PUBLIC TEXT CATEGORIZATION

DANIELA GÎFU[1], DAN CRISTEA[1, 2]

[1] *"Alexandru Ioan Cuza" University, Faculty of Computer Science, Iași – România*
[2] *Institute for Theoretical Computer Science, Romanian Academy - Iași branch.*

{daniela.gifu, dcristea}@info.uaic.ro

Abstract

To analyze public discourse (media, politic, religious, etc.), means to analyze two dimensions: rational and emotional. This paper introduces an important natural language processing (NLP) problem, text categorization from the perspective of public language. Classification or categorization is the task of assigning words from a text corpus to two or more classes. The goal in text categorization is to classify the theme of a text, but, also, the dominant tonalities in a discourse. Our sets of semantic classes (33 for this version) are the extracts from many dailies, which we monitored in time (especially, in different crisis contexts). Here we present a computational tool, *Discourse Analysis Tool* (DAT), based on natural language processing (NLP) techniques for the interpretation of the public discourse. The idea behind it is that the vocabulary betrays the speaker's orientation (emotional or rational). Practically, the receptor identifies with the transmitter (journalist, politician, priest and so on), who becomes the legitimate voice of common ideals. When the object of study is the public discourse in print media, an investigation on these dimensions could put in evidence features influencing the auditory. Our purpose was to develop a computational platform able to offer to researchers in the humanities or social sciences, to the public at large (interested to consolidate their options before any public confrontation), and, why not, even to public speakers themselves, the possibility to measure different parameters of a written public discourse.

1. *Introduction*

Public discourse can be characterized from a rhetorical perspective, depending on its specific strategies: orientation to change opinions or to determine action, ratio between rational (*logos*) and emotional (*pathos*), etc. The main directions of research of public language are content analysis, with quantitative investigations of vocabulary (key words, frequent words) and rhetorical-pragmatic analysis of discursive strategies (presence of the person I, preference for vague statements, generics, etc.). In USA, the tradition of quantitative analysis is rather strong, starting from Lasswell (Lasswell, 1936); in Europe the interest grew more for discursive-rhetoric analysis. The situation, already described by Desideri (1984a: 11-13), hasn't changed very much in the meantime. On the other hand, the American analyses are often neutral, technical, comparative, while the European analysis (especially the model CDA[1]) has a critic component and a strong enough ethicist.

[1] "Critical theories, thus also CDA, are afforded special standing as guides for human action. They are aimed at producing «enlightenment and emmancipation». Such theories seek not only to describe and explain, but also to root

The current approaches in analyzing the public language are based on Natural Language Processing (NLP) techniques designed to investigate syntactic, lexical-semantic and pragmatic 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 (Liddy, 2001). To be able to interpret correctly a public phenomenon we must take into account the past events. Each public event "is an action that always tends to alter a pre-existing condition" (Perelman and Tyteca, 1970: 72). We can consider a public speech like an "aggressor", because it promotes and supports the programs and values of a group, capable to answer the auditor’s expectations. Receptors demand from speakers the logicians’ opened mind, the philosophers’ deep meditation, the poets’ metaphoric expression, the jurists’ bright memory, tragedians’ penetrated voice and, I’d say, a famous actor’s gestures (Cicero, 1973: 51). The basic assumption in the public analysis is that any text isn't merely a string of signs placed randomly. Any group of signs is hierarchically organized, the signs can define various informational and interaction relations (Fox, 1987). Our analysis is meant to highlight the relevance and to understand different forms of communication, as captured by the print media in different contexts. Print media discourse may mean that the actual object (theme, context of a word, sentence) sometimes appears incoherent, incomplete, etc., as more general rules and principles, which does not mean it cannot be interpreted, at least in part, its purpose being to convince. In fact, the deviation in terms of rules of construction may be, on one hand, deliberate, so as to achieve specific rhetorical or aesthetic purposes, or, on the other hand, may be an expression of social and cognitive characteristics of those who use language such as memory limitations, the strategic aspects of speech production, etc.

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 characterizations of the public discourse. A linguistic portrait of an author is drawn by putting together features extracted from the following linguistic layers: lexicon and morphology (richness of the vocabulary, rare co-occurrences, repetitions, use of synonyms, coverage of verbs’ grammatical tenses, etc.) and semantic (semantic classes used).

The paper is structured as follows. Section 2 shortly describes the previous work. Section 3 discusses the lexical and semantic features having rhetorical values and section 4 presents the platform for multi-dimensional public 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

The aim of an interdisciplinary approach such as analyzing the language of public speeches is to define and explain different discursive contexts (political, social,

out a particular kind of delusion. Even with differing concepts of ideology, critical theory seeks to create awareness in agents of their own needes and interests” (Wodak, 2006).

economic, etc.), in this case, reflected in the print media. The studies in this direction have mainly concentrated on three tasks: the first had to do with a cognitive side and, often, with an emotional side, of how humans acquire, produce, and understand language. The second aimed at understanding the relationship between the linguistic utterance and the world, and the third – at understanding the linguistic structure of the language as a communication device. Linguistics has usually treated language as an abstract object which can be accounted for without reference to social or political concerns of any kind (Romaine, 1994). Noam Chomsky (1968) and a whole range of scholars following him have given incentive ideas over topics that are placed on the immediate horizon today, their perspective on structural linguistics being at the origin of a whole range of theories in modern linguistics. From a different perspective, another reference model for communication theory was formulated by Habermas[2] (Stevenson, 1995). His thesis is that the public domain, in which we communicate, comes increasingly under the control of private business interests, either through direct and interactive forms, such as phone or Internet, or by means of mass communication, centrally controlled, such as audiovisual and print media.

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[3] 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 (9500) 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 current version includes 33 semantic classes, chosen to fit optimally with the necessities of interpreting the public discourse, five of them being added recently (`failures`, `nationalism`, `moderation`, `firmness`, `spectacular`). The second range of differences between the two platforms regards the user interface. In DAT, the user is served by a friendly interface, offering several 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 authors. 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. Gîfu and Cristea (2011) report similar approaches of human validation.

[2] Habermas's thesis is applied to the evolution of the British press, which said that industry trade press led to two types of journalism: quality journalism (for a small audience, educated and informed and with great power to attract publicity) and scandal journalism (for a group with low incomes and low power to attract advertisers).

[3] www.liwc.net.

3. Lexical and semantic features with rhetorical values

The use of language in public sphere has a “sanctifying” role (Edelman, 1964/1985) in the tentative to gain the trust of the auditor. The object of language could seem sometimes incoherent, unfinished, deprived of sense, etc., if confronted against general rules or principles of the language, but it can still be deciphered and function adequately. 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. (van Dijk, 1972). The trajectory of rhetoric's (as a theory of discourse persuasion) has been intimately interlinked with the public discourse since Antiquity till our days. The only means to impose yourself in the public life is to convince by spreading your word. Today, the art of rhetorical discourse is understood only in correlation with performance, by combining in a highly elaborate way four ingredients: be rational, have ideas, master the language, and use an adequate style. It is extremely difficult to make an objective evaluation of this magic mélange of methods, but at least some parts of it can be measured. It is what we try to do in this research.

3.1. The context

The public discourse is, especially, a contextual discourse. Therefore, the analysis of public discourse, spoken or written, involves the analysis of the context in which it is transmitted. It becomes a context of speech, the whole reality that surrounds a sign, a verbal act or a discourse, as a “science” of speakers, physical presence and activity. We distinguish three context types (Coșeriu, 2004: 319-324):

1. the *idiomatic* context, created by the language itself, as a background of the speech. In other words, inside of the idiomatic context, each word meaning is defined in a smaller context, which is its field of meanings. Thus, a name of color, such as *portocaliu* (orange), has a meaning in relation to other color names of the same language (e.g. *roșu* (red), *albastru* (blue), etc.).

2. the *verbal* context is the speech itself. For each sign and discourse sequence, it becomes “the verbal context” not only what was said before, (Bally, 1950:43-44), but, also, what will be said in the same discourse. Thus, *Crin’s bank account* and *the bank account in Switzerland* include contextual elements, highlighting the significance of the phrase *bank account*.

3. the *extra-verbal* context consists of all non-linguistic circumstances that are directly perceived by the speakers. We distinguish several subtypes: *physical* (things that are in the visual sight of the speakers or to which a sign adheres), *empirical* (objective things, which are known by those who speak in a certain place and moment, although they are not in the sight of the speakers), *natural* (totality of possible empirical contexts), *occasional* (occasional speech), *historical* (historical circumstances known to the speakers), *cultural* (cultural tradition of a community). Given their importance in establishing the semantic classes and, also, the correct interpretation of each entry in these classes, in our analyses we specify, particularly, this extra-verbal context,

especially the last three. Thus, the global economic crisis is an occasional extra-verbal context which gives a strong significance for the public discourse, from 2007 to present.

3.2. The lexical-semantic perspective

The speaker in a public space is determined to collect empathy and to convince the auditor. Yet, placing himself within the general limits of the public goals, very often a skilful speaker 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, achievements). The present day public 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 public metaphor may lose the virtues of poetical metaphor, becoming vulgar (the class injuries).

But often, words have multiple senses. Among the number of senses words are registered within 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: the public discourse (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.

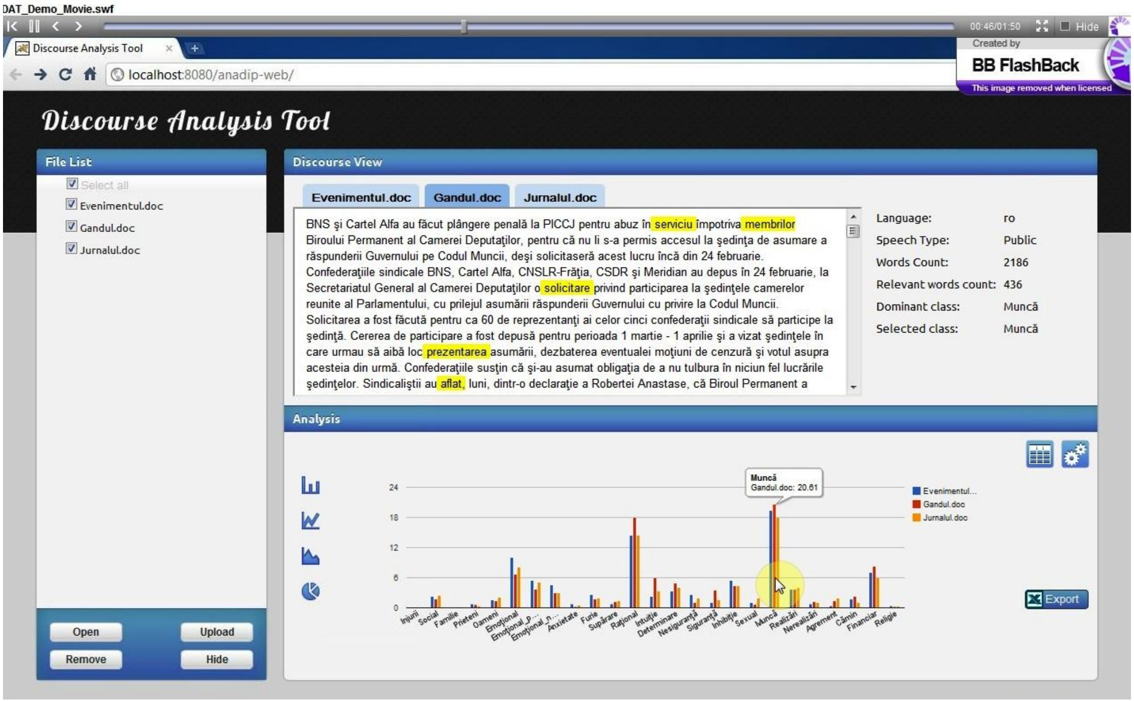
3. The selected senses have been looked for in the electronic version of the biggest dictionary for Romanian language, eDTLR (Cristea et al., 2007). This dictionary includes for each sense of each word a great number of citations selected from writings of Romanian authors.

4. The citations attributed to the selected senses of the selected words have been copied from eDTLR and processed (by lemmatizing 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 sentence 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 (Lesk, 1986), (Kilgarriff, Rosenzweig, 2000).

4. *The DAT platform*

The *Discourse Analysis Tool* (DAT, currently at version 3) considers the public discourse from two perspectives: lexical and semantic. We describe shortly our platform which integrates a range of language processing tools, with the intent to build complex characterizations of the public 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 lexical-semantic analysis, the platform incorporates two alternative views: graphical (pie, function, columns and areas) and tabular (Microsoft Excel compatible).



☑ Evenimentul.doc
☑ Gandul.doc
☑ Jurnalul.doc

BNS şi Cartel Alfa au făcut plângere penală la PICCJ pentru abuz în serviciu împotriva membrilor Biroului Permanent al Camerei Deputaţilor, pentru că nu li s-a permis accesul la şedinţa de asumare a răspunderii Guvernului pe Codul Muncii, deşi solicitaseră acest lucru încă din 24 februarie. Confederaţiile sindicale BNS, Cartel Alfa, CNSLR-Frăţia, CSDR şi Meridian au depus în 24 februarie, la Secretariatul General al Camerei Deputaţilor o solicitare privind participarea la şedinţele camerelor reunite al Parlamentului, cu prilejul asumării răspunderii Guvernului cu privire la Codul Muncii. Solicitarea a fost făcută pentru ca 60 de reprezentanţi ai celor cinci confederaţii sindicale să participe la şedinţă. Cererea de participare a fost depusă pentru perioada 1 martie - 1 aprilie şi a vizat şedinţele în care urmau să aibă loc prezentarea asumării, dezbaterea eventualei moţiuni de cenzură şi votul asupra acesteia din urmă. Confederaţiile susţin că şi-au asumat obligaţia de a nu tulbura în niciun fel lucrările şedinţelor. Sindicaliştii au aflat, luni, dintr-o declaraţie a Robertei Anastase, că Biroul Permanent a

Language:	ro
Speech Type:	Public
Words Count:	2186
Relevant words count:	436
Dominant class:	Muncă
Selected class:	Muncă

Figure 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, a plot chosen from a range of graphical styles is displayed. By selecting a specific class in the middle window, all words assigned to that class are highlighted in the text.

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.

The vocabulary of the platform covers 33 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, moderation, firmness, spectacular), considered to fulfill optimally the necessity of interpreting the public discourse in different contexts. Some of these categories are placed in a hierarchical relation.

Linguistic processing begins by tokenization, part of speech tagging and lemmatization. Only the words belonging to the lexicon are considered relevant and therefore count in establishing the weights of different semantic classes. 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 maximizes 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), and a report of the lexical-semantic analysis.

Our interest went mainly in determining those discursive attitudes able to influence the audience decision. But the system can be parameterized 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 *jurnalist* (journalist), 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*

Given that political discourse is a type of public discourse, we propose below a comparative analysis starting from political texts of two liberal leaders that we have found in print media.

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 to have led the basis of the liberal ideology in Romania, one of the most complex personalities of the Romanian history. Patriotic values were very important in influencing the auditory in the 19th century. The main theme of the speech is integrated in the class nationalism. The second political actor was chosen based on similar criteria: Crin Antonescu, a contemporary liberal political leader. Amid a permanent crisis (economic, political, moral, etc.), the Romanian political discourse contains many arguments for improving living standards. The main theme of the speech is integrated in the class `work`. We are, this way, putting on the balance two styles of political discourse that are distant in time by one century and a half, interval which witnessed many changes in the state (the union of the Romanian provinces, wars, economical crises, etc.). For the elaboration of preliminary conclusions over the two Romanian elections processes, conducted in December 1858 (Marinescu and Grecescu, 1938) and November 2009, we collected, stored and parsed manually and automatically, political texts published by four national publications having similar profiles[4]. This corpus includes a collection of 1548 political sentences/phrases (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. To exemplify, one type of graphics considered for the interpretation was the one-to-one difference, as given by Formula (1), included in the DAT Mathematical Functions Library:

$$Diff_{x,y}^{1-1} = average(x) - average(y) \quad \mathbf{(1)}$$

where *x* and *y* are two streams; *average(x)* and *average(y)* are the average frequencies of *x* and *y* over the whole stream, and the difference is computed for each selected class. Since a difference can lead to both positive and negative values, these particular graphs should read as follows: values above the horizontal axis are those prevailing at the candidate Brătianu versus the candidate Antonescu, and those below the horizontal axis show the reverse prominence. A zero value indicates equality.

So, the graphical representation in Figure 2, in which the present day politician is compared against the outstanding politician of the past should be interpreted as follows: Ion C. Brătianu's was interested more on Romanian specific aspects (the `nationalism` class) uttered in an emotional tone (the `positive` class) than Crin Antonescu, whose discourse had an argumentative (the `rational` class) attitude.

[4] 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).

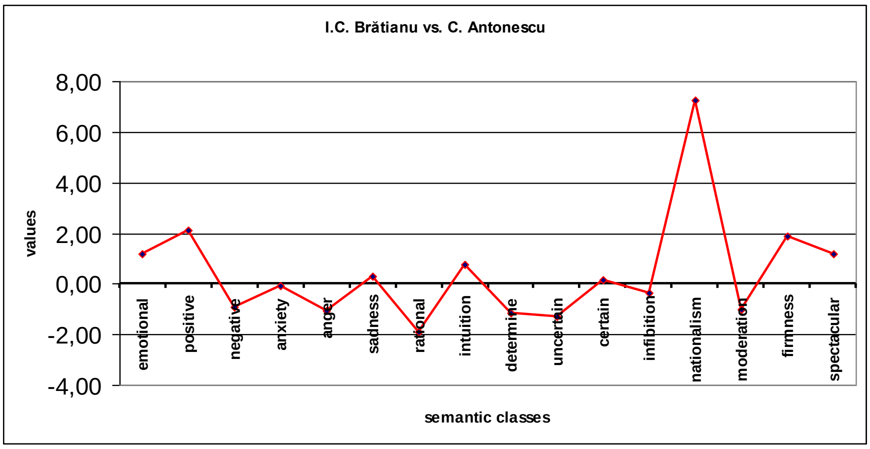


Figure 2: The average differences in the frequencies for each parent class (>0.5%) after processing political discourses, between Ion C. Brătianu and Crin Antonescu.

6. *Conclusions*

Surely, the problem of characterizing the public text receives no final solution with our approach. We believe, however, that our method sheds an interesting light and opens new perspectives. 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 historical 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 public speakers (politicians, columnists and so on).

In the future, new features will be added to the platform, with a special emphasis on the syntactic and rhetorical level analysis. The new release of DAT should help the user to identify and count relations between different parts of speech and to put in evidence patterns of use at the syntactic and rhetorical level.

The collection of manually annotated texts should also be augmented. 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 public discourse.

Acknowledgments. The DAT platform has been developed by Mădălina Spătaru at the Faculty of Computer Science, “Alexandru Ioan Cuza” University of Iași. In order to perform this research the first author received financial support from the POSDRU/89/1.5/S/63663 grant. The work of the second author is partially supported from the ICT-PSP project METANET4U.

References

Bally, Ch. (1950). Linguistique générale et linguistique française. Berna, 43-44.

Chomsky, N. (1968). Language and Mind, Harcourt Brace Jovanovich, Inc., chapter III.

Cicero, (1973). Opere alese. *Ed. Univers*, București, II, 51.

Coșeriu, E. (2004). Teoria limbajului și Lingvistică generală, *Ed. Enciclopedică*, București, 319-324.

Cristea, D., Răschip, M., Forăscu, C., Haja, G., Florescu, C., Aldea, B., Dănilă, E. (2007). The Digital Form of the Thesaurus Dictionary of the Romanian Language. In Proceedings of SpeD 2007 *Speech Technology and Human - Computer Dialogue*, Iasi, May 10-12, 2007.

Desideri, P. (1984a). En marge du discours politique. *Degrés*, 12, 37, 1-9.

Edelman, M. (1985). The Symbolic Uses of Politics. *Urbana: University of Illinois Press*. Originally published in 1964.

Fox, B.A. (1987). Discourse structure and anaphora. *Written and conversational English*. Cambridge: Cambridge University Press.

Gîfu, D., Cristea, D. (2011). Computational Techniques in Political Language Processing: AnaDiP-2011. *In J.J. Park, L.T. Yang, and C. Lee (Eds.), FutureTech 2011*, Part II, CCIS 185, 188–195.

Kilgarriff, A., Rosenzweig, J. (2000). English SENSEVAL: Report and Results. In *Proceedings of the 2nd International Conference on Language Resourcesand Evaluation*, LREC, Athens, Greece.

Lasswell, H. D. (1936). Politics: Who Gets What, When, How., McGraw-Hill, New York.

Lesk, M. (1986). 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.

Liddy, E.D. (2001). Natural Language Processing in *Encyclopedia of Library and Information Science*, 2nd Ed. NY. Marcel Decker, Inc.

Marinescu, G., Grecescu, C. (ed.) (1938). Ion C. Brătianu. Acte și cuvântări, vol. I – part I (june 1848 = decembrie 1859). *Cartea Românească*, Bucharest, 228-237.

Perelman, C., Olbrechts-Tyteca, L. (1972). Traité de l'argumentation. *Éd. de l'Institut de Sociologie de l'Université Libre de Bruxelles*, 72.

Romaine, S. (1994). Language in society. An Introduction to Sociolinguistics. *Oxford University Press Inc.*, New York.

Stevenson, N. (1995). Understanding Media Cultures: Social Theory and Mass Communication, Londra: Sage.

van Dijk, T.A. (1972). Textual Structures of News in the Press. *Working notes*, University of Amsterdam, Department of General Literary Studies, Section of Discourse Studies, 14.

Wodak, R. (2006). Critical Linguistics and Critical Discourse Analysis. *Handbook of Pragmatics*, Benjamins.