Multi-dimensional analysis of political language

Daniela Gîfu[1], Dan Cristea[1,2],

[1]„Alexandru Ioan Cuza“ University, Faculty of Computer Science,
16, General Berthelot St., 700483, Iaşi
{daniela.gifu, dcristea}@info.uaic.ro
[2]Institute for Theoretical Computer Science, Romanian Academy - Iaşi branch,
2, T. Codrescu St., 700481, Iaşi

Abstract. This paper presents a method for the valuation of discourses from different linguistic perspectives: lexical, syntactic and semantic. We describe a platform (Discourse Analysis Tool – DAT) which integrates a range of language processing tools with the intent to build complex characterisations of the political discourse. The idea behind this construction is that the vocabulary and the clause structure of the sentence betray the speaker's level of culture, while the semantic classes mentioned in a speech characterises the speaker's orientation. When the object of study is the political discourse, an investigation on these dimensions could put in evidence features influencing the electing public. The method is intended to help both political speakers to improve their discourse abilities, by comparing their speeches with those of personalities of the public life in the past, and the public at large by evidencing hidden aspects of the linguistic and intellectual abilities of their candidates.

Keywords: political language, linguistic coverage, treebanks, syntactic structure, discourse structure, semantic classes.

1 Introduction

The motivation for our study relies on the need for objectivity in the interpretation of the political language situated at the intersection of three important symbolic spaces: the political space, the public space and the communicational space, as well as on the need to measure to what extend a discourse can influence its direct receptor, the electorate and in what ways. The current approaches in analysing the political language are based on Natural Language Processing (NLP) techniques designed to investigate lexical-semantic aspects of the discourse. The domain of NLP includes a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications [5].

In this paper, we describe a platform (Discourse Analysis Tool – DAT) which integrates a range of language processing tools with the intent to build complex characterisations of the political discourse. A linguistic portrait of an author is drawn by putting together features extracted from the following linguistic layers: lexicon and morphology, semantics and syntax. Among the resources used for the study of natural

language syntax, of a tremendous importance are the treebanks, large collections of sentences annotated by human experts at syntactic structures. We used an interactive graphical tool which allowed easy annotation, visualisation and modification of syntactic trees, initially obtained as a result of an automatic parsing process.

The paper is structured as follows. Section 2 shortly describes the previous work. Section 3 discusses the lexical, semantic and syntactic features having rhetorical values and section 4 presents a platform for multi-dimensional political discourse analysis. Next, section 5 discusses an example of comparative analysis of discourses very distant in time, elaborated during elections. Finally, Section 6 highlights interpretations anchored in our analysis and presents conclusions.

2 Previous work

As we will see, one aspect of the platform that we present touches a lexical-semantic functionality, which has some similarities with the approach used in Linguistic Inquiry and Word Count (LIWC), an American product used on the American elections in 2008. There are, however, important differences between the two platforms. LIWC-2007[1] is basically counting words and incrementing counters associated with their declared semantic classes.

A previous version of DAT performs part-of-speech (POS) tagging and lemmatization of words. The lexicon contains a collection of lemmas (7700) having the POS categories: verb, noun, adjective and adverb. In the context of the lexical semantic analysis, the pronouns, numerals, prepositions and conjunctions, considered to be semantically empty, have been left out. Our version includes 30 semantic classes, chosen to fit optimally with the necessities of interpreting the political discourse, two of them being added recently.

The second range of differences between the two platforms regards the user interface. In DAT, the user is served by a friendly interface, offering a lot of services: opening one or more files, displaying the file/s, modifying/editing and saving the text, functions of undo/redo, functions to edit the lexicon, visualization of the mentioning of instances of certain semantic classes in the text, etc. Then, the menus offer a whole range of output visualization functions, from tabular form to graphical representations and to printing services. Finally, another important development for the semantic approach was the inclusion of a collection of formulas which can be used to make comparative studies between different subjects. A special section of the lexicon includes expressions. An expression is defined as a sequence: <root-list> => <semlist>, in which <root-list> is a list of roots of words, therefore each optionally followed by the '*' sign [2] report similar approaches of human validation.

Completely new in DAT in comparison with other approaches is the facility for analyzing the political discourse from the syntactic point of view. DAT helps the user to identify and count relations between different parts of speech, to put in evidence patterns of use at the semantic and syntactic level, etc.

1 www.liwc.net

3 Lexical, semantic and syntactic features with rhetorical values

The study of political language should necessarily be rooted in an interdisciplinary approach, in which the sciences of the rhetoric, communication and politology cooperate with linguistics. The use of language in politics has a "sanctifying" role in the tentative to gain the trust of the electorate. The deviation from the rules of language construction can be intended, in which case it is commanded by some rhetorical or aesthetic goals, expressing thus strategic aspects of the production of discourse, or can represent social or cognitive characteristics of the speakers, "as memory limits, lacks in culture, etc." [9].

3.1 The lexical-semantic perspective

The political speaker is determined to collect empathy and to convince the public. Yet, placing himself within the general limits of the political goals, very often a skilful politician studies the public for fixing the type of vocabulary and the message to be delivered. He might exploit connections between more daring ideological categories (as is for instance the class `nationalism`) and those generally accepted (for instance, belonging to the classes `social`, `work`, `home`). The present day political language puts in value the virtues of the metaphor, its qualities to pass abruptly from complex to simple, from abstract to concrete, imposing a powerful subjective, i.e. emotional dimension to the discourse (the class `emotional`). Nonetheless, the political metaphor may loose the virtues of poetical metaphor, becoming vulgar (the class `injuries`).

But often words have multiple senses. Disambiguating their senses in context proved to be a necessary, although difficult, phase. Among the number of senses words are registered with in dictionaries we have retained only those considered relevant for the semantic classes selected. As such, each semantic class is mapped against a lexicon of word senses. Thus, the disambiguation task resides in using the context of a word occurrence for making a forced choice among the retained connotations. For sense disambiguation we have used the classical bag-of-words paradigm. The following preliminary steps have been followed to prepare the corpus against which word sense have been disambiguated:

1. A number of semantic classes have been retained, considered relevant for the type of discourses we have concentrated on: political (see section 4 for a list of these classes).
2. For each of these semantic classes, we have selected a number of words (actually lemmas), to each of them retaining the appropriate, intended, sense for the semantic class at hand.
3. The selected senses have been looked for in the electronic version of the biggest dictionary for Romanian language, eDTLR [1]. This dictionary includes for each sense of each word a great number of citations selected from writings of Romanian authors.
4. The citations attributed to the selected senses of the selected words have been copied from eDTLR and processed (by lemmatising and eliminating the stop words)

in order to build the “master” sense vectors to be used in further word sense comparisons.

The interpretation of word senses in our approach follows a perspective in which words of a document are having a narrow semantic spectrum. This means that all occurrences of the same word in the same text are supposed to have the same sense. As such, when a focus word *w* is to be decided its sense in the text, all words belonging to its occurrences (windows of a fixed size around the occurrences of *w*) are collected to assemble a test vector, which is compared against the master vectors of the recorded senses, by using a simplified-Lesk algorithm [4], [3].

3.2 The syntactic perspective

Regarded as one of the most developed branches of semiotics, syntactic analysis aims at studying the relations between signs and the logical and grammatical structure at the sentence level [6]. Outside the sentence, rhetorical relations identify particular interdependencies that can hold between adjacent spans of text. Based on relations, the rhetorical schemes define patterns in which a particular span of text can be analyzed. But if we remain at the level of syntax, the text is formed out of an ordered sequence of language signs which are governed by a set of combinatorial rules [8].

From this perspective, the syntactic analysis of political language aims at identifying patterns or idiosyncrasies in the written phrase of an author: the repeated use of some syntactic relations or linking expressions, their characterization as coordination or subordination, degree of breaking the grammar rules, etc.

4 A platform for multi-dimensional political discourse analysis

The Discourse Analysis Tool (currently at version 2) considers the political discourse from two perspectives: lexical-semantic and syntactic. We describe shortly our platform (Discourse Analysis Tool – DAT) which integrates a range of language processing tools with the intent to build complex characterisations of the political discourse. The concept behind this method is that the vocabulary used by a speaker opens a window towards the author’s sensibility, his/her level of culture, her/his cognitive world, and, of course, the semantic spectrum of the speech, while the syntax may reveal the level of culture, intentional persuasive attitudes towards the public, etc. Some of these means of expression are intentional, aimed to deliver a certain image to the public, while others are unintentional. Figure 1 shows a snapshot of the interface showing a semantic analysis, during a working session.

To display the results of the semantic analysis, the platform incorporates two alternative views: graphical (pie, function, columns and areas) and tabular (Microsoft Excel compatible).

In DAT, the user has an easy-to-interact interface, offering a lot of services: opening of one or more files, displaying the file/s, modifying/editing and saving the text, functions of undo/redo, functions to edit the lexicon, visualizing the mentioning of instances of certain semantic classes in the text, etc. Then, the menus offer a whole

range of output visualization functions, from tabular form to graphical representations and to printing services.

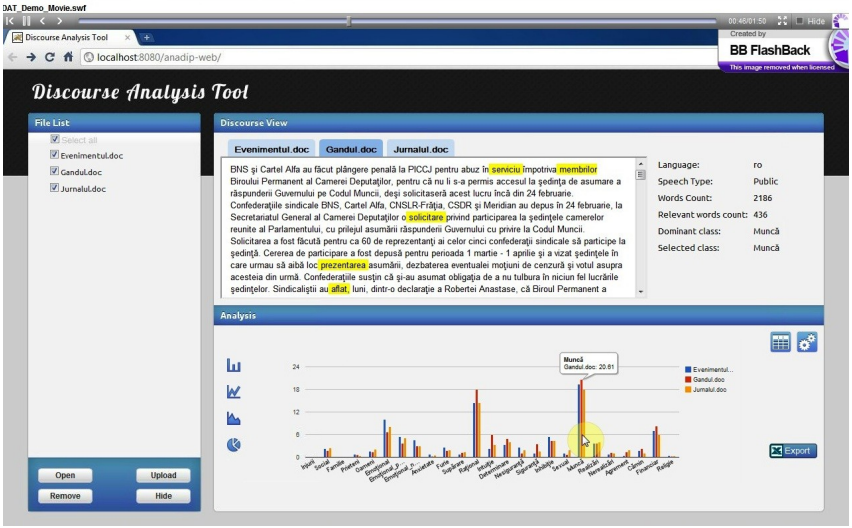

Fig. 1. The DAT interface: in the left window appear the selected files, in the middle window - the text from the selected file - and in the right window, information about the text (language, word count, dominant class, etc.). Bellow it appears a plot selected from different graphical types. By selecting a specific class in the middle window, all words assigned to that class are highlighted in the text.

The vocabulary of the platform covers 30 semantic classes (`swear`, `social`, `family`, `friends`, `people`, `emotional`, `positive`, `negative`, `anxiety`, `anger`, `sadness`, `rational`, `intuition`, `determine`, `uncertain`, `certain`, `inhibition`, `perceptive`, `see`, `hear`, `feel`, `sexual`, `work`, `achievements`, `failures`, `leisure`, `home`, `financial`, `religion`, `nationalism`), considered to fulfil optimally the necessity of interpreting the political discourse in electoral contexts. The hierarchy of these categories preserves the structure of a tree.

Linguistic processing begins by tokenization, part of speech tagging and lemmatization. Only the relevant words count in establishing the weights of different semantic classes, as given by the lexicon. Since the lexicon maps senses of words to different semantic classes, depicting a semantic radiography of the text should follow a phase in which words are sense disambiguated. As mentioned already, our hypothesis is that in all the occurrences of a multi-sense word in a text the word displays the same sense. This hypothesis facilitates the disambiguation process, because all contexts of occurrence of a word participate in the disambiguation and that sense is selected which maximises a bag-of-words-like analysis among all recorded possible choices.

In response to the text being sent by the user, the system returns a compendium of data which includes: the language of the document, the number of words, and the type

of discourse detected, a unique identifier (usually the file name), a report of the lexical-semantic analysis and a report of the syntactic analysis.

Our interest went mainly in determining those political attitudes able to influence the voting decision of the electorate. But the system can be parameterised to fit also other conjunctures: the user can define at will her/his semantic classes and the associated lexical, which, as indicated, are partially placed in a hierarchy. As an example, for the lemma *lucrător* (worker), the following classes are assigned: 2 = *social* and 5 = *people.* The class *people*, is a subclass of the class *social*. Whenever an occurrence belonging to a lower level class is detected in the input file, all counters in the hierarchy, from that class to the root, are incremented. In other words, the lexicon assigned to superior classes includes all words/lemmas of its subclasses.

5 A comparative study

5.1 The corpus

The corpus used for our investigation was configured to allow a comparative study over the discursive characteristics of two political leaders, both embracing liberal convictions, although in quite distant periods. The first one, I. C. Brătianu, is known as having led the basis of the liberal ideology in Romania, one of the most complex personalities of the Romanian history. For the second political actor we have chosen a modern liberal party leader: Crin Antonescu, right now in power. This way, we wanted to put on the balance two styles of political discourse, which, although representing the same orientation, are quite distant in many other respects, since the two political leaders are separated by one century and a half of dramatic history: the union of the three Romanian provinces, wars, economical crises, changes of political regimes, cultural and linguistic developments, etc.

For the elaboration of preliminary conclusions over the two elections processes, conducted in December 1858 [7] and November 2009, in Romania, we collected, stored and parsed manually and automatically, political texts published by four national publications having similar profiles[2]. The corpus includes a collection of 1548 political sentences (units), each containing one or more clauses.

5.2 The lexical-semantic analysis

We present below a chart with two streams of data, representing the political texts in electoral context between the two liberal leaders mentioned above. Our experience shows that an absolute difference value below the threshold of 0.5% should be considered as irrelevant and, therefore, ignored in the interpretation. Apart from simply computing frequencies, the system can also perform comparative studies. The assessments made are comprehensive over the selected classes because they represent averages on collections of texts, not just a single text. So, the graphical representation in Figure 2, in which the present day politician (in red), is compared against the

2 National newspapers of general informations, are presented as a tabloid with a circulation of tens of thousands of copies per edition: Românul (19th century), Evenimentul zilei, Gândul and Ziua (our days).

outstanding politician of the past (in blue) should be interpreted as follows: Ion C. Brătianu's was interested more on Romanian specific aspects (the *nationalism* and *family* classes) uttered in an emotional tone (the *positive* class) than Crin Antonescu, whose discourse had an argumentative (the *rational* class) attitude, being also very much centred on labour aspects (the *work* class).

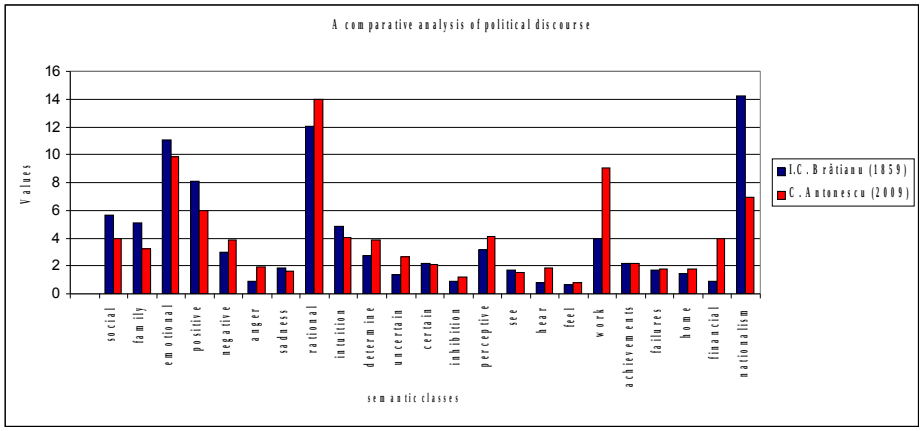

Fig. 2. The average differences in the frequencies for each parent class (>0.5%) after processing political discourses, between Ion C. Brătianu and Crin Antonescu.

5.3 The syntactic analysis

In order to proceed with the syntactic analysis, the text bodies were annotated with syntactic information, in XML. Two sources of information have been used, involving manual (200 units) and automatic (600 units) annotation for each political actor monitored. The manually annotated treebanks (see Table 1) included the adjectival relations with the respective frequencies.

Table 1 Occurrence of adjectival relations for political speeches corresponding to the two electoral contexts

Adjectival type of relation (abrev.)	Ion C. Brătianu 1859		Crin Antonescu 2009	
	Number	Percentage	Number	Percentage
a.adj.	536	8,1%	600	8,5%
a.subst.	312	4,7%	280	4,0%
a.adv.	96	1,4%	112	1,6%
a.vb.	272	3,9%	144	2,0%
Total	1216	18,10%	1136	16,10%

We can notice that adjectival structures make up 18.1% of all syntactic relations at the first author and 16.1% – at the second. The adjectives not only bring a contextual, albeit new, information, but enhance the enounce by detailing it and developing it. When the adjectival groups is placed in the thematic position it's role is emphatic, usually associated with a particular tone, but, generally, it does not change the content of the message. The relation reveals a certain taste for belletrist culture of the authors.

6 Conclusions

It is clear that some of the differences at the level of discourse which we have evidenced as differentiating the two political actors should be attributed only partially to idiosyncratic rhetorical styles, because they have also historic explanations. Moreover, speeches of many public actors, especially today, are the product of teams of specialists in communication and, as such, conclusions regarding their cultural universe, for instance, should be uttered with care. We believe that the platform helps to outline distinctive features which bring a new, and sometimes unexpected, vision upon the discursive characteristics of political authors or columnists. In the future, we will extend the analysis of political discourse to the public sphere. The collection of manually annotated texts is only at beginning, a starting point for an efficient automatic annotation. We'll manually correct all the automatically annotated texts, improving thus the behaviour of the parser. Another line to be continued regards the evaluation metrics, which have not received enough attention till now. We are currently studying other statistical metrics able to give a more comprehensive image on different facets of the political discourse.

Acknowledgments: The DAT platform has been developed by Mădălina Spătaru from the Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iaşi. In order to perform this research the first author received financial support from the POSDRU/89/1.5/S/63663 grant.

References

1. Cristea, D.: "eDTLR - an accomplished project", a conference at the Department of Sciences of the "Alexandru Ioan Cuza" University, May 17th 2010, Iaşi, Romania.
2. Gîfu, D., Cristea, D.: "Computational Techniques in Political Language Processing: AnaDiP-2011" in J.J. Park, L.T. Yang, and C. Lee (Eds.): FutureTech 2011, LNCS, Part II, CCIS 185, Springer, Heidelberg, 188–195 (2011).
3. Kilgarriff, A and Rosenzweig, J: "English SENSEVAL: Report and Results". In Proceedings of the 2nd International Conference on Language Resourcesand Evaluation, LREC, Athens, Greece (2000).
4. Lesk, M.: "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone". In SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation, New York, NY, USA. ACM, 24-26 (1986).
5. Liddy, E.D.: "Natural Language Processing" in Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc. (2001).
6. Mann, W.C., Thompson, S.A.: „Rhetorical Structure Theory: Toward a functional theory of text organisation", Text 8(3), 243-281 (1988).
7. Marinescu, G., Grecescu, C.(ed.): "Ion C. Brătianu. Acte şi cuvântări" (Documents and speeches), vol. I – part I (June 1848 = December 1859), Cartea Românească, Bucharest, 228-237 (1938).
8. Plett, H. F.: „Ştiinţa textului şi analiza de text" (The science of text and text analysis), Ed. Univers, Bucharest, 55 (1983).
9. van Dijk, Teun A.: "Textual Structures of News in the Press. Working notes", University of Amsterdam, Department of General Literary Studies, Section of Discourse Studies, 14 (1972).