VIRTUAL CIVIC IDENTITY

DANIELA GÎFU[1], DAN STOICA[2], DAN CRISTEA[1,3]

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

{daniela.gifu, dcristea}@info.uaic.ro, dstoica_ro@yahoo.com

Abstract

The paper presents a study on a typology of civic identities of public contributors of online articles on forums and their possibilities of automatic identification. We analyse the dialogic means and exploration of automatic extraction of features from forum utterances. The research suggests new perspectives for defining types of online commentators of public discourses addressing domains such as politics, arts, education, etc. In the investigation we apply some pragmalinguistics approaches on communication, mainly taken from polyphony and enunciation areas. The classification of user profiles make use of criteria that take into consideration: common topics, expression of sentiments, style features, lexical n-grams, morphosyntactic analytics and pragmatic features. Our purpose was to lay the basis for a thorough classification of categories of publics and to suggest ways of their automatic identification, in the benefit of editors of media institutions, specialists in public communication, intelligence agencies, political structures, etc.

Keywords: civic identity, pragmalinguistics, semantic classes, journals forums, editors.

1. Introduction

Nowadays, a part of the reality has moved in the cyberspace. And the same happened with the bar or side-way chats, traditional in older times. Almost every public page we visit is cast into a stream of ongoing discussions, comments, gossips, thus becoming a property jointly owned by its composer and any person who may want to back react. The civic identity seems to be manifested on the Internet without constraints.

This study attempts to identify a model of identification of the *civic identity* of an individual, as revealed through online channels, by evidencing decision features whose values can be extracted automatically. The investigation focuses on a corpus of online journals' forums, from where the commentators' profiles are being extracted. Profiling the *civic identity* of readers of articles should exploit their inputs, therefore now seen in the position of writers of forums' short comments. The process puts at its basis a panoply of pragmatic markers, extracted by linguistic methods at the following levels: lexical (tracking patterns of specific vocabulary), syntactic (grammar errors, punctuation, enumerations, repetitions, use of emoticons, etc.), semantic (frequent use of some semantic classes), discourse (rhetorical markers). Using these features, the resulted portrait should be characterised along the following dimensions: the capacity to stay within article's topic, the capacity to express opinions on another forumist's comment, tendency of presenting themselves rather then following the forum's debate, degrees of assuming their respective identity as individuals,

preoccupation for really participating in the debate opened by the article (vs. just the desire to assert vague, general ideas).

When the media product is on the Internet, the actors of the cyberspace who decide to interact with it have tremendously numerous possibilities to shadow or even hide their identities, and, of course, their communicative intentions. As such, the attempt to determine the *civic identity* of people hiding their identities as individuals seems impossible. Up to date, there are no consistent instruments or studies on the different nature of forums' users, and the statistics are used just to group up reactions to the journalistic material. More than this, the lack of studies on the true nature of the forums' writers makes it impossible to apply advanced statistical calculi in order to rank positions or attitudes. A smart argument put out in well-formed phrases could reveal a civic activist, but also a good PR from a political party, trying to influence the readers of the forum; an upset man who wants to let it out on any subject could reveal a shy person accepting to express himself from behind the protection of the anonymity. Basic criteria like age, gender, education level are insufficient for determining the *civic identity* of people under study. The markers used by specialists in pragmalinguistic analysis could reveal in one's discourse a lot more on the personality of the writer than the writer would accept to unveil. Based upon this kind of findings, a typology of forums' users from the point of view of their respective civic identity is possible. This approach shows the importance of a natural language processing system capable to extract basic linguistic features from large amounts of online texts and to organize them as a collection of pragmatic knowledge aiming to inventory the profile of online commentators. The outcome of the study could provide tools for public speakers to be used for improving their future discourses. This is why the effort to mentally represent the interlocutor – and if not the actual interlocutor, the general profile s/he belongs to – is important in improving one's ability to communicate. There are many ways one could enhance his/her capacity of well representing the others before or during an online interaction. One of them is to analyze their public discourse, in order to extract information to be used in orienting your own discourse, making it efficient. A good apprehension of media products' respective publics, for example, could serve to improve their editorial politics and so be of better use for the communities they serve.

Section 2 presents the state of the art. Section 3, after a short description of the corpus analyzed, during the two hot months of the presidential crisis (July – August 2012), presents the methodology applied in identifying lexical-semantic and pragmatic features of the civic identity online. Finally, Section 4 presents some conclusions and directions for the future work.

2. State of the art

Our study combines automatic user profiling techniques (opinion mining, authorship classification) with pragmatic and linguistic studies of computer-mediated communications. In this moment, many systems collect various information about millions of people on the Web. Some of the current systems rely on the information manually provided by users. In others, information is obtained often from users' actions. In this case, user profiling requires inferring acquired information, both observable and unobservable data, such as, users' behaviour (Schiaffino and Amandi, 2009), (Zukerman & Albrecht, 2001). His/her behaviour and profile can be obtained from this information using different techniques like machine learning and statistical methods. Thus we have a wide range of techniques that were used to create user profiles, such as Bayesian networks (Nurmi, 2006), (Withby *et. al.*, 2005), (Weiwei *et. al.*, 2007), (Mui *et. al.*, 2001), (Garcia *et. al.*, 2007), fuzzy models (Grishchenko, 2004), (Sabater *et. al.*, 2002), (Manchala, 1998), association rules (Adomavicius & Tuzhilin,

2001), mechanisms of text classification (Trandabăţ *et. al*., 2012), (Gîfu & Cristea, 2012), and more.

Discourse/text output of users (posts, comments, forum messages) is used to infer elements about authors' identity (gender, sex, age, level of education and much more). In these text productions a user expresses his/her opinions about a given topic and interacts with other users. Content analysis is used in several applications to identify conflicts (Denis *et. al.*, 2012), or to detect various opinions (Grivel et Bousquet, 2011). The challenge is to involve theories of pragmalinguistics, mainly from the works on polyphony and enunciation (Ducrot et. Anscombre, 1989, Plantin, 2005 and also Tuţescu, 2005, Kerbrat-Orecchioni, 1999, Maingueneau, 2000). Language is no longer seen as a means to represent the world (referential function of the language), but as a means of argumentation in linguistic interactions among human beings. Enunciation is making a choice from the infinite offers of a given language: a choice of words, a choice of the order in which the words are uttered, a choice in the tone, the intensity of the voice and so on and so forth. Making those choices reveal a social profile of the enunciator will be our aim, and this is what we will try to track down, in order to set up patterns. We will search for patterns of linguistic behaviour that reveal patterns of social profiles. Trying to situate our research, we shall mention that the French revue HERMÈS published along the years papers on communication and the Internet, on social relations and the online communication, or on civic exchange in the cyberspace (Loh, 2009, Akiyoshi, 2009, Cardon, 2007, Oliveri, 2011), and also that (Holt, 2004) might be a model of how to use particularities of language use to determine the kind of citizen the speaker is. Email discussion messages are often expressed in a familiar register, with slang, abbreviations, and profanity and their composers frequently seem to delight in disregarding traditional rules such as those governing syntax, conventional logic, evidence and idea development, is the idea expressed by Holt in his *Dialogue on the Internet*. (Mortensen, 2003) discusses the use of language productions to understand the mind of a player. (Stoica, 2001) comments on the degrees of liberty authors have when writing for traditional, printed scientific journals and when they write for the web.

Pragmatic and rhetoric studies identify several relevant features for characterizing specific genres focusing on the expected audience (scientific articles vs. popular science articles (Hyland, 2009). Some research projects collect new media communication documents (Lin, 2007) (Stark and Dürscheid, 2011) to study their features for classification purposes.

3. A case study

The methods, the techniques and the tools in the development phase of the project create the premises for a thorough investigation of categorisation of online civic identities, drawn from statistics on large amount of textual data. The approach has a high degree of generality that makes it applicable to other types of investigations, provided they rely on text analytics.

3.1. The corpus

For the elaboration of preliminary conclusions on the configuration process of the online „civic identity", we collected, stored and processed 11,100 relevant texts/day/newspaper (summing up 146,000 words)[1], published during July-August 2012 (July 01-06, 2012 – a week before President's suspension; July 07-11, 2012 – a week after President's suspension; August 11-16, 2013 – a week before President's return at the Cotroceni Palace) by three

[1] We are aware that the actual dimension of our corpus is still insufficient to obtain an accurate categorization of the clasiffication criteria, but in this study we are merely interested to investigate a research methodology than to arive to precise conclusions over types of civic identities, as revealed by text analytics.

important Romanian online newspapers having similar profiles[2] (*Evenimentul zilei*, *Gândul*, *Jurnalul Naţional*) but usually displaying totally disjoint opinions and journalistic styles on any political topic. We talk about the hot political period when the President was suspended.

3.2. Methodology

In the following, we briefly describe the steps of our analysis:

- by attentive reading, we identified 10 typologies of commentators, that can be called: the-decent, the-porn-aggressive, the-incitator, the-linkable, the-affected, the-author-attacker, the-supporter, the-intellectual, the-rational, the-irrational (see. Table 1).

- after manually processing the whole corpus, it resulted that 6 that out of the 10 profiles were rather accidental (too few data): the-decent, the-porn-aggressive, the-linkable, the-author-attacker, the-supporter, and the-intellectual). As the average of their occurrences was under 5%, we eliminated these texts. Only the remaining 4 profiles are quantitatively analysed below.

- we established a number of features (belonging to the lexical, syntactic, semantic and discourse levels of analysis) that are, more or less, subject to automatic extraction: declared ID (hide, partial expose, expose, invented, etc.); making use of emoticons, familiarity in dialog, jokes, punctuation, etc.; the semantic classes of being rational emotional (with their sub-classes), and swear; comments that follow the topic, that have no correlation with the topic, that are connected to other comments, that are aggressive, etc.; number of appearances of the ID / article and the number of appearances ID in other online publication(s);

- all comments belonging to the same type, irrespective of their actual identity, have been put in the same folder, as belonging to the same type;

Table 1**:** Profile's typology after manually annotations

Abbreviations		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Style		caps lock, politeness formulas, quotes	porn formulas, k instead of c, sh, tz	familiar formulas, swears	links	emotional tone	familiar and aggressive formulas	too more punctuation marks	historical arguments	more info, official names	abbreviation, repeated punctuation marks
Rational		x					x		x	x	x
Positive				x							
Negative		x	x	x		x	x	x			
The nature of comment	related to topic	x							x	x	x
	off-topic		x	x		x					
	related to others						x	x			x
	incite commentators		x	x							
No. ID/articles		123	230	47721	230	19982	378	402	134	86650	32850
Profile's commentator		Decent	Porn-aggressive	Incitator	Linkable	Affected	Author-attacker	Supporter	Intellectual	Rational	Irrational

[2] These are national dailies of general information, tabloids with a circulation of tens of thousands of copies per edition, each. The newspapers were monitored on their websites: *Evenimentul zilei* – www.evz.ro, *Gândul* – www.gandul.info, *Jurnalul naţional* – www.jurnalul.ro.

- consequently, we adorned all texts with values on the established features, either manually or automatically. For instance, the semantic level has been automatically annotated with values for each of the semantic classes residing under the general classes: emotional, rational and swear, in total, 12 semantic classes;
- these data are discussed below as possible input for training a classifier to recognise the civic identities (portrait types).

3.3.Lexical-semantic features

After eliminating six of the manually annotated profiles, as identified initially, together with their comments, the remaining corpus was processed with the DAT[3] tool (initially intended to analyse political discourses). Out of the 33 semantic classes in DAT, arranged hierarchically – see two examples of XML class definitions in (1) –, we selected only those noticed to have dominant tonalities: rational, with 5 subclasses (uncertain, inhibition, intuition, certain, and determine), emotional with 2 subclasses (positive and negative), each of them having other 3 subclasses (positive with moderation, firmness and spectacular, and negative with anxiety, anger and sadness), and swear.

```
<class name="negative" id="8">
<class name="anxiety" id="8" parent="9">
```
(**1**)

The placement of classes in hierarchies makes that, when 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, be incremented.

For instance, in *Evenimentul zilei*, we can see the results outputted by DAT (Fig. 1), when analysing the streams of textual data for each semantic class. So, we analysed 4 profiles of online commentators (abbreviated with "C"), that we have considered to be predominated in cyberspace as follows:

- the first type of commentator, C5, predominate the self-confidence (the class certain), he is, rather, the type of dynamic blogger (the class emotional). In general, he comments in line with the subject, being convinced about his ideas (the class firmness);
- the second type, C10, is unsure (the class uncertain). He comments in line with the subject, because he looks for a way to get himself into the dialog;
- the third type, C3, has an insulting language (the classes swear, negative, anger). He prefers to shock the audience, in general he is out of subject or binds onto other commentators;
- the last type, C9, adopts a rational discourse (the class rational), with sustainable arguments (the class determine), and, often, he has a moderate tone (the class moderate) about the political topics.

[3] DAT (Discourse Analysis Tool) has some similarities with LIWC (Linguistic Inquire and Word Count), used during the American presidential elections in 2008 (Pennebaker, 2001). The Romanian lexicon resourcing DAT contains a collection of over 9,500 entries (roots and lemmas).

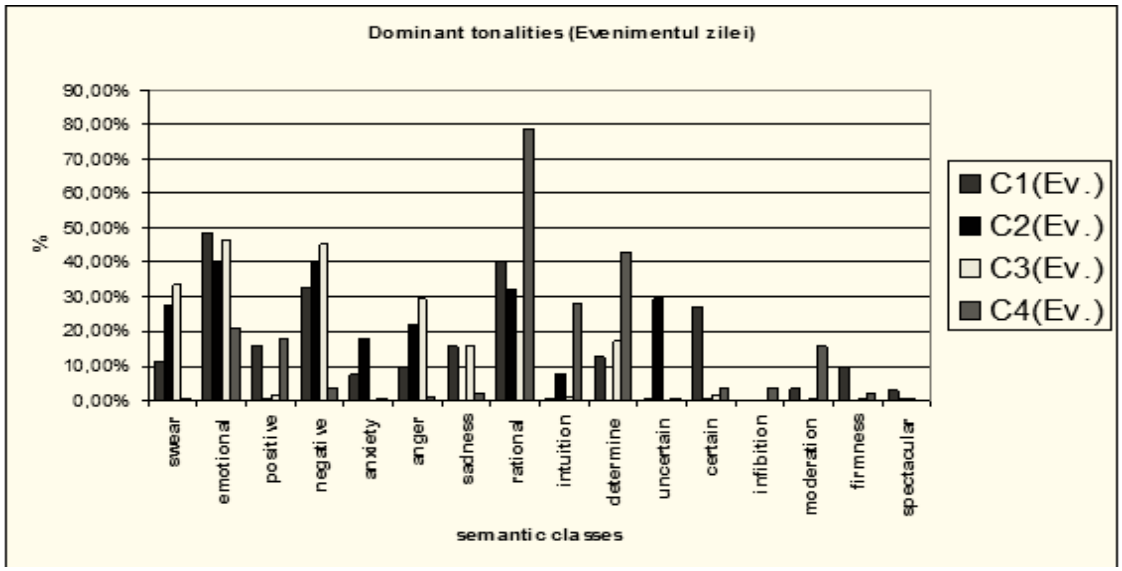

Figure 1: Analysis of user's profiles in *Evenimentul zilei* journal

3.4. A comparative lexical semantic analysis between two profile's online journals

We present below a chart with two streams of data, collected during the presidential crisis, representing comments between the two profile's online journals, *Gândul* and *Evenimentul zilei*. Our experience shows that an absolute difference value below the threshold of 0,75% 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 (2), included in the DAT Mathematical Functions Library:

$$Diff_{x,y}^{1-1} = average(x) - average(y) \qquad (2)$$

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. So, the graphical representation in Figure 2, where the commentator C1 of *Gândul* is compared against the commentator C1 of *Evenimentul zilei*, should be interpreted as follows:

- the first profile, C1, is much better argued than the second one (the classes rational, firmness), predominating self-confidence (the class certain), and uttered in an affective tone (the classes emotional, negative);
- the second profile, C1, is more emotionally implicated in comments, manifesting upset, even anxious (the classes anxiety and anger). He prefers to comment with sustainable arguments (the class determine), but, often with a precaution tonality (the class moderate) because he has no intention to start a dispute with the others.

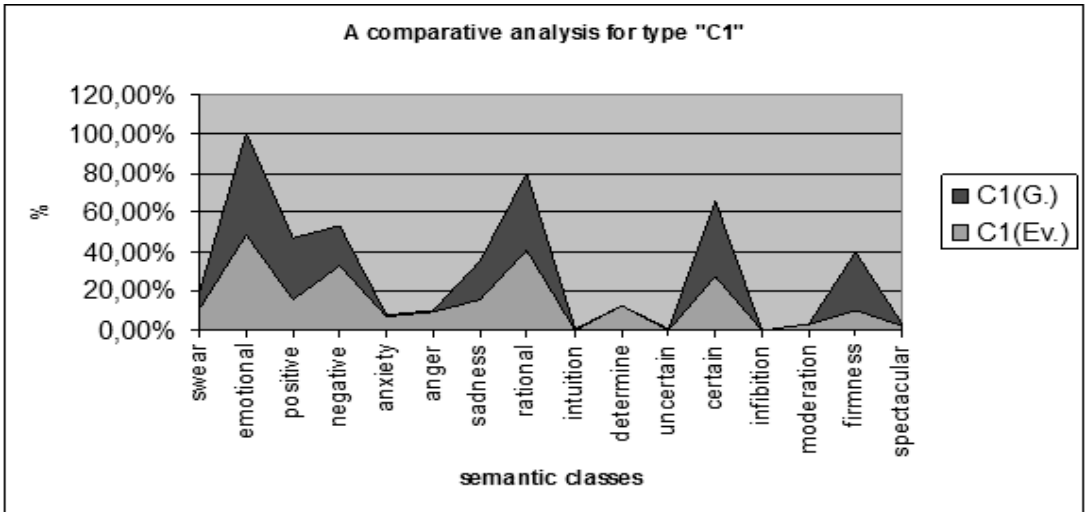

Figure 2 A comparative analysis between the users profiles in the journals *Gândul* and *Evenimentul zilei*

3.5. The pragmatic perspective

The pragmatic analysis should be based on the knowledge of the civic intentions of the commentator in connection with the meanings of the article or of the other comments. Only a good knowledge of the civic aspirations of the receptors and knowing that the editor knows himself this spectrum of civic aspirations, could make a human analyst succeed in interpreting the whole range of subtleties of a comment. It is clear that pragmatics makes a good deal of the forums interpretation process. It is nevertheless true that an experienced human analyst would succeed to acquire these facets of the pragmatic context of a comment 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 civic behaviour of the speaker or of the receptor from the text itself.

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

4. Conclusions

The discourse is a place where the personality is disclosed, but not at a level of certainty that could lead to establishing incontestable patterns. Furthermore, on Internet, possibilities for manipulating information are endless. Manipulation by people who design web sites or participate in discussion groups can give the clues whether the information on a site is reliable. However, by using statistical tools and pragmatic methods we will challenge these risks on safer ground than before.

Some features are mentioned earlier only for a theoretical reason, as their effective recuperation in the text by the technology is still out of the present day possibilities. For instance, the intentions of the online commentator, a feature falling into the pragmatic perspective, are not yet technologically feasible. An author of a text is conscious that he wakes a reaction onto the reader's mind, so her/his message has an intentional component (we talk mainly about conscious intentions, as they can be reflected in the author's and/or the editor's convictions about how the reader could be influenced). However, the automatic

detection of the authors' intentions, apart from the line of research triggered by the Attentional State Theory (Grosz & Sidner, 1986) and Rhetorical Structure Theory (Mann & Thompson, 1988), are still far from being conclusive.

This research opens a new direction for the study of online journals' commentators in areas such as: politics, culture, education, etc. The DAT tool becomes a necessary instrument of editorial policy and public relations departments. The study presented shows how one could shape profiles of commentators on forums of online publications (which are in a permanent dynamism). As the cyberspace is the perfect environment for hiding one's identity, some risks occur from this.

An analysis on the lines presented in this study could prove helpful to different categories of beneficiaries, mainly media editors and PR specialists. They could use the results of such analyses to better plan their policy, to adapt to different categories of public they might not even imagine be part of the general public (as they call it). Public segmentation is a continuous activity for PR specialists, and it has to be performed by using adequate criteria for each topic they want to develop in a discourse.

This kind of research could and will be continued further on: as society changes, media techniques change, the relation between media and their reader's changes all the time, and last but not least, the civic identity changes, but the need to know whom you can count on remains of paramount importance.

Acknowledgments: In performing this research, the first author was supported by the POSDRU/89/1.5/S/63663 grant.

References

Adomavicius, G., Tuzhilin, A. (2001). Using Data Mining Methods to Build Customer Profiles, *IEEE Computer* 34:2.

Akiyoshi, M. (2009). Les Japonais en ligne: le prisme des générations et des classes sociales, in HERMÈS (55).

Cardon, D. (2007). Le style délibératif de la «blogosphère citoyenne», in HERMÈS (47).

Denis, Al., Quignard, M., Freard, D., Detienne, F., Baker, M. and Barcellini, F. (2012). Détection de conflits dans les communautés épistémiques en ligne? Grenoble.

Ducrot, O. et. Anscombre, J.-C. (1989). Logique, structure, énonciation. Lectures sur le langage, Minuit.

Garcia, P., Amandi, A., Schiaffino, S., Campo, M. (2007). Evaluating Bayesian Networks' Precision for Detecting Students' Learning Styles. *Computers and Education* 49:3.

Gîfu, D., Cristea, D. (2012). Multi-dimensional analysis of political language. Future Information Technology, Application, and Service: FutureTech2012 (volume 1) Springer, Netherlands (James J. , Jong Hyuk Park, Victor Leung, Taeshik Shon, Cho-Li Wang Eds.)

Grishchenko, V. (2004). A fuzzy model for context-dependent reputation, at the Trust, Security and Reputation, *Workshop at ISWC*, Hiroshima, Japan.

Grivel, L., Bousquet, O. (2011). A discourse analysis methodology based on semantic principles - an application to brands, journalists and consumers discourses, *Journal of Intelligence Studies in Business* 1.

Grosz, B.J., Sidner, C.L. (1986). Attentional State Theory, *Journal of Computational Linguistics*, 12:3, 175-204, The MIT Press Cambridge, MA, USA.

Hyland, K. (2009). Academic Discourse: English In A Global Context (Continuum Discourse).

Holt, R. (2004). Dialogue on the Internet. Language, Civic Identity, and Computer-Mediated Communication, Westport, Conn., PRAEGER.

Lin, J. (2007). Automatic Author Profiling of Online Chat Logs, M.S. Thesis, Naval Postgra duate School, Monterey.

Loh, C. (2009). Une ancienne députée de Hong Kong sur la Toile: le site «Civic Exchange» (entretien avec Éric Sautedé), in HERMÈS, (55).

Kerbrat-Orecchioni, C. (1999). L'énonciation : De la subjectivité dans le langage. 4-ème édition. Paris: Armand Colin.

Maingueneau, D. (2000). Analyser les textes de communication. Paris: Nathan.

Manchala, D.W. (1998). Trust metrics, models and protocols for electronic commerce transactions, *Proceedings of the 18th International Conference on Distributed Computing Systems*.

Mann, W.C., Thompson, S.A. (1988). Rhetorical Structure Theory. Toward a functional theory of text organization, *Text - Interdisciplinary Journal for the Study of Discourse*. 8: 3, 243–281.

Mortensen, T. E. (2003). Pleasures of the player. Flow and control in online games, Volda University College.

Mui, L., Mohtashemi, M., Ang, C., Szolovits, P., Halberstadt, A. (2001). Ratings in distributed systems: a Bayesian approach, *Proceedings of the Workshop on Information Technologies and Systems* (WITS).

Nurmi, P. (2006). A Bayesian framework for online reputation systems, *Proceedings of the Advanced Int'l Conference on Telecommunications and Int'l Conference on Internet and Web Applications and Services*.

Pennebaker, J. W., Francis, Martha E., Booth, R. J. (2001). Linquistic Inquiry and Word Count – LIWC2001, Mahwah, NJ, Erlbaum Publishers.

Plantin, C. (2005). L'Argumentation, PUF, Que sais-je?

Pragmatics, (2006)/(2011). Metaphysics Research Lab, CSLI, Stanford University.

Oliveri, N. (2011). La cyberdépendance: un objet pour les sciences de l'information et de la communication, in HERMÈS (59).

Sabater, J., Sierra, C. (2002). Social ReGreT, a reputation model based, on social relations, SIGecom Exchanges 3.1.

Schiaffino, S., Amandi, A. (2009). Intelligent user profiling, *Artificial intelligence*, Lecture Notes In Computer Science, Vol. 5640. Springer-Verlag, Berlin, Heidelberg (Max Bramer ed.).

Stark, A., Dürscheid C., (2011). SMS4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland, *Crispin Thurlow/Kristine Mroczek (Hrsg.): Digital Discourse. Language in the New Media*. Oxford: Oxford University Press.

Stoica, D. (2001). Modalités de la communication scientifique, în NOESIS. Travaux du Comité Roumain d'Histoire et de Philosophie des Sciences, vol. XXVI. Bucureşti, Editura Academiei Române.

Trandabăţ, D. Irimia, E., Barbu Mititelu, V., Cristea, D., Tufiş, D. (2012). The Romanian Language In The Digital Age. META-NET White Paper Series, Springer.

Tuţescu. M. (1998). L'argumentation. Introduction a l'étude du discours, Bucureşti, Ed. Universităţii.

Zukerman, I. and Albrecht, D. (2001). Predictive Statistical Models for User Modeling. User Modeling and User-Adapted Interaction, 11(1-2).

Weiwei, Y., Donghai, G., Sungyoung, L., Young-Koo, L., Heejo, L. (2007). Bayesian Memory-Based Reputation System, in Proceedings of the 3rd international conference on Mobile multimedia communications.

Withby, A., Josang, A., Indulska, J. (2005). Filtering Out Unfair Ratings in Bayesian Reputation Systems, Journal of Management Research.