Computer Science Journal of Moldova, vol.**21**, no.1(61), 2013

Towards an Automated Semiotic Analysis of the Romanian Political Discourse*

Daniela Gîfu, Dan Cristea

Abstract

As it is known, on the political scene the success of a speech can be measured by the degree in which the speaker is able to change attitudes, opinions, feelings and political beliefs in his auditorium. We suggest a range of analysis tools, all belonging to semiotics, from lexical-semantic, to syntactical and rhetorical, that integrated in the exploratory panoply of discursive weapons of a political speaker could influence the impact of her/his speeches over a sensible auditory. Our approach is based on the assumption that semiotics, in its quality of methodology and meta-language, can capitalize a situational analysis over the political discourse. Such an analysis assumes establishing the communication situation, in our case, the Parliament's vote in favour of suspending the Romanian President, through which we can describe an action of communication.

We depict a platform, the Discourse Analysis Tool (DAT), which integrates a range of natural language processing tools with the intent to identify significant characteristics of the political discourse. The tool is able to produce comparative diagrams between the speeches of two or more subjects or analysing the same subject in different contexts. Only the lexical-semantic methods are operational in the platform today, but our investigation suggests new dimensions touching the syntactic, rhetorical and coherence perspective.

Keywords: political discourse, natural language processing, president's suspension, lexical-semantic, syntax, rhetorical analysis, coherence of discourse.

* This work was supported by the POSDRU/89/1.5/S/63663 grant, and the ICT-PSP projects METANET4U #270893 and ATLAS #250467.

1 Introduction

One of the major recent developments in the evaluation of the political language and its related facets (rhetoric, political science, journalism, sociology, etc.) is the increasing attention being paid to the objectivity and relevance of the semiotic dimensions.

Theoretical approaches in the semiotics of discourses, involving pragmatic aspects (the dynamics of relations between individuals and signs), semantic (conceptual conglomerate met in the meanings of terms), and syntactic (relations between signs) showed a significant strengthening after the '80s. The current approaches in analysing the political language (the applicative dimension) 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 proficiency in the interpretation of language for a range of tasks or applications [12].

In this paper we start by describing a platform, the Discourse Analysis Tool (DAT), which integrates a range of language processing tools with the intent to build complex characterisations of the political discourse and show how its functionality can be prolonged with more complex features. A linguistic profile 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.), semantics (semantic classes used) and syntax (complexity of syntactic constructions, the frequency of relative clauses, length of the sentences, number of clauses in sentences, subordinate/coordinate structures, frequent use of certain type of syntactic relations, etc.).

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. The collection described in this paper refers to the Romanian language and has been acquired with the help of an interactive graphical tool which

allowed easy annotation, visualisation and modification of syntactic trees, initially obtained as a result of an automatic parsing process.

Our purpose was to develop a computational platform able to offer to researchers in mass-media and political sciences, to political analysts, to the public at large (interested to consolidate their options before any political context analysed), and, why not, even to politicians themselves, the possibility to measure different parameters of a written political discourse.

The paper is structured as follows. Section 2 shortly describes the previous work. Section 3 discusses a number of lexical-semantic, syntactic, rhetorical and pragmatic features on which an automatic analysis is capable to manipulate values in view of drawing statistics. Section 4 presents a platform for multi-dimensional political discourse analysis. Section 5 discusses an example of comparative analysis of discourses collected during the presidential crisis of July 2012, when the Parliament voted in favour of suspending the Romanian President. Finally, section 6 highlights interpretations anchored in our analysis and presents conclusions.

2 Previous work

The aim of an interdisciplinary approach such as analysing the language of political speeches is to define and explain different discursive contexts, in this case, reflected by the online 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 [19].

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) [16]. There

are, however, important differences between the two platforms. LIWC-2007 is basically counting words and incrementing counters associated with their declared semantic classes. In the lexicon, words can be given by their long form, as a complete string of characters, or by their roots. For each text in the input, LIWC produces a set of tables, each displaying the occurrences of the word-like instances of the semantic classes defined in the lexicon, as sub-unitary values. For each semantic class, such a value is computed as the number of occurrences of the words corresponding to that class divided by the total number of words in the text. It remains in the hands of the user to interpret these figures. Also, LIWC has no support for considering lexical expressions.

A previous version of DAT [8] performs part-of-speech (POS) tagging and lemmatization of words. The lexicon contains a collection of lemmas (9.000) 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. In contrast with LIWC-2007, which includes 64 semantic classes (classified into 4 categories: linguistic – 22 classes, psychological – 32 classes, socio-professional preoccupations – 7 classes and paralinguistic – 3 classes), DAT.v3 works with 33 semantic classes, out of which 5 are newly introduced, chosen to fit optimally with the necessities of interpreting the political discourse.

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 more 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 semantic approach was the inclusion of a collection of formulas which can be used to make comparative studies between different subjects. The lexicon entries are coded in XML, following one of the patterns: `<word stem="`*wordStem*`" classes="`*semList*`">`, or `<word`

`lemma="`*wordLemma*`" classes="`*semList*`">`, in which *wordStem* is the stem of a word (therefore symbols optionally followed by the '*' sign), *wordLemma* is the lemma of a word, and *semList* is a list of semantic classes (each indicated by a unique identifier). The following line shows such an example of lexical entry:

```
<word lemma="deportare" classes="30,11"/>
<word stem="conspiraţioni*" classes="30,1,5,10"/>
```

3 Semiotic features of political discourse

As meta-language, the semiotics explain the evolution of different types of object-languages, from physical to linguistic (among those – the political discourse). It helps to understand the way the humans apply these systems with the intend to "designate states of possible worlds or to criticize and even change the structure of systems" [6].

The three analytical horizons are: structural analysis of the levels / hierarchical relations of (macro)sign, the triadic analysis (syntactic, semantic, pragmatic), and the analysis of the communication situation taken for investigation. In the following we will focus on one of the three horizons of analysis assumed by the semiotic methodology, namely, on the triadic analysis. Conforming to this view, any text/discourse can be analysed from three perspectives [15]: syntactic (the relation between signs), semantic (the relation between signs and reference), and pragmatic (the relation between participants in the communication). Highlighting methodological operations presumed by such a perspective offers as many (re)signifying strategies of political contexts.

We will adopt analytical techniques developed by the NLP field to a semiotic study over political texts, in the classical sense [17], that go back to the methodology proposed by Ferdinand de Saussure [20], in order to show that the results can be significantly comparable and, therefore, there are good reasons to trust the computational techniques.

3.1 The lexical-semantic perspective

A lexical-semantic perspective is supposed to focus on the following targets:

1. establishing meanings that a political speech includes, as a whole or at the level of its content units (negative/positive, affirmative/adversative, etc.); determining the correlation degree (motivation) between the orientation of the political speech and the language (code) used (adequate, partially adequate, inadequate);

2. a qualitative-semantic analysis of content units, that could be operated on two dimensions: denotative (what is said explicitly about the topic discussed), focusing on the intelligibility of the political text, by assessing its lexical-semantic connectedness [18], or by counting the originality, oddity or banality of the used lexicon, as well as the phrase length, the number of subordinate sentences, parentheses, etc.; connotative (what are the side suggestions, the sayings in-between the lines, the symbolistics of the language used), aiming to highlight the possible hidden semantic meanings of a speech and determining the most likely ones by taking into account all circumstantial factors (situational), and specifying the gap between the explicit and implicit intentions expressed;

3. a quantitative-semantic analysis focusing on determining of the frequency of key concepts encountered in the political text, highlighting the frequency of certain themes in the speech, identifying the frequency of emotionally charged terms, etc.; building a dictionary of symbols (for key-concepts) specific to the political discourse that helps to frame it in terms of semantic categories.

4. a discourse and para-language analysis considering the identification of the rhetorical aspects of the verbal language (spectacular, suggestive, allusive, emotional, metaphoric factors, etc.), and the characteristics of the nonverbal language which have a significant weight in the political discourse.

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 and emotional dimension to the discourse (the class `emotional`). The political metaphor may loose the virtues of poetical metaphor, becoming injurious (the class `swear`).

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 [13]. The sentence is composed out of an ordered sequence of language signs, which are governed by a set of combinatorial rules.

From this perspective, the syntactic analysis of a text aims at: segmenting the text onto information units (sentences, clauses, phrases, words and punctuation markers), identifying the constituency structure of the sentence (recurrent levels of constituency), emphasising the dependency structure of a sentence (putting in evidence the unique syntactic head of each word and the relation linking it to its head in a tree-like dependency structure [21], etc. The syntax may reveal the level of culture, intentional persuasive attitudes towards the public, iritation or rude passion during the production of speech, etc.

Then, a combination of syntactic and semantic means of investigation could bring forward the semantic verbal roles in sentences (see, FrameNet [2]), as well as the balance between given and new or rheme and theme [10].

The final goal of a combined syntactic-semantic analysis is the inference of a logical-form of the sentence, which would give a formal

expression of the content.

3.3 The discourse-level perspective

Beyond the sentence, at the discourse level, a rhetorical analysis identifies relations or interdependencies holding between adjacent spans of text. Then, the arguments of a relation (discourse units, or spans of text) could be compared one to the other in terms of their importance (nuclearity). The rhetorical relations and their nuclearity are grouped in rhetorical schemes, as general patterns in which spans of text can be recurrently analyzed.

The main regard of discourse theories are towards explaining the structure of a text (how is a text organised in segments and these ones – in sub-segments, and how this compositional structure influences the comprehension of the meaning), its degree of difficulty (for instance, why are certain texts easier to interpret than others [9]), its cohesion (or what makes that different components of a text look like being glued together [11]) and coherence ("Intuitively, coherence is a semantic property of discourse, based on the interpretation of each individual sentence relative to the interpretation of other sentences." [22]), and, finally, what is the relationship between coherence, cohesion and discourse structure [4]. Summarisation issues are nonetheless immerged onto a discourse-level analysis.

3.4 The pragmatic perspective

The pragmatic analysis should be based on the knowledge of the political intentions (of both the speaker, and the receiver) in connection with the ideological meanings of a speech. Only in good knowledge of the political aspirations of the hearers and knowing that the speaker knows himself this spectrum of political aspirations, a human analyst would succeed in interpreting the whole range of subtleties of a political speech. It is clear that pragmatics makes a good deal of the political speeches interpretation process. It is nevertheless true that an experienced human analyst would succeed to acquire these facets of the pragmatic context of a political speech even having little direct

knowledge on them. It is like in an act of reverse engineering in which the analyst is able to infer the political ideology of the speaker and of the auditorium from the speech itself.

A closer look on a pragmatic analysis of a political discourse reveals the following aspects: interpretation of the text in terms of psychological distance between the partners, opponents, etc.; defining the transmitter's political attitude before and after the communication; determining the receptor's political attitude (i.e. being pro, against or undecided); pursuing echoes of the political communication in the audience (immediately), or in the society (after a delay), etc.; discovering the political speaker's intentions by evidencing the semantic roles of different sentence constituents (reiterations, expressions, etc.).

4 A platform for multi-dimensional political discourse analysis

In this section we briefly describe the Discourse Analysis Tool (DAT), a platform which integrates a range of language processing tools, with the intent to build complex characterisations of the public discourse. Out of the discussed perspectives of semiotic analysis, DAT (currently at version 3) implements only lexical-semantic features. The concept behind the lexical-semantic analysis in DAT is that the vocabulary used by a speaker opens a window towards the author's sensibility, towards his/her level of culture, her/his cognitive world. Some of these means of expression are persuasive, aimed to convince the public on the own opinions, while others are manipulative, aimed to induce a false perspective on an issue. Figure 1 displays a snapshot of the interface showing a semantic analysis, during a working session. The platform incorporates two alternative views for presenting the results of the lexical-semantic analysis: graphical (pie, function, columns and areas) and tabular (Microsoft Excel compatible).

The vocabulary of the 33 semantic classes (detailed in Figure 2) of DAT.v3 are considered to fulfil optimally the necessity of interpreting the political discourse of our corpus.

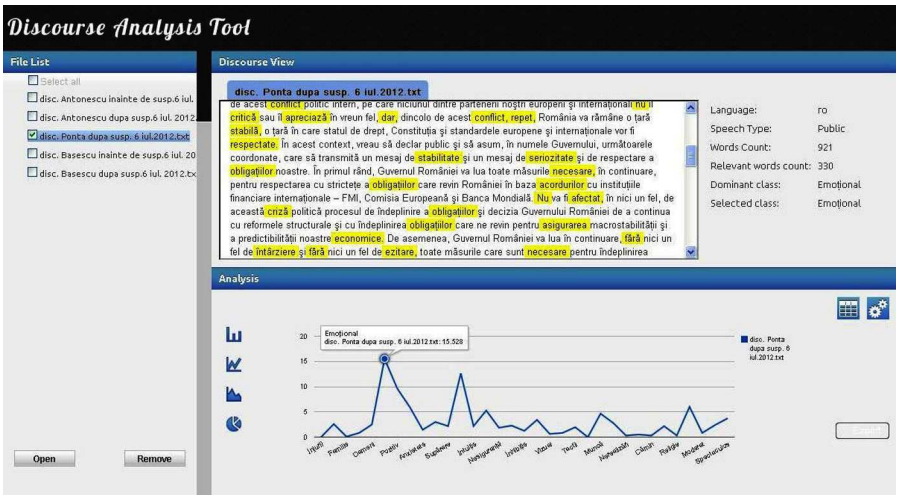

Figure 1. The DAT interface: in the left window the selected files appear, in the middle window – the text in the selected file, and in the right window – information about the text (language, word count, dominant class, etc.). Below, a plot (form chosen from a range of graphical tools) is displayed. By selecting a specific class in the middle window, all words assigned to that class are highlighted in the text.

```
<?xml version="1.0" encoding="UTF-8"?>
-<classes>
<class name="swear" id="1"/>
<class name="social" id="2"/>
<class name="family" id="3" parent="2"/>
<class name="friends" id="4" parent="2"/>
<class name="people" id="5" parent="2"/>
<class name="emotional" id="6"/>
<class name="positive" id="7" parent="6"/>
<class name="negative" id="8" parent="6"/>
<class name="anxiety" id="9" parent="8"/>
<class name="anger" id="10" parent="8"/>
<class name="sadness" id="11" parent="8"/>
<class name="cognitive" id="12"/>
<class name="intuition" id="13" parent="12"/>
<class name="determine" id="14" parent="12"/>
<class name="uncertain" id="15" parent="12"/>
<class name="certain" id="16" parent="12"/>
<class name="inhibition" id="17" parent="12"/>
<class name="perceptual" id="18"/>
<class name="see" id="19" parent="18"/>
<class name="hear" id="20" parent="18"/>
<class name="feel" id="21" parent="18"/>
<class name="sexual" id="22"/>
<class name="work" id="23"/>
<class name="achievements" id="24"/>
<class name="failure" id="25"/>
<class name="agreement" id="26"/>
<class name="home" id="27"/>
<class name="financial" id="28"/>
<class name="religion" id="29"/>
<class name="nationalism" id="30"/>
<class name="moderate" id="31"/>
<class name="firmness" id="32"/>
<class name="spectacular" id="33"/>
</classes>
```

Figure 2. Semantic classes in DAT.v3

Our interest went mainly in determining those political attitudes able to influence the voting decision of the auditorium. But the system can be parameterised to fit also other conjunctures. As such, the user can define at will her/his semantic classes, which, as can be noticed in Figure 2, are partially placed in a hierarchy.

The development of the lexicon associated with these classes was done in several phases. We started with a small vocabulary (mainly looking for translation equivalents in Romanian of a subset of the LIWC-2007 classes). Then, the words of this initial lexicon have been used as seeds in a trial to enrich the lexicon automatically by using the morphological database of DEX-online, an online electronic dictionary for Romanian language.

To prepare the integration of syntax in DAT, a dependency parser for Romanian is in the process of being trained on a dependency treebank. This corpus of syntactic trees (incorporating now over 4,000 tree structures) has been partially developed manually, by using a graphical editing tool (TreeAnnotator) and, later on, by the dependency parser itself, in a bootstrapping manner. After the corpus reached the size of 100 structures, the development of the resource continued in a bootstrapping manner: the new sentences belonging to the interim president were first parsed by the parser and then manually corrected by the first author of this paper. This way, the development of the corpus gained very much in speed. The format of the stored trees is XML, with the following elements:

- `sentence` – marking the sentences; its attributes are: a unique identifier and the name of the annotator who lastly worked over the sentence;
- `word` – marking individual words of the sentence; its attributes are: a unique identifier, the morphological tag, the lemma form of the inflected word, the ID of its parent word (the head in the dependency structure) and the name of the relation linking the word to its parent.

The following version of DAT is planned to integrate also a syntactic parser, offering to the user the possibility to identify and count relations

between different parts of speech, to put in evidence patterns of use at the semantic and syntactic level, discursive behaviours, etc.

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 three political leaders, Traian Băsescu, Crin Antonescu, and Victor Ponta. Traian Băsescu was the Romanian's president since 2004 (with an interruption in the summer of 2012, when he was suspended, period monitorized in this study), one of the most complex personalities of the Romanian political arena of the last decade. The second political actor, Crin Antonescu, is an ex-leader of the Liberal Party, for a short while – President of the Senate and then – the Romania's interim President (during Băsescu's suspension). The last political actor, Victor Ponta, is an ex-leader of the Social Democrat Party, the actual Romanian prime minister, and represents the new political generation. His party and Antonescu's party form the USL coalition (The Social-Liberal Union). This coalition, with a social-liberal ideology is a premiere in Romania.

We are, this way, putting on the balance three styles of political discourse that, at least in principle, are perceived as being different as ideologies (democrat-liberal, liberal, and social-democrat). But more than comparing political discourses belonging to different ideologists, the year 2012, so politically dense, offers the opportunity to study how the stress of the political battle from the edge of a crises is reflected in these major opponents' speeches, as evidenced by a semiotic analysis. Indeed, 2012 was the year of governmental changes in Romania. After the January street protests and following President Băsescu's request, the Boc Government resigns (20 January) and is replaced by the Ungureanu Government (6 February). Permanently contested and sanctioned by the public opinion, less than 3 months later, the Ungureanu Government falls, following a vote of confidence from the Parliament, put forward by the opposition block PSD-PNL-PC (27 April).

The President will designate a new premier, Victor Ponta, the head of the principal opposition party, PSD, sustained by Crin Antonescu, the liberals' head. The two politicians make the bases of a new coalition, USL, whose principal objective is the removal of the President, following thus one of the demands of the protestants. On 10 June, the local elections will completely change the political map of Romania: the governmental coalition becomes legitimate in the principal cities and districts of the country. The next step will be the relegation of Băsescu, preceded by a motion of censure (6 July), when the President is suspended. This will trigger the political crisis, around which our analisys gravitates.

For the elaboration of preliminary conclusions on the crisis process, we collected, stored and processed, partially manually, partially automatically, political texts published by three national on-line publications having similar profiles. A small part of this corpus which includes a collection of 100 political sentences, thoroughly chosen, each containing one or more clauses, has been syntactically annotated.

5.2 The lexical-semantic analyses

Apart from simply counting frequencies of mentions of semantic classes of one author, the system can also perform comparative studies between two or more authors or for the same author in different periods of time.

To exemplify, we present below different charts with two streams of data, representing the political speeches in the context of the political crisis (before Băsescu's suspension), belonging to the three important political leaders mentioned above. In fact, our analysis makes a two by two comparison of the three political actors mentioned. In each of the diagrams that follow, for each semantic class, the values corresponding to one subject are subtracted from the other. Our experience shows that an absolute difference value below the threshold of 0.5% should be considered as irrelevant and is, therefore, ignored in the interpretation. For this reason, these classes are not represented in the chart.

The graphical representation in Figure 3, in which Traian Băsescu, President of Romania before the temporary suspenssion (figured above

the Ox axis) is compared against Crin Antonescu, the President of the Senate at that time (figured below the Ox axis), should be interpreted as follows: Traian Băsescu was interested more on the labour market in Romania (the class `work`), uttered in an intuitive tone (the class `intuition`), than Crin Antonescu, whose discourse had patriotic accents (the class `nationalism`), and who developed a comparative analysis between failures (the class `failures`) and achievements (the class `achievements`) during Băsescu's presidency.

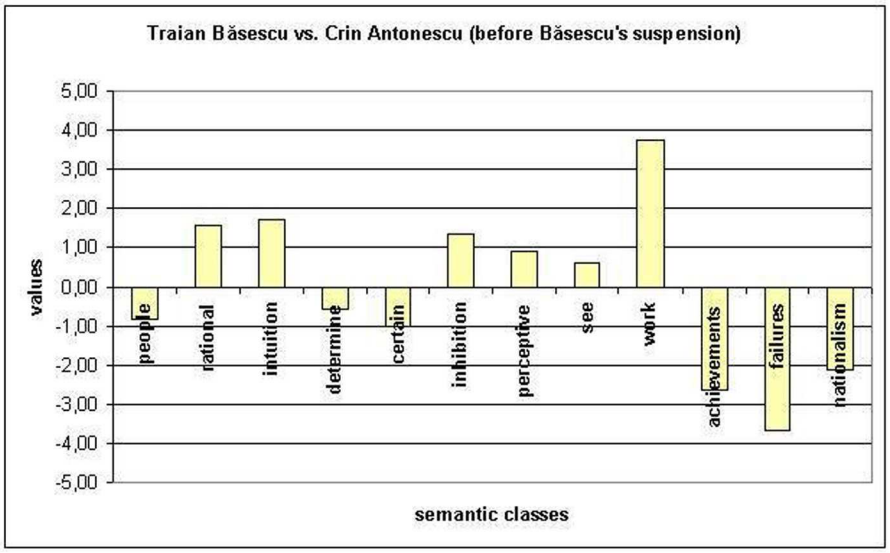

Figure 3. The average differences in the frequencies of all classes (that cumulate more than 0.5 % occurrences) in the political discourses of Traian Băsescu and Crin Antonescu, before the initiation of the crisis.

It is interesting to see how quickly the discursive spectrum changes after Băsescu's suspension: in the same day, Băsescu becomes negative, and Antonescu positive. In fact, a normal attitude... as the first subject was suspended after the vote of the Parliament, and the second subject will become the interim President, triggered by his quality of President of the Senate.

This new situation is narrated by the chart in Figure 4, which shows again two streams of data belonging to the same subjects, but this time after the moment the crisis erupted (after Băsescu's suspension). Our reading of the diagram is as follows: Traian Băsescu had a negative tone (the class `anger`), but he kept a more rational attitude (the class `intuition`) than Crin Antonescu, who becomes full of hope (the class `positive`) and who has a stronger patriotic attitude (the class `nationalism`).

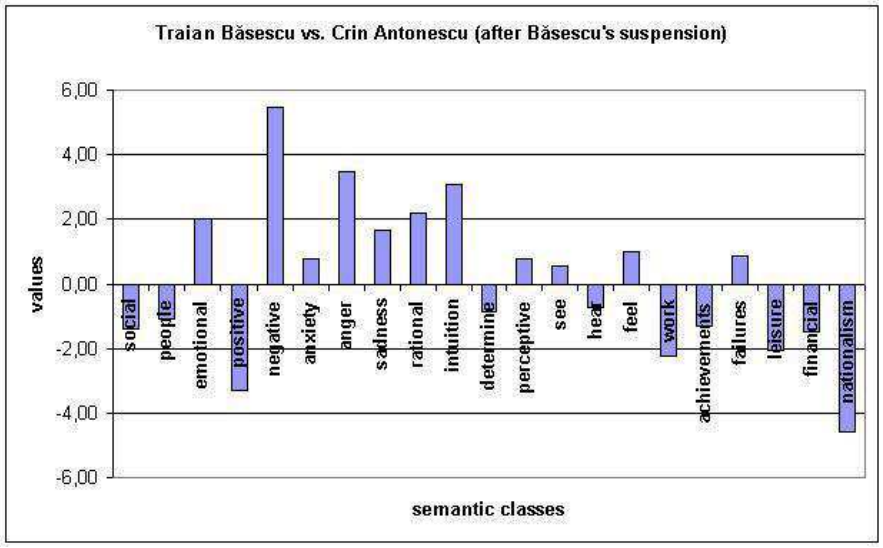

Figure 4. The average differences in the frequencies of all classes (that cumulate more than 0.5% occurrences) in the political discourses of Traian Băsescu and Crin Antonescu, after the initiation of the crisis.

The inedited element was the absence of Romanian Prime Minister, Victor Ponta, at the meeting of Parliament. He preferred to have a short statement after Băsescu's suspension.

It is also interesting to make a comparative radiography of the other two political opponents – Traian Băsescu and Victor Ponta in a critical moment, i.e. immediately after the political crisis has been fired. The

chart in Figure 5 compares the political texts of Traian Băsescu (above the Ox axis) and Victor Ponta (below the Ox axis). Our reading is the following: Traian Băsescu had a negative tone (the classes `negative`, and `anger`), but he kept a rational attitude (the classes `rational`, and `intuition`), while Victor Ponta was satisfied with the results (the class `positive`).

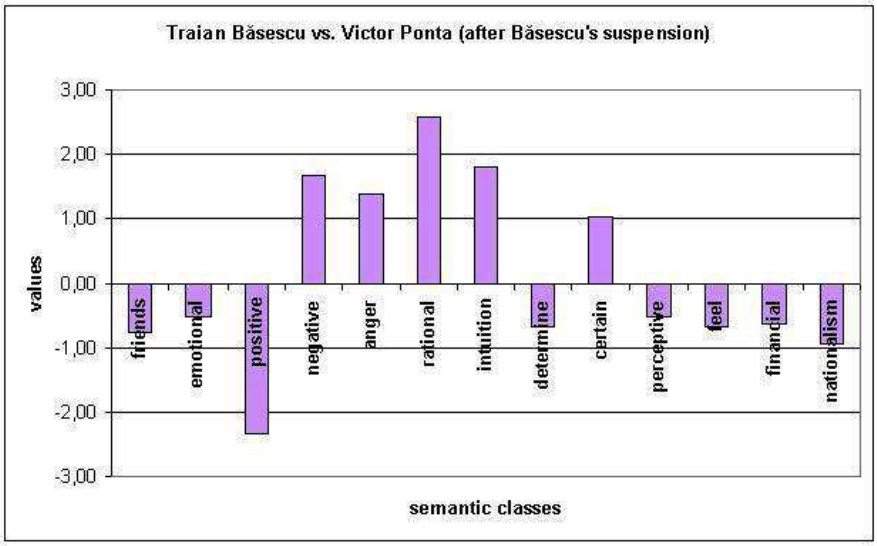

Figure 5. The average differences in the frequencies of all classes (that cumulate more than 0.5% occurrences) in the political discourses of Traian Băsescu and Victor Ponta, after the initiation of the crisis.

One of the interesting studies which we have in attention is to perform comparative studies for the same political actor in different periods of time, in our case, before and after the initiation of the crisis that resulted in the Romanian President's suspension. For exemplification, we have chosen Băsescu's speeches.

The graphical representation in Figure 6, in which the President Traian Băsescu's speech (above the Ox axis) is compared against the suspended President Traian Băsescu's speech (below the Ox axis)

should be interpreted as follows: before his suspension, the subject accentuated more on social aspects, his discourse was positive and insisted on the achievements. On the contrary, after being suspended his discourse became emotional, negative, with eruptions of anger and sadness, while still preserving a rational tone. For instance, before his suspension, Băsescu used expressions such as: "se pare că eu nu reuşesc" (*it seems that I don't succeed*), "decât atingerea scopurilor politice" (*other than attaining political purposes*), "Eu cred că este o greşeală" (*I consider being a mistake*), etc. After president's suspension, Băsescu changed the discursive tone preferring expressions, such as: "în concluzie, mergem la Referendum" (*in conclusion, we're going to Referendum*), "dar, să vedem" (*but let's see*), "Dar înainte de a merge la referendum" (*but before going to Referendum*), etc.

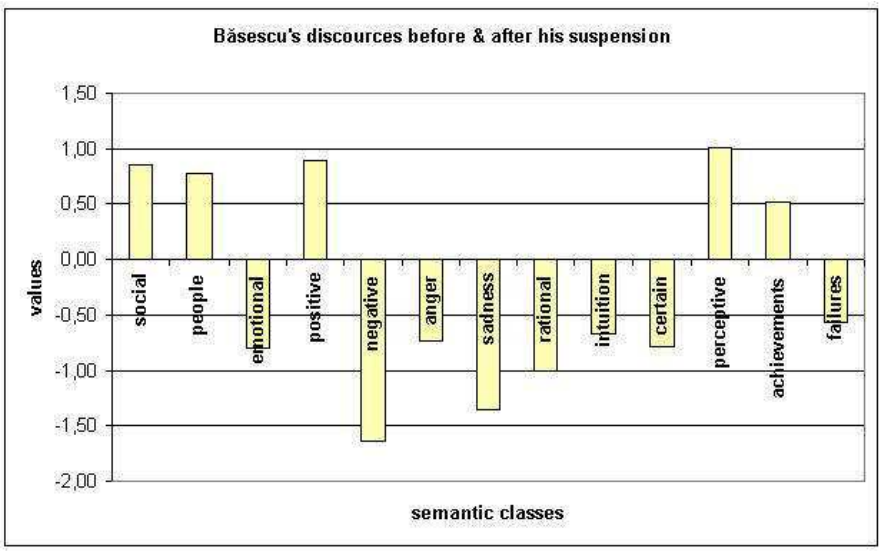

Figure 6. Băsescu's versus himself, before and after the suspension.

5.3 The syntactic analyses

In order to proceed with a syntactic level investigation, the text bodies have been pre-processed automatically by an NLP processing flow that included: sentence splitting, tokenisation, part-of-speech tagging and lemmatisation. Then, two thirds of the corpus were automatically parsed at the FDG structure, and the remaining part was manually annotated using the TreeAnnotator interface. Both resulted in heavily annotated XML files. Table 1 shows the size of the corpus used in these syntactic analysis.

Table 1. The corpus of texts annotated for syntax in Crin Antonescu's speeches

Number of sentences	Number of words	Number of annotated sentences	Number of words in the annotated sentences
123	3,960	100	3,286

We concentrated our analysis on three types of syntactic relations that we believe have a rhetoric role in the crisis context: adjectival, appositional and anacoluthic [7] (Table 2 displays absolute and relative values for all types of relations). Note that none of these relations are compulsory in the syntax of the phrase (the same as with the overtly expressed pronouns on the position of subject, in Romanian, for instance). Even more than that, the anacoluthic constructions are considered errors in a cultivated speech, although, properly mastered, they could have rhetorical value. Therefore, the use of all these relations is strictly a matter of personal choice.

The adjectival structure (marked as `a.adj`, `a.subst`, `a.vb` and `a.adv` in Table 2; 19.5% of all relations in the corpus) means adjectival, nominal, verbal and adverbial attributes: the adjectives add colour to the discourse. The orator not only that brings a contextual, albeit new, information, but enhances the enouncement by detailing it and developing it. The adjectival group is usually part of the rheme (the

Table 2. Occurrence of dependency relations for Crin Antonescu's political speeches corresponding to the crisis context

Relation	Number	Percentage
coord.	286	11.1%
prep.	320	12.4%
a.adj.	156	**6.0%**
c.d.	198	7.7%
punct.	100	3.9%
sbj.	96	3.7%
part.	120	4.6%
c.i.	76	2.9%
a.subst.	198	**7.7%**
a.vb.	112	**4.3%**
det.	90	3.5%
c.c.m.	98	3.8%
n.pred.	60	2.3%
aux.	84	3.3%
a.adv.	40	**1.5%**
refl.	120	4.6%
anacol.	98	**3.8%**
c.c.t.	40	1.5%
neg.	80	3.1%
ap.	102	3.9%
c.c.l.	46	1.8%
comp.	40	1.5%
c.c.scop.	24	0.9%
Total	2584	100

new information), not the theme (the old), being placed (in Romanian) usually after the theme element. When it 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 from the part of the author.

In Figure 7 the arrows highlight the presence of two adjectival structures: "România adevărată" (*Real Romania*), "România normală" (*normal Romania*).

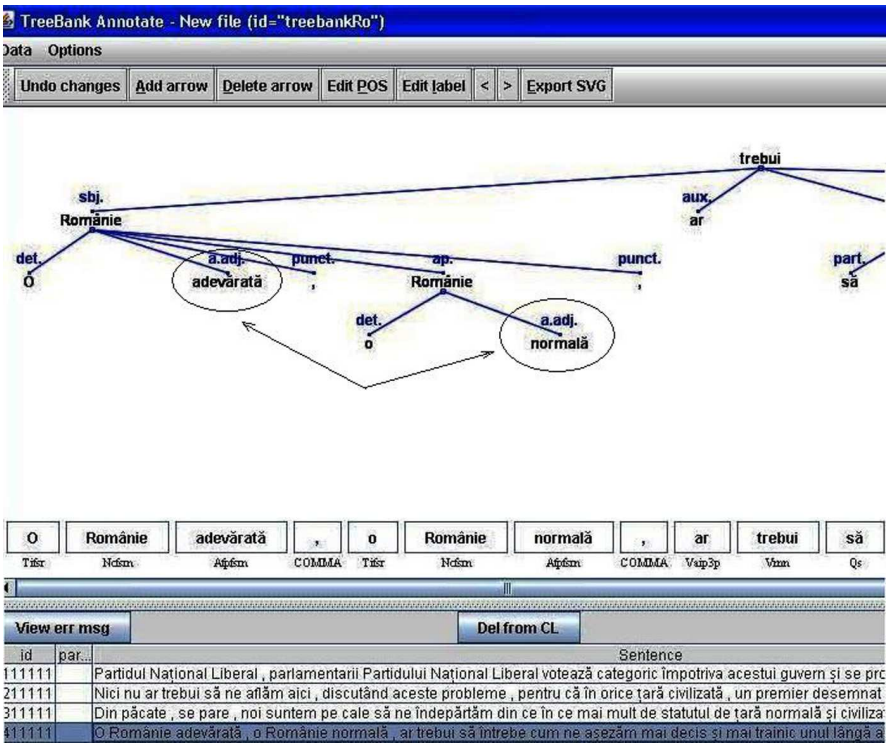

Figure 7. An adjectival structure on a dependency tree visualised with TreeAnnotator

The apposition structure (`ap.` in Table 2; 3.9%): this is the depen-

dency relation that holds between two lexical sequences, called base and apposition (the apposition being open to an unlimited number of terms), the second one giving a complementary information on the first one.

The apposition structure should be delimited from the syntactic relations of subordination and coordination, because between the base and the apposition there is no syntactic hierarchy. However, by convention, in our dependency structures, the appositive term is represented attached to the base.

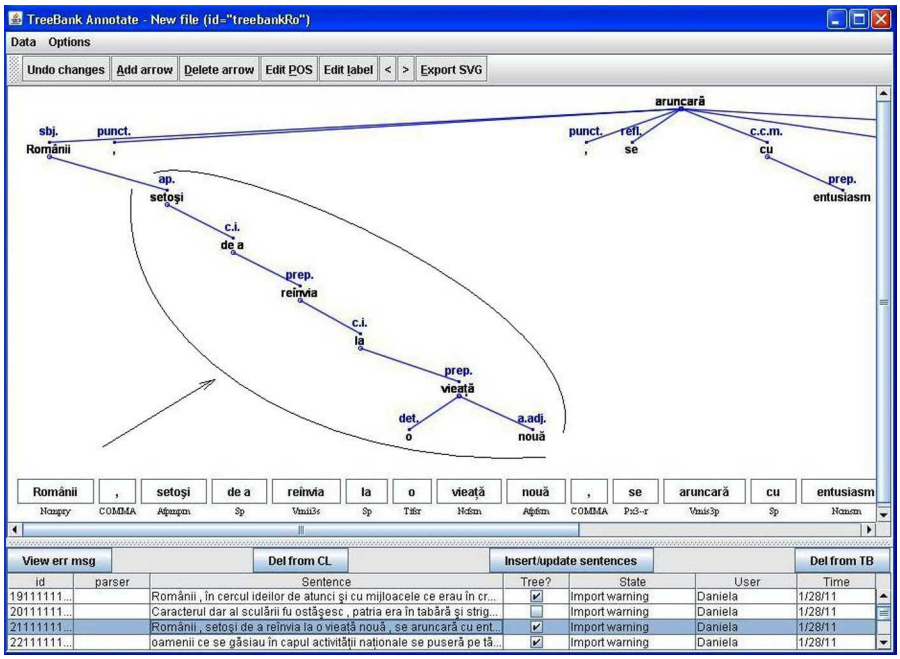

Figure 8. An apposition structure visualised with TreeAnnotator

In Figure 8, the arrow highlights an apposition structure. The sentence "Românii se aruncară cu entuziasm..." (*The Romanians jumped with enthusiasm...*) is interrupted by the apposition "setoşi de a reînvia la o viaţă nouă" (approx. *thirsty to be reborn in a new life*), which add contextual information to the main subject "Românii" (*Romanians*).

The anacoluthic structure (anacol. in Table 2; 3.8%) marks an interruption of a syntactic construction (clause, phrase) and continuation with another construction. In general, the anacoluthon is considered an error in the grammar books. So, strictly grammatical it is forbidden. To evidence it automatically in texts is extremely difficult because it is rare and a parser needs many occurrences in order to develop the ability to put it in evidence. In long sentences it is difficult even for an experienced annotator to note these intentional (or unintentional) errors, because the interspersed components have such diverse structures.

In the example in Figure 9, the principal sentence "După dânsul, veni mai târziu Regulamentul" (*After him, the Regulation came later*) is followed by the anacoluthon "căci el" (*because it*), which represents a suspended nominative (nominativus pendens) relation. The author feels the need for a change in the discourse theme, after upgrading the nominative "el" (*it*), seeming to have the function of subject near a predicate which is never uttered afterwards. The experienced political actors use anacoluthic structures strategically in communication with the intend to focus the discourse or to highlight a particular element. In this example, the politician focuses on "Regulamentul" (*the Regulation*), and the subordinate concessive sentence "deşi fu impus de străini" (*although having been imposed by foreigners*).

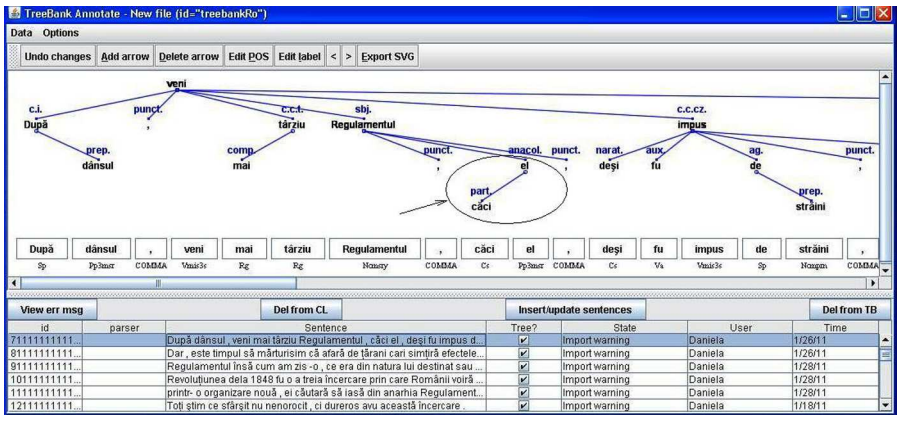

Figure 9. An anacoluthic structure, visualised with TreeAnnotator

5.4 The rhetorical and pragmatic analysis

As mentioned, a discourse-level type of analyses should reveal elements of coherence and cohesion of the text, together with the identification of the rhetorical structure of the discourse. Some of these aspects are technologically feasible with different degree of accuracy. Discourse level techniques are applied at the very end of a long processing chain, which should include: segmentation into sentences, tokenisation, pos-tagging and lemmatisation, and segmentation at the clause level. Optionally, in a more developed type of analysis, it should also include: shallow parsing (for the identification of noun-phrases), name-entity recognition, and anaphora resolution.

Counting different types of rhetorical relations in a political speech could reveal a lot over the rhetorical strategy of the author and the dynamics of the discourse. A rhetorical parser is usually trained to recognise complex rhetorical trees out of a corpus manually annotated with these structures [14]. The discourse parser developed in the NLP-Group@UAIC-FII builds rhetorical structures based on the identification of cue words and other discursive features [3, 1]. The outputted trees of the current implementation, however, miss the names of the relations, but they can retain the cue-words and the nuclearity.

Perfectly feasible with the present day technology are also the identification of some cohesion and coherence elements of a political speech, as mentioned in Section 3.3. Centering parsers (see [5, 1], for instance) could measure the coherence of a text on a scale from 0 to 4 [4]. Scaling up an exploratory tool for the purpose of our investigation would be an interesting research objective, which should take into consideration that a high quality human discourse is not always one that reaches a maximum on the coherence scale, because that one would also be very boring [5], the same as it should not be a randomly generated one, because this would be completely incoherent. It's a pharmacy chemistry that the great orators know to master, combining in proper quantities, as the discourse unfolds, the fulfilment of expectations with the unexpected and surprise.

Present day techniques make feasible the development of a number

of automatic techniques in the area of rhetorical and coherence analysis. It will be our further objective to concentrate on this type of investigation.

6 Conclusions

The analysis we proposed in this paper aims at verifying if a semiotic perspective anchored in natural language processing techniques could be of value in valuating political speeches. If this performance proves to be feasible, than semiotics would become a very applicative science, with interesting virtues in the optimization of the political discourse. Rhetorical weapons in the hands of a political actor should be: the diversity of the lexicon and a proper mastering of the semantic classes, the syntactic form, the emancipation of the expression, the coherence and the proper mastering of the comprehensibility. It is our conviction that the present day linguistic technology can successfully cover many of these facets.

However, we are aware that this study only sketches a way to go, and a lot more should be studied until a reliable discourse interpreting technology will become a tool in our hands. We should also be aware of the dangers of false interpretation. For instance, if we take as example the three orators we used in our experiments, differences at the level of lexicon and syntax, which we have evidenced as differentiating them, should be attributed only partially to their idiosyncratic rhetorical styles, because these differences could also have ideological roots. 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. It remains yet to be decided the impact that the use of certain syntactic structures, such as adjectival, appositional and anacoluthic, could have over an auditory in the political discourse.

Different politicians could raise the use of these measures to the level of a rhetorical strategy, therefore exploiting them perhaps too much in the benefit of the aimed goals. In other words, this study shows that technological instruments are able to detect tendencies of manipulation

of the receiver with the evident role of detouring the attention of the audience from the actual communicated content in favour of the orator. The software allows online editing of a yet-to-be-delivered speech, in order to make it fit to the audience profile (public of large, journalists, different levels of culture).

Many interpretation facets are pertinent to the specific context a discourse is being uttered. For instance, in a crisis context a political discourse should be evaluated in function of the balance between the agenda of an orator that happens to be on the site of the political power, versus the opposite agenda. Different intensities of emotional levels could also be evidenced, and we prepare a more fined grade scale of emotional expressions. It is a known fact that the audience can be manipulated easily (e.g., the class `sadness`) by political actors when their themes are treated with excessive emotional tonalities.

We are aware that many technological aspects remain yet to be refined and enhanced. One of the most important is the determination of the senses of words and expressions in context. In the future we intend to include a word sense disambiguation module in order to determine the correct senses, in context, of those words which are ambiguous between different semantic classes, or between classes in the lexicon and outside the lexicon (in which case they would not have to be counted). Also, negations could completely reverse the semantic class a certain expression belongs to in certain contexts and need therefore special treatment.

The collection of manually annotated texts is only at beginning, a starting point for an efficient automatic annotation. In the near future we will 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.

We believe that the platform has a range of features that make it attractive as a tool to assist any kind of political campaigns. It can be rapidly adapted to new domains and to new languages, and its inter-

face is user-friendly and offers a good range of functionalities. It helps to outline distinctive features which bring a new and, sometimes, unexpected vision upon the discursive characteristics of political authors.

Acknowledgments: In performing this research, the first author was supported by the POSDRU/89/1.5/S/63663 grant, and the second author – by the ICT-PSP projects METANET4U # 270893 and ATLAS # 250467. Alex Moruz helped the first author to clean the DAT Romanian lexicon in an early phase. Afterwards it has been largelly extended by Radu Simionescu after importing the Romanian morphology from the DEX-online database. We are grateful to Cătălin Frâncu and Radu Borza for offering this database. The DAT platform has been developed by Mădălina Spătaru, as a post-master activity in the Faculty of Computer Science of the "Alexandru Ioan Cuza" University of Iaşi. All the Romanian NLP components mentioned in this paper were developed in the NLP-Group@UAIC-FII.

References

[1] D. Anechitei, D. Cristea, I. Dimosthenis, E. Ignat, D. Karagiozov, S.Koeva, M. Kopeć, C. Vertan. (2013, to appear). *Summarizing Short Texts Through a Discourse-Centered Approach in a Multilingual Context*. In Neustein, A., Markowitz, J.A. (eds.), Where Humans Meet Machines: Innovative Solutions to Knotty Natural Language Problems. Springer Verlag, Heidelberg/New York.

[2] C.F. Baker, C.F. Fillmore, J.B. Lowe. (1998). *The Berkeley Framenet project*. In Proceedings of the COLING-ACL 1998, Montreal, Canada.

[3] A. Belogay, D. Karagyozov, S. Koeva, C. Vertan, A. Przepiórkowski, D. Cristea, P. Raxis. (2012). *Harnessing NLP Techniques*, in Walter Daelemans, Mirella Lapata Lluis Marquez (Eds.) Processes of Multilingual Content Management, Proceedings of EACL 2012 – the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, pp. 6–10, ISBN: 978-1-937284-19-0.

[4] D. Cristea, N. Ide, L. Romary. (1998). *Veins Theory. An Approach to Global Cohesion and Coherence*. In Proceedings of Coling/ACL '98, Montreal.

[5] D. Cristea, A. Iftene. (2011). *Grounding Coherence Properties of Discourse*. In ALEAR Final Report, vol. II. Embodied Cognitive Semantics, Berlin, April.

[6] U. Eco. (1996). *Limitele interpretării* (Limits of interpretation), Ed. Pontica, Constanţa.

[7] *Gramatica limbii române* (The Grammar of the Romanian Language). (2005). Vol. II, Enunţul (The statement), Ed. Academiei Române, Bucureşti, 105–113, 619–31, 743–747.

[8] D. Gîfu, D. Cristea. (2012). *Multi-dimensional analysis of political language*, in "Future Information Technology, Application, and Service", in James J. (Jong Hyuk) Park, Victor C.M. Leung, Cho-Li Wang, Taeshik Shon (eds.), volume 1/164, Springer Science+Business, Media Dortdrecht.

[9] B.J. Grosz, A.K. Joshi, S. Weinstein. (1995). *Centering: A framework for modeling the local coherence of discourse*. In Computational Linguistics, 12(2), 203–225.

[10] Eva Hajicŏvá, B.H. Partee, P. Sgall. (1998). *Topic–Focus Articulation, Tripartite Structures, and Semantic Content*. In Studies in Linguistics and Philosophy, 71, Dordrecht, Kluwer.

[11] M.A.K. Halliday, R. Hasan. (1976). *Cohesion in English*. Longman, London.

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

[13] W.C. Mann, S.A. Thompson. (1988). *Rhetorical Structure Theory: Toward a functional theory of text organization*, in Text 8(3), 243–281.

[14] D. Marcu. (2000). *The theory and practice of discourse parsing and summarization*, The MIT Press, Cambridge, Massachusetts.

[15] Ch. Morris. (1938). *Foundations of the Theory of Signs*, The University of Chicago Press. Pennebaker, J. W., Francis, Martha E., Booth, R. J. (2001). Linguistic Inquiry and Word Count "LIWC2001, Mahwah, NJ, Erlbaum Publishers.

[16] J.W. Pennebaker, M.E. Francis, R.J. Booth. (2001). Linguistic Inquiry and Word Count LIWC2001, Erlbaum Publishers, Mahwah, NJ, 2001.

[17] H.F. Plett. (1983). *Ştiinţa textului şi analiza de text* (The science of text and the text analysis), Ed. Univers, Bucharest.

[18] N. Rescher. (1973). *The coherence theory of truth*, Oxford UP, London.

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

[20] Ferdinand de Saussure. (1916). *Cours de linguistique générale*, Payot, Paris.

[21] L. Tesniére. (1959). *Elements of structural syntax*, Editions Klincksieck.

[22] T. Van Dijk. (1977). *Text and Context. Explorations in the semantics and pragmatics of discourse*, Longman, New York.

Daniela Gîfu, Dan Cristea, Received July 24, 2012

Daniela Gîfu
"Alexandru Ioan Cuza" University of Iaşi
Faculty of Computer Science
16, Berthelot St., 700483 Iaşi, Romania
Phone: +40.232.201724
E–mail: *daniela.gifu@info.uaic.ro*

Dan Cristea
"Alexandru Ioan Cuza" University of Iaşi
Faculty of Computer Science
16, Berthelot St., 700483 Iaşi, Romania
Phone: +40.232.201542
E–mail: *dcristea@info.uaic.ro*

Institute of Computer Science
Romanian Academy, the Iaşi branch
2, T. Codrescu St., 700481-Iaşi, Romania