms of accuracy, showing that the tool is able to analyse texts written in the past century without major issues. Overall, we observed that De Gasperi's language could be analysed with good accuracy also with NLP tools trained on news, and that the strategies available to improve classification (e.g. the use of 'white lists' for NER) could be effectively employed to cope with performance drop.

4.2 Corpus-based analysis and findings

Based on the ALCIDE platform, scholars performed different corpus explorations, resulting in findings that would be hardly achieved without NLP support. Few examples are reported below.

Taxonomies. Using the taxonomies described in Section 3, it was possible to look at lexical differences between propaganda speeches and official documents. It was also possible to compare the content of the documents issued by De Gasperi when he was Prime Minister with those written when he was just an activist of the Christian-Democratic Party, and check if the key-concepts he dealt with vary when he changed his role. For example, in the speeches uttered during party conferences, the most frequent key-concepts are *direttorio*/'board of directors', *direzione*/'leadership', *tripartitismo*/'three-party system' while in the official documents keywords related to the international situation prevail, e.g. *autorità francesi*/'French authorities', *governo militare alleato*/'Allied Military Government', *cooperazione*/'cooperation'.

Paths of exploration. By combining different platform functionalities, it was possible to find new research paths. For example, searching for the frequency of the lemma *libertà*/'freedom', a peak is observed in 1943 (Fig. 2), when freedom was severely limited by the fascist regime. Co-occurrences of the lemma in that year provide a closer look to its context of use: different types of freedom are mentioned, e.g. political, economic, civil, of conscience, together with the expression *giustizia sociale*/'social justice'. This expression is particularly frequent in a document from 1943 named "Political Testament", in which the author outlines his ideas for the economic and political reconstruction after the tragic events of WW2. In order to give strength to his argument, De Gasperi makes reference to several Italian personalities of the past, ranging from the political to the literary and the religious domain, such as Balbo, Manzoni, S. Tommaso, Leone XIII. Indeed, looking at the persons' co-occurrence network (Fig. 3), we observe that Manzoni is not only mentioned in the corpus with other important Italian artists (e.g., Dante, Michelangelo), but also with representatives of the so-called Neo-Guelphism movement (e.g., Cesare Balbo, Gino Capponi), thus playing also a political role in De Gasperi's discourse.

Ingroup-outgroup distinction. The ingroup and outgroup polarization is a peculiarity of political discourse, in that ingroups and their members including allies and friends are generally described in positive terms, while outgroups, enemies and opponents are described in negative terms. This strategy plays a role in the persuasion-reception dimension of political communication, and, according to (Van Dijk 2006), is a central characteristic of all ideologies. De Gasperi's speeches are built along this line, since he tries to share with the audience his point of view, for example by using the first person plural *noi*/'us', so to enhance empathy through the identification between speaker and hearer. Ingroups change over the long political career of De Gasperi: looking at co-occurrences of the pronoun *noi* before the annexation of Trentino by Italy in 1919, the social groups in which he identifies are *trentini*/'people from Trentino',

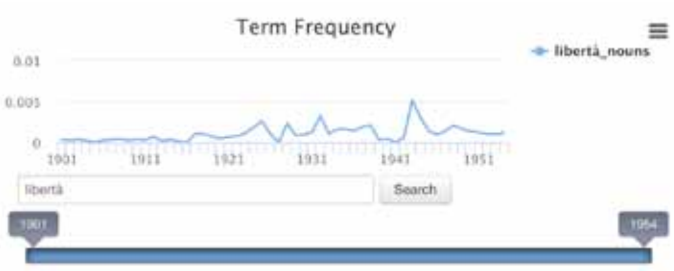

Figure 2
Frequency of *libertà*/'freedom' in the corpus with a peak in 1943.

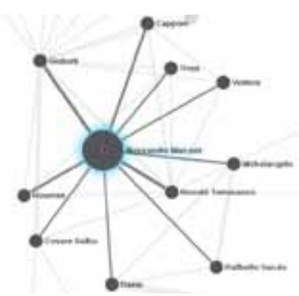

Figure 3
Persons' network of Manzoni

cattolici/'Catholics' and *studenti*/'students'[2]. After World War II, new groups emerge: *italiani*/'Italians', *democratici*/'democrats', *europei*/'Europeans'. Together with this conceptualisation of group identity, De Gasperi marks the distance from the outgroup, delegitimizing his opponents. Looking at co-occurrences of the lemma *nemico*/'enemy' in the first part of his political life (1901-1919), we observe that he considers enemies those opposing suffrage and religion, and supporting Pan-Germanism. In the last part (1945-1954), instead, enemies are those against the Republic, the Constitution but also communists and fascists. This shows clearly a shift in De Gasperi's construction of consensus, driven by changes in his political role and by external events.

5. Second project phase: Ongoing work and future directions

The second project phase, ending in 2017, limits the scope of the project, and has the goal to track De Gasperi's attitude in different contexts and roles. In particular, one of the main issues is his use of the past, present and future dimension in public documents. In this phase, we will make explicit use of the document type and role labels described in Section 3. This study, even if lexically grounded, makes mainly use of tools extracting semantic information. In particular, in order to capture the temporal dimension of texts, we plan to combine three different layers of information: *i)* persons and named events mentioned in the documents, which we will automatically link to Wikipedia[3] and anchor to a period of time, *ii)* verb morphology information (mood and tense) obtained with the Tint tool (Palmero Aprosio and Moretti 2016) and *iii)* (normalized) temporal expressions extracted with the HidelTime tool (Strötgen, Zell, and Gertz 2013). In this context, we will use the set of De Gasperi's documents annotated with temporal information for the EVENTI task at Evalita 2014 (Caselli et al. 2014) as a gold standard. Combining these three information sources together, possibly assigning them different weights, will allow us to assess whether present, past or future is prevalent in each document, and then to aggregate this information comparing different subcorpora (e.g. propaganda speeches versus Parliamentary debates). This will be an important step towards the understanding of De Gasperi's rhetorical strategies. Indeed past, present and future have important argumentative and stylistic functions in political discourse. For example, as stated by Aristotle in the first book of the "Rhetoric" (2010), future

[2] De Gasperi was the leader of the student movement in Tyrol.
[3] `https://bitbucket.org/fbk/twm-lib`

represents the time for political action and thus it is used to influence the behavior of the audience. On the other hand, references to the past are used to highlight the continuity between elements of a collective history (e.g., famous figures of nineteenth-century Catholicism as the ones in Fig. 3) and the present, so to produce a charismatic effect among the public (Shamir, Arthur, and House 1994).

Another research direction we are exploring is motivated by the need to aggregate information related to keywords and provide an overview of the main topics dealt with by De Gasperi over time. Similar studies applied to State of the Union discourse have been presented in (Rule, Cointet, and Bearman 2015). To this purpose, we extended the KD tool for keyword extraction (Moretti, Sprugnoli, and Tonelli 2015) with information from WordNet domains (Magnini et al. 2002) in order to generate a weighted list of domains from the ranked keyword list, extracted from each document (for details see (Moretti, Sprugnoli, and Tonelli 2016). Although researchers in the Digital Humanities community have mainly used topic modelling (Blei 2012) for similar tasks, our approach is easier to interpret, makes use of a well-estabilished domain hierarchy and does not require the user to set a priori the number of domains to be extracted. As an example, we report in Fig. 4 the outcome of a first analysis considering the two top domains for each document issued between 1914 and 1918.

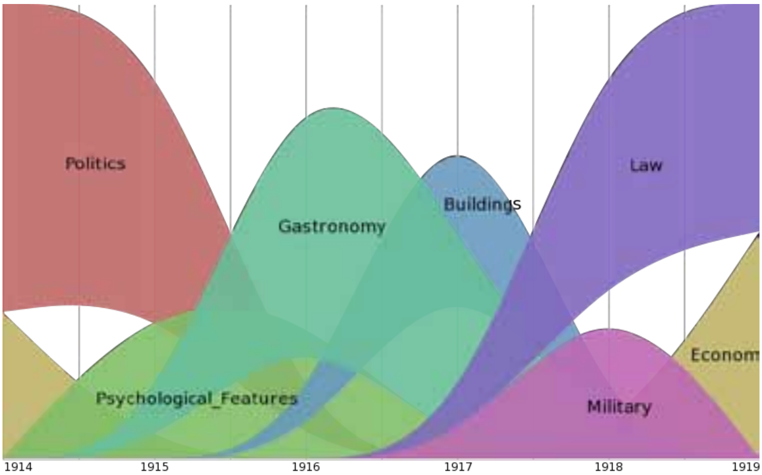

Figure 4
Top domains in De Gasperi's documents from 1914 to 1918

The analysis shows clearly how topics change over the five years, with peaks corresponding to major events in De Gasperi's political life. In the Summer of 1914, the First World War starts in Trentino that was part of the Austrian Empire. Documents published before the entrance into war are mainly about the political situation in Austria and Trentino (`Politics` domain) having, for example, keywords related to Trento municipal election: *rappresentanza proporzionale*/'proportional representation', *campagna elettorale*/'election campaign'. After the beginning of the war, De Gasperi wrote newspaper articles encouraging readers not to lose hope and wishing a quick conclusion of the conflict: in this documents, keywords such as *fede*/'faith', *fratellanza*/'brotherhood', *forza*/'strength' are identified by the `Psychological_Features` domain. Starting from 1915, De Gasperi was appointed delegate of the Refugee Committee and he regularly drew reports from the refugee camps. The main problems addressed in these reports are related to food supply and the living conditions. For this reason, the top domains are `Gastronomy` (e.g., *farina*/'flour', *razione*/'ration') and `Buildings` (e.g., *nuove dimore*/'new homes', *scuola*/'school'). In 1917 the Austrian Parliament reopened: De Gasperi fought to pass a laws to regulate the treatment and to increase subsidy to war refugees. This explains the peaks of the `Law` (with keywords such as *illegittimità*/'illegitimacy', *tribunale amministrativo*/'administrative court') and `Economy` (with keywords such as *rincaro*/'inflation', *sussidio in contanti*/'cash subsidy') domains.

In the next project steps, we will enrich this analysis with information related to language complexity, and investigate the connection between topic, audience and readability level of the documents. This will imply tuning existing readability metrics, which have been developed for contemporary language, to the language used in the first half of the XXth Century.

6. Conclusions

In this work, we described an ongoing project related to the analysis of De Gasperi's public documents with NLP tools. The project foresees two phases: in the first one, which ended in 2015, most effort was devoted to the implementation of an infrastructure allowing the automated analysis of large corpora. This was performed with the help of history scholars, who defined typical research questions and evaluated the suggested solutions, also in terms of usability. The second phase, ending in 2017, has focused on a specific research topic, i.e. how De Gasperi's attitude changed in different contexts, in particular how he referred to past, present and future when addressing different audiences.

The continuous interaction with history scholars has shaped the design choices of the developed tools, in favour of analyses that are easy to interpret compared to more sophisticated outputs. For example, the use of WordNet domains attached to keywords has been preferred over topic modelling. Also approaches using word embeddings to explore the semantic space around given persons or concepts was deemed interesting but difficult to connect with more traditional 'close reading' studies. On the other hand, this inter-disciplinary scenario has given the possibility to combine different analyses in novel ways for knowledge distillation. For example, temporal processing and entity linking are being integrated to convey information about the present, past or future dimension of the documents.

Acknowledgments

We thank the colleagues from the Italian-German Historical Institute at Fondazione Bruno Kessler for their help in analysing De Gasperi's corpus and shaping the research described in this work. The project has been partially supported by Fondazione Cassa di Risparmio di Trento e Rovereto [4] and Fondazione Cassa di Risparmio delle Province Lombarde [5].

References

Aristotle. 2010. *Aristotle: Rhetoric:*, volume 1 of *Cambridge Library Collection - Classics*. Cambridge University Press, Apr.

Balahur, Alexandra, Zornitsa Kozareva, and Andrés Montoyo. 2009. Determining the polarity and source of opinions expressed in political debates. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 468–480. Springer.

[4] http://www.fondazionecaritro.it/

[5] http://www.fondazionecariplo.it/

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

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

Cardie, Claire and John Wilkerson. 2008. Text annotation for political science research. *Journal of Information Technology & Politics*, 5(1):1–6.

Caselli, Tommaso, 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.

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

Cimino, Roberta, Tim Geelhaar, and Silke Schwandt. 2015. Digital Approaches to Historical Semantics: new research directions at Frankfurt University. *Storicamente*, 11(7):1–16.

Cortelazzo, Michele A. and Ivano Paccagnella. 1981. Tipologia del testo politico. In D. Goldin, editor, *Teoria e analisi del testo*. Padova, Cleup, pages 205–220.

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

De Gasperi, A. 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.

De Gasperi, A. 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.

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

Delmonte, Rodolfo, Daniela Gîfu, and Rocco Tripodi. 2013. Extracting opinion and factivity from italian political discourse. In *Proceedings 10th International Workshop NLPCS, Natural Language Processing and Cognitive Science, Marseille*, pages 162–176.

Desideri, Paola. 1984. *Teoria e prassi del discorso politico*. Bulzoni.

Gold, Valentin, Mennatallah El-Assady, Tina Bögel, Christian Rohrdantz, Miriam Butt, Katharina Holzinger, and Daniel Keim. 2015. Visual Linguistic Analysis of Political Discussions: Measuring Deliberative Quality. *Digital Scholarship in the Humanities*.

Guerini, Marco, 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*. Springer, pages 86–98.

Hirst, Graeme, 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.

Howarth, David R., Aletta J. Norval, and Yannis Stavrakakis. 2000. *Discourse theory and political analysis. Identities, hegemonies and social change*. Manchester University Press,.

Magnini, Bernardo, Carlo Strapparava, Giovanni Pezzulo, and Alfio Gliozzo. 2002. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8(4):359–373.

Moretti, Franco. 2013. *Distant Reading*. Verso, London.

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

Moretti, Giovanni, 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)*, pages 198–203.

Moretti, Giovanni, Rachele Sprugnoli, and Sara Tonelli. 2016. KD Strikes Back: from Keyphrases to Labelled Domains Using External Knowledge Sources. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, pages 216–221.

Palmero Aprosio, Alessio and Giovanni Moretti. 2016. Italy goes to Stanford: a collection of CoreNLP modules for Italian. *ArXiv e-prints*, September.

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

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

Rule, Alix, Jean-Philippe Cointet, and Peter S Bearman. 2015. Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. *Proceedings of the National Academy* *of Sciences*, 112(35):10837–10844.

Shamir, Boas, Michael B. Arthur, and Robert J. House. 1994. The rhetoric of charismatic leadership: A theoretical extension, a case study, and implications for research. *The Leadership Quarterly*, 5(1):25–42.

Speranza, Manuela. 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.

Stranisci, Marco, Cristina Bosco, Viviana Patti, and Delia Irazu Hernández Farıas. 2015. Analyzing and annotating for sentiment analysis the socio-political debate on #labuonascuola. In *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*, pages 274–279.

Strapparava, Carlo, 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.

Strötgen, Jannik, Julian Zell, and Michael Gertz. 2013. Heideltime: Tuning english and developing spanish resources for tempeval-3. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, pages 15–19, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Sudhahar, Saatviga, Giuseppe A. Veltri, and Nello Cristianini. 2015. Automated analysis of the US presidential elections using Big Data and network analysis. *Big Data & Society*, 2(1):1–28.

Thomas, Matt, 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.

Van Dijk, Teun A. 2006. Ideology and discourse analysis. *Journal of political ideologies*, 11(2):115–140.

Vinciguerra, Antonio. 2016. La delegittimazione dell'avversario politico nei discorsi di Alcide De Gasperi per la campagna elettorale del 1948. *Lingue e Linguaggi*, 17:277–296.

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

Wodak, Ruth. 1989. *Language, power and ideology: Studies in political discourse*, volume 7. John Benjamins Publishing.

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

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

Zanoli, Roberto and Emanuele Pianta. 2009. A multistage PoS-tagger at the EVALITA 2009 PoS-tagging Task. In *Proceedings of the EVALITA 2009 Workshop on Evaluation of NLP Tools for Italian*, Reggio Emilia, Italy.