INT J COMPUT COMMUN, ISSN 1841-9836
7(5):832-839, December, 2012.

Public Discourse Semantics. A Method of Anticipating Economic Crisis

D. Gîfu, D. Cristea

Daniela Gîfu
Alexandru Ioan Cuza University of Iaşi, Faculty of Computer Science
16, General Berthelot St., 700483 Iaşi, Romania
E-mail: daniela.gifu@info.uaic.ro

Dan Cristea
1. Romanian Academy - the Iasi branch, Institute for Theoretical Computer Science
2, T. Codrescu St., 700481 Iaşi, Romania
and
2. Alexandru Ioan Cuza University of Iaşi, Faculty of Computer Science
16, General Berthelot St., 700483 Iaşi, Romania
E-mail: dcristea@info.uaic.ro

Abstract:
This paper provides a proof that anticipation of an economic crisis by analysing public discourses (in particular, speeches on economic issues) is feasible. It proposes a method of text classification and semantic interpretation based on natural language processing techniques that could be used to trace, over a period of time, the print press discourses, with the aim to valuate the perspective of occurrence of crises. Classification is the task of assigning tags (words, expressions) to the texts that make up a corpus. In our case, we were interested to identify among the texts under scrutiny those belonging to classes like *financial*, *economic*, *nationalism*, etc. This approach is sustained by the fact that public discourses can be characterized from a rhetorical perspective, depending on the specific strategies their authors have chosen: orientation to change opinions or to determine action, ratio between rational (*logos*) and emotional (*pathos*), etc. We are sugesting an automatic analysis of the content of the public language, by using quantitative measures. Our purpose was to develop a computational tool able to offer to researchers in the economic, social or political sciences, but, not less, to the public at large, the possibility to measure the acuity of different accents of a written public discourse (*financial*, *emotional*, etc.), as mean to anticipate the threat of financial waves. Such a tool could help the processes of decision making in the analysis of crisis. Although our analysis used as data the journalistic and economic environments of Romania, it could easily be extrapolated to other languages/countries.
Keywords: public language, text categorization, semantic analysis, economic crisis.

1 Introduction

In the atempt to divulge ante-factum crises in public discourse, primarily the voices of those entities must be listen to which are most influencial on the financial and economic domains. These entities, clearly, are: The Romanian National Bank (in the internal context) and the World Bank (in the international context)[1]. The voices of these entities are best listen to in the public speeches of governors and, in many cases, of journalists specialised on economic-financial issues.

A public discourse arguing on some extremely important moment-related issue is, most of time, an amalgam of arguments, rational forms, descriptions, stylistic procedures, which are

[1] "In times of internal or international crisis (...), we talk about managing various symbolic aspects of the role of: guardian of institutions, guarantor of national unity, moderator." [3]

intended to inform or to prepare a receptor in front of a problematic reality. But, as close to the subject a discourse would be, often it hides, in subtle ways, the true nature of the subjective thinking of the emitter. For instance, an exaggerated trust in the fresh energy of the society, in the benefits the loans on mortgage could bring to ordinary people, on the exceptional rise of the rate of interests, or on the incredible high bonuses certain banks are offering to their highly ranked employees could simultaneously bring the negative news, that something wrong is in the air, that a crisis is insinuating. Decoding of this hidden message, which is most of the time transmitted unintentionally, could be done only by someone extremely sensible to all facets of the financial and economic life. Signals for economic crises are issued by the central banks (e.g. Federal Reserve System, Central Bank of U.S., European Central Bank, etc.). During the period 2001-2008, when the banking system issued large but artificially cheap credits, there have been many public rhetoric appearances favourable to this behaviour which tried to set up an economic development investment with questionable prudence. Slogans like "a home for every American" (U.S.) or "credit with only an ID" (Romania), addressing a wide range of borrowers but having extremely low interest rates, could have been taken as signals of an economic potential crisis.

In this study we address the question: *Can an economic crisis be anticipated by evaluating public discourses from a lexical-semantic perspectives?*

We are interested to pursue a content analysis of the public language, using for that investigation tools that belong to the domain of natural language processing (NLP) and addressing: vocabulary (key words, frequent words), semantics (classes of concepts arranged in a hierarchy) and rhetorical-pragmatic discursive strategies (presence of the person I, preference for vague statements, generalities, etc.).

In U.S., the tradition of quantitative analysis is very strong, its roots being defined by Lasswell [5]. In Europe the interest grew more towards theoretical investigation of the semiotics of discourse ([1], [10], [1]). Modern content analysis is not only an illustration of a theory of text, but, should be rooted on empirical data. On the other hand, the American analysis is often neutral, technical, comparative, while the European analysis (especially the Critical Discourse Analysis model[2]) has a critical component and a strong enough ethicist.

In the perspective of our study, we are interested on public discourses (speeches), in written form, given by specialists on economy or by journalists, on economic issues. It is known that economy crises succeeds either a period of economic thrive or, as happened recently, a previous crisis. In our investigation we have used texts produced by most pertinent spokesmen which appeared in press materials issued by the Romanian National Bank (BNR), the most legitimate voice on economy issues in Romania. The other major filter in selecting the texts that should populate our corpus was the economic context (e.g. economic stability vs. economic crisis). A text categorization application filtered a stream of news that was considered of interest for our research. Some of the topics of interest have been: "credit ID only", "real estate boom", "mortgages", and "transactions with land or housing".

On another hand, at the base of our quantitative investigation was laid a lexical-semantic database. In order to assure generality, in acquiring it we had to use rather neuter sources, not necessarily tight to our specific corpus of texts. As such, the lexicon and the semantic classes have been collected from different sources usually dealing with economy themes: the BNR publications, already mentioned, but also a collection of dailies, *Ziarul financiar*, *Curierul Naţional*, *Bursa*, that have been monitored for a long period of time.

Current empirical approaches in analysing the public language put at work NLP techniques,

[2] "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 out a particular kind of delusion. Even with differing concepts of ideology, critical theory seeks to create awareness in agents of their own needs and interests." [11]

by which a multitude of features of the discourse were extracted and interpreted. 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. In this paper we describe a platform (*Discourse Analysis Tool* - DAT) specialised in the interpretation of the public discourse, which integrates a range of language processing tools with the intent to build complex characterisations of the public discourse. The idea behind it is that the vocabulary betrays discursive tonalities, this way allowing interpretations over the speakers orientation.

The paper is structured as follows. Section 2 shortly describes the previous work. Section 3 presents the DAT software and section 4 discusses an example of comparative analysis of economic discourses, elaborated during one year (2007-2008). Finally, Section 5 highlights interpretations anchored in our analysis and presents conclusions.

2 Previous Work

The aim of an interdisciplinary approach such as analysing the language of public speeches is to define and explain different discursive contexts, in our 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 [9].

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 software used to analyse the elections in United States in 2008. There are, however, important differences between the two platforms. LIWC-2007 basically counts words and increments counters associated with their declared semantic classes. DAT performs part-of-speech (POS) tagging and lemmatization of words. The lexicon contains a collection of lemmas (over 8800) for the POS categories of verb, noun, adjective and adverb, each being associated with one or more semantic classes. In the context of the lexical semantic analysis, the pronouns, numerals, prepositions and conjunctions, considered to be semantically empty, have been left out. Then, a special section of the lexicon includes expressions. An expression is defined as a sequence of stems of words. DAT includes now 33 semantic classes, chosen to fit optimally with the necessities of interpreting the public discourse, five of them having been added recently (`failures`, `nationalism`, `moderation`, `firmness`, `spectacular`). Then, another range of differences between the two platforms regards the user interface. In DAT, the user is served by a friendly interface, offering a range of services: opening and displaying one or more files, editing and saving the text, functions of undo/redo, functions of editing the lexicon, visualization of the mentioning of occurrences of certain semantic classes in the text, etc. The menus offers a whole range of output visualization functions, from the tabular form to graphical representations and to printing services. And finally, and most importantly, to help the user to interpret different authors simultaneously, she/he can chose among a collection of formulas that facilitate comparative studies.

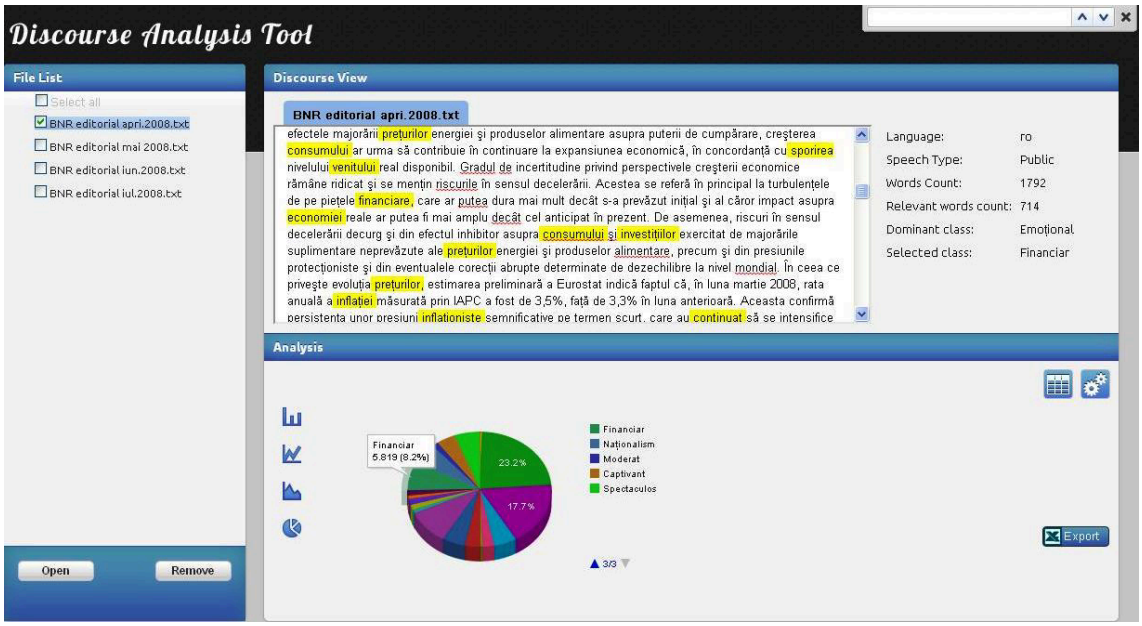

Figure 1: The DAT interface: in the left window appear the selected files, in the middle window - the text from the selected file, and in the right window, information about the text (language, word count, dominant class, etc.). Bellow, a plot chosen from a range of graphical styles is displayed. By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

3 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 characterisations of the public discourse. The concept behind this method is that the vocabulary used by a speaker betrays the authors sensibility, her/his level of culture, her/his cognitive world, and, by this, to the semantic spectrum of her/his speeches, 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).

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 fulfil 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. 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

Table 1: Examples of phrases on economy issues, on BNR editorials

Classes		Original in Romanian	English equivalent
financial	positive	creşterea PIB, expansiunea economiei mondiale, investiţii, scăderea ratei şomajului, expansiunea economică	PIB growth, global economic growth, investments, unemployment has declined, economic growth
	negative	moderarea ritmului de creştere a salariilor, gradul de incertitudine, turbulenţe pe pieţele financiare, efectul inhibitor asupra consumului şi investiţiilor	moderate the wage growth, uncertainty, financial markets turmoil, dampening impact on consumption and investment

report of the lexical-semantic analysis.

Our interest went mainly in determining those discursive attitudes able to betray an approaching recession. But the system can be parameterised to fit also other conjunctures: the user can define at will her/his semantic classes, which, as indicated, are partially placed in a hierarchy. Thus, for example, for the lemma *economist*, the following classes are assigned: 2 = `social` and 5 = `people`. The class `people`, is a subclass of the class `social`. These classes and their hierarchy are defined in a XML-like manner:

```
<class name="social" id="2">
<class name="people" id="5" parent="2">
```

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.

4 A Comparative Study

4.1 The corpus

The corpus used for our investigation was configured to allow a comparative study over the discursive characteristics of economic-financial themes, by including economy texts published on the BNR site in three different periods:

1. April-June 2007, when Romania crossed a period of economic stability.
2. April-June 2008, when Romania was near the economic crisis.
3. July 2008, when the Romanian president declared the economic recession.

Table 1 presents examples of phrases in the economy domain that exhibit two different discourse moods: positive emotional and negative emotional.

The analyzed texts were essentially dealing with the topics social and financial. After processing the texts with the DAT software, the following classes proved to have preponderant occurrences: `financial`, `social`, `work`, `emotional` (`positive` and `negative`), `rational` (`intuition`, `determine`, `uncertain`, `certain` and `inhibition`) and `nationalism`. To stress the distinguishing features, only these classes were finally left on the graphics.

4.2 The lexical-semantic analysis

We show in this section the results outputted by DAT when analysing the streams of textual data belonging to the three sections of the corpus (presented in section 4.1). For that, we have

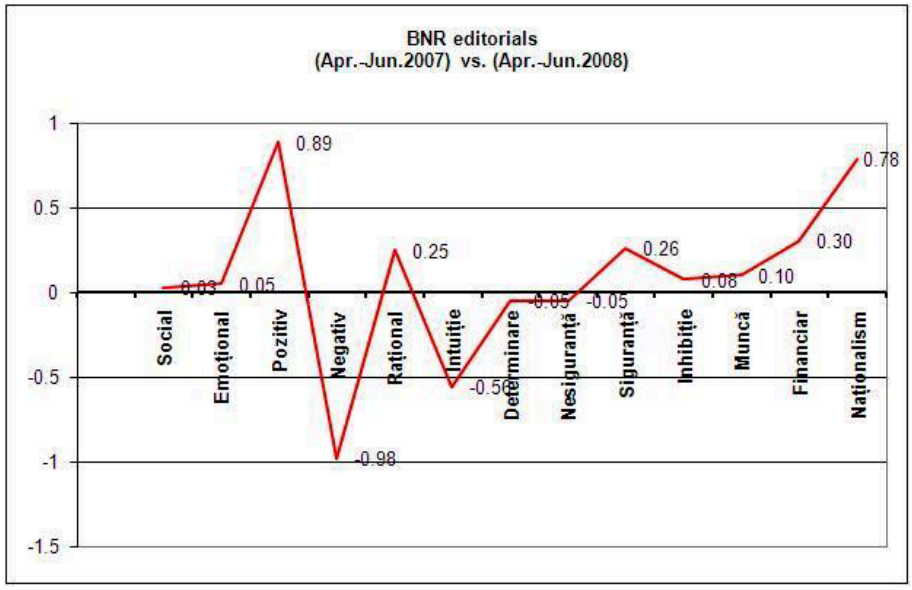

Figure 2: Difference between the occurrence of semantic classes in BNR editorials: one year before the economic recession versus three months before.

used the DAT feature of performing comparative studies. The values are supposed to reflect correctly the indicated classes, because they were computed by averaging on the whole collections of texts, not just a single text. The graphics considered for the interpretation computed one-to-one differences, as given by Formula 1, included in the DAT Mathematical Functions Library:

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

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 first element more than at the second element, and those below the horizontal axis show the reverse prominence. A zero value indicates equality. Our experience showed that values below the threshold of 0.5% should be considered as irrelevant and, therefore, were ignored in the interpretation.

So, the graphical representation in Figure 2, in which the editorials (Apr.-Jun. 2007) are compared against the editorials (Apr.-Jun. 2008) should be interpreted as follows: in 2007 the BNR discourse was extremely optimistic (high difference values of the class `positive`) and they were giving high importance to Romanian specific aspects (class `nationalism`), while in 2008 (nearly recession time) the BNR discourse had become rather pessimistic (class `negative`) and speculative (class `intuition`) with respect to the Romanian economic future.

In the following we will compare the same 2007 discourse against their discourse immediately after the recession.

The graphical representation in Figure 3, in which the editorials (Apr.-Jun. 2007) are compared against the editorials (July 2008) should be interpreted as follows: the difference in optimism between the BNR discourse one year before the recession and that of the moment the crisis was officially declared (class `positive`) is more pregnant (1.25% here versus 0.89% in Figure 2). However, although the pessimistic tone (class `negative`) is more pronounced in July 2008 than in the period of stability, it has weakened in intensity. We could say that BNR is caution to push too much on the distress pedal, because its voice could influence the fixing and, by that, worsen

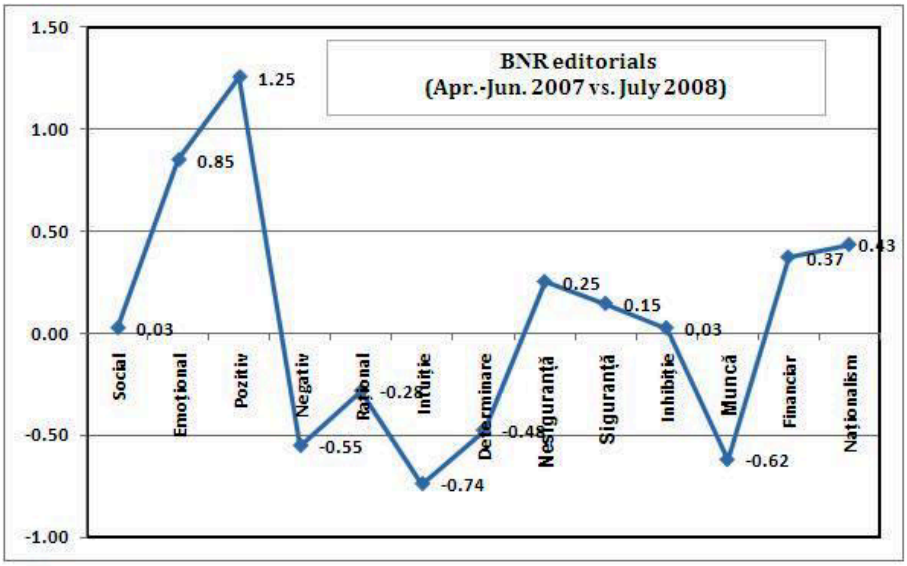

Figure 3: Difference between the occurrence of semantic classes in BNR editorials: one year before the economic recession versus one month after the economic recession.

the financial market even more. Moreover, BNR is offering a possible immediate solution, by accenting on the job sphere (class `work`).

5 Conclusions and Future Work

In this paper we presented a quantitative method and an application that strengthen the idea that crises can be anticipated by monitoring public speeches produced by representative entities.

We are aware that some of the differences which we have evidenced in our comparative study should partially be attributed to idiosyncratic rhetorical styles. However, when the traits inventoried acquire the regularities of patterns, then they could be used as measure apparatuses and, properly used, could emit useful signals to a receptive society.

There are a number of ways in which we think our research could be continued. First, we want to add new features to the platform, with a special emphasis on the syntactic and rhetorical levels of analysis. The new release of DAT should help the user to identify and count patterns of use at the syntactic and rhetorical level. 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. A weakness of the present system is the fact that the unequal sizes of the lexicons characteristic to semantic classes can influence the decisions: the more entries in the lexicon a certain class contains, the higher its influence could be foreseen. To this problem, the solution is not to balance the classes in their number of entries, because the language makes them intrinsically unequal, but to find calibration techniques that bring their values on equivalent ranges, irrespective of the dimensions of the lexicons. Let's note that in the present study we have counterbalanced somehow this skew by using the difference-based formulas (and thus avoiding absolute values).

Surely, the problem of characterising public speeches receives no final solution with our approach. We believe, however, that our method sheds an interesting light on possibilities of automatically interpreting discourses and, equally, it opens new perspectives.

Acknowledgements

In pursuing this research the authors had partial support from the projects POSDRU-63663, ICT-PSP 250467-ATLAS and ICT-PSP 270893-Metanet4U.

Bibliography

[1] Bürger, C., *Textanalyse als Ideologiekritik, Zur Rezeption zeitgenössischer Unterhaltungsliteratur*, Frankfurt am Main, Athenäum, 1973.

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

[10] Dijk van, T. A., *Sémantique générale et théorie des textes*, Linguistics, 62, 66-95, 1970.

[3] Gerstlé, J., *Comunicarea politică*, trad. Gabriela Cămară Ionesi, Institutul European, Iaşi, 94, 2002.

[4] 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*, Part II, CCIS 185, 188-195, 2011.

[5] Lasswell, H. D., *Politics: Who Gets What, When, How*, McGraw-Hill, New York, 1936.

[6] Lazarsfeld, P. F., Berelson, B., Hazel, G., *The Peoples Choice: How the Voter Makes up His Mind in a Presidential Campaign*, 3d ed., New York, Columbia University Press, 1944.

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

[1] Plett, H. F., *Ştiinta textului şi analiza de text*, trad. Speranţa Stănescu, Ed. Univers, Bucureşti: 72, 1983.

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

[11] Wodak, R., *Critical Linguistics and Critical Discourse Analysis*, Handbook of Pragmatics, Benjamins, 2006.