Monitoring and predicting journalistic profiles

Daniela Gîfu¹, Dan Cristea^{1,2},

¹„Alexandru Ioan Cuza“ University, Faculty of Computer Science,
16, General Berthelot St., 700483, Iași
{daniela.gifu, dcristea}@info.uaic.ro

²Institute for Theoretical Computer Science, Romanian Academy - Iași branch,
2, T. Codrescu St., 700481, Iași

Abstract. The paper develops a pilot study aiming at finding a methodology for identifying features of journalistic writing profiles. The study is based on capturing dominant discursive tonalities, knowing that journalistic discourse entails public legitimacy. We made use of a series of natural language processing tools as preliminary steps in revealing three egocentric journalistic identities. Based on quantitative analysis, such as syntactic and lexical-semantic, we put in evidence qualitative pragmatic aspects. The final goal of this study is to configure a tool for automatic analysis of journalistic profiles. We have concentrated on three top Romanian journalists, whose newspaper publications have been monitored over a period of three months and semi-automatically analyzed. As a result, a database with their profiles is collected and interpreted. Such a tool could be of interest to mass-media, but also to specialists in communication and public relations, to political parties and to the public opinion, in general.

Keywords: journalism, lexicon, syntactic analysis, semantic analysis, egocentrism features.

1 Introduction

The automatic identification of the features of journalistic profiles, although in the attention of researches in Natural Language Processing, is currently a problem only partly solved. We rely on the human capacity to assign significances¹[21] of words, accepting the argument that the human beings have the potential to understand the natural language. That is why, current methods for identification of discursive tonalities, by using statistical or symbolic algorithms, approximate the human ability to classify, to identify the author's position and the nature of his opinions. The interest is focused on identifying the linguistic modalities to persuade or seduce an auditorium about the truth of the speaker/writer's ideas. A well-crafted argument in the architecture of a discourse can reveal the intellectual nature of a journalist which has as principal purpose to serve editorial policies through his articles.

A technique for profiling the author of a text is given by the use of linguistic markers that can reveal much more about the author's personality, often, more than he

¹ Here, the significance concept is synonymous with understanding.

reveals by other ways (e.g. in an interview). This approach emphasizes, in fact, the importance of using a natural language processing system able to extract basic linguistics features on a large amount of texts, that can be arranged as a collection of pragmatic knowledge in order to inventory the features of journalistic profiles.

Such a methodology could be useful for communication and public relations specialists who build, suggest, etc. discursive structures of different public actors that represent them (press, politicians, economists, secret services and so on). One of the reasons for the effort to represent mentally journalistic portraits (identifying the discursive nature of a journalist) is important in improving communication. Another reason is that it can help to improve the editorial policies and thus to offer a better adequacy of the journalistic thesis to the expectations of the community and the needs they represent.

Section 2 presents the state of the art. Section 3 depicts a new vision about linguistics boundaries between *I* (*we*) and *you*, then in section 4, after a short description of the corpus analyzed, we present the methodology applied in identifying the *egocentrism* features of the journalist profile. Finally, Section 5 presents some conclusions and directions for the future work.

2 State of the art

Our study combines tools that enable classification of the texts [7] and automatic recovery methods of text information monitored applicable pragmalinguistics studies that is used in journalistic language in order to identify features of journalistic profiles [15], [22]. Of major importance are some researches based on the collection of new media texts, used in identifying characteristics underlying implementation of text classifiers [11], [17], [13]. A comparative analysis of n texts signed by n different journalists provides the criteria or heuristics which can be implemented in a platform for natural language processing for characterization of each document and each signatory of text. When two or more collections of texts, each having a single author are compared, it can be determined statistically, to which typology it belongs. These criteria of differentiation can be: the use of longer sentences, with specific ways of structuring the sentence/paragraph; the use of special means to divide words at the end of the row, using the abbreviations (with points between each letter, with no points, all letters are uppercase), the use of particular expressions or phrases with a significantly higher frequency than the common language, using words and/or phrases in another language, use of quotes; using a specific number/relationship of words (adjectives, adverbs) or preferences for specific classes of pronouns; an author prefer some specific morphological variations and spellings of words (replacing diacritics with similar groups of letters); the use of punctuation, or emoticons that become elements of nonverbal or paraverbal language etc.

Simple features such as the (punctuation, functional words, and n-grams) are too common to be used in order to characterize the style of an author. Some researchers use the frequency of part-of-speech (POS) tags [14] to classify user profiles from a transcribed dialogical relationship. Features like: average number of words per sentence, average number of letters in a word and use of punctuation marks are often encountered in the studies leaned towards the profile identification of authors [4]. In

this context, it is important to remember the lexical aspects and types of discourse (including the choice of words, the type of participation in communicative act) for the identification of the discursive features [6]. A number of linguistic and pragmatic studies aiming the identification of characteristics of a particular type of speech, or an author [10] make reference to: argumentation [3], preference for the personal pronouns (first person, singular and plural and second person plural), exclamations and rhetorical questions, etc. Some platforms are known as to man-machine mediated interaction [2], [12], based on automatic learning to enhance the skills of detection for attributes in the identification of authors profiles for certain texts [20].

One aspect in our study is the profiling of a discursive style starting from the type of opinion and how it is expressed, manual annotation being a first step in this sense. In the production of the text to which we make reference, a journalist is expressing her/his opinions about a particular subject and interacts with an imaginary audience. Content analysis is used in many applications to capture potential conflicts or to detect discursively the various opinions of the signatory [8]. For the Romanian language it was introduced a similar system of annotation [18], which does not exclude, however, subjective aspects. These approaches use specific dictionaries [9] or emotional intensity [5].

By making these choices we can trace the socio-pragma-linguistics profile of the enunciator, otherwise a part of our purpose, correlating certain computational techniques with pragmalinguistics theories.

3 Pragmalinguistics boundaries

If social boundaries are strategies of social policy, by extrapolation, pragmalinguistics boundaries are strategies of discursive intentionality. These expression processes are involved in the construction of the identity of the speaker, being it wilfully, coerced or involuntary. We have introduced here the *boundary* concept just to clarify that we are not talking of limit² but the separation of the two protagonists, which define an act of communication (transmitter and receiver). The representation of the communication between the two parts of this split communication space is here decisive: "boundaries are not air tight, are never occlusive, but more or less fluid, moving and permeable [19]. The dividing line [1], which separates "I"/"we" from "you", is vital for their identity. For example, the relation "I-we" allows the definition of membership of the journalist to the editorial team, known as the signers of the articles of the print media apt their speech to the political editorial which they represent.

Pragmalinguistics boundaries become tools that make it possible to identify inclusions and potential exclusions from the editorial board (we exemplify with the journalist Ion Cristoiu, who, for a while, was one of the most vocal opponents of media that he serves now). In Figure 1, we have sketched the process of communication initiated by a journalist: the publication – represented by "I"/"we" – is addressed to the public opinion – "you". The linguistic boundaries, in fact pragmalinguistics, are distinguishing traits that the addressee of the article would

² Limitations in the communication sciences, known as communicational barriers, are a series of obstacles that arise between transmitter and receiver and that reduce the efficiency of messages sent within a process of communication.

want to identify in the language of the journalist, but which can also be a barrier for deciphering the correct message. This may be due either to a wilful intent on the part of the media player (the journalist or the publication, through its policies) or an unconscious intention (the use of a journalistic code, understood by the receiver³). The intentioned discursive purpose when using “I” is a modality of shaping a pragma-discursive dimension.

Fig. 1. Communicational process (“I” ± “We” and “You”) .

The use of “I”/“we”, on one hand, and “you”, on the other, are signs that can help in the process of classification of an egocentric typology. In this study we restrain at doing a simple statistics that would subsequently help in advancing a more rigorous investigation over the distance that the writer interposes between himself (or his publication) and his readers.

4 A case study

The method and instruments for processing natural language used in this paper confirm the premise that *investigating journalists' identity in pragmalinguistics terms can be captured based on statistics depicted from a corpus of texts*. Since language suffers a perpetual metamorphosis, and the one of the virtual media is even more rapidly changing, it would be good if the corpus would be acquired over different periods of time. Although we believe that the proposed approach has an important degree of generality, allowing for inclusion of more investigation axis, we are also aware that the analyzed corpus in this initial research is still insufficient for drawing crisp egocentrism classification conclusions. What we propose, therefore, is merely a methodology for typifying journalistic styles, focused on egocentrism elements. The exercise is drawn around some of the Romanian editorialists with great outlet to the public.

³ For instance, when you write in other language than that spoken by the receiver.

4.1 The corpus

For the elaboration of preliminary conclusions on the configuration process of the journalist identity, we collected, stored and processed 1,463 relevant articles (summing up 28,879 words), published by three journalists: Cristian Tudor Popescu (CTP), Marius Tucă (MT), Ion Cristoiu (IC), during January – March 2013 by three important Romanian newspapers having similar profiles⁴, but usually displaying totally disjoint opinions and journalistic styles on any topic.

4.2 Methodology

We present briefly the accomplished steps:

- by attentive reading, we identified 3 typology of egocentrism journalistic, that can be called: *egocentric (self-)ironic*, *egocentric puffy*, and *egocentric all-knowing*.
- we established a number of features (belonging to the syntactic, lexical-semantic and pragmalinguistics levels of analysis) that are, more or less, subject to automatic extraction: the use of the personal pronouns, familiarity in communication, ironies, punctuation, etc.; the semantic classes of being rational and emotional (with their sub-classes), nationalism, sexual; comments that are aggressive, etc.; number of enumerations/article and the expressions with emotional content;
- we have tagged all texts for POS in order to highlight the personal pronouns first person⁵, singular and plural, but also second-person, plural. All sentences containing personal pronoun forms have also been semantically analyzed (using the DAT software) and pragmatically (qualitative analysis).

4.3 The pragmalinguistics analysis

In order to proceed with the syntactic analysis, the text bodies were annotated with syntactic information, in XML. Two sources of information have been used, involving manual (200 units⁶ for each journalist monitored) and automatic annotation (350 units – for the first, 210 – for the second, and 483 – for the third). The manually annotated segments (see Table 1) included the interrogative and exclamatory count, and all pronoun forms, first person, singular and plural, but, also, second person, plural (nominative, dative, and accusative cases). After highlighting the characteristics of the specified syntax, we were interested to see in what phrase structures are the personal pronouns used, operation performed using POS tagging [16].

⁴ 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 național* – www.jurnalul.ro.

⁵ Here we have also reserved the lemma *subsemnatul* (undersigned), which replaces the pronoun "I" in the nominative case.

⁶ Sentences.

We can conclude about the syntactic structure (Table 1) for each journalist as follows:

- the first type of journalist, CTP, prefers medium length texts (approx. 22 sentences/article). His interrogative (50) and exclamatory (32) sentences have a rhetorical purpose; the audience can or should be able to reflect to CTP's opinions.

Certain pragmalinguistics boundaries can be identified in the speech of this journalist:

1. personal pronoun, first-person, singular - evokes in the sentence, through personal pronouns "I", entities present implicitly or explicitly in the universal speech, the communication situation being defining (e.g. *because I want to contribute with what I can to bring viewers in the Romanian theatre*).
2. personal pronoun, first-person, plural - empathises with public opinion (e.g. *The funny parliamentary Becali is invited several times a day to televisions by us in order for him to make a mockery of the rule of law, common sense and human dignity, to insult women, to curse men*).
3. personal pronoun, second-person, plural - assuming the role of very good connoisseur of governmental management (e.g. *The reasoning seems wrong for you, dear readers, sick or healthy?*).

Table 1. Syntactic descriptors used in this research

Descriptors	CTP	MT	IC
I, sg. (<i>eu, îmi, mi, m, mine, mă</i>)	29	25	118
I., pl (<i>noi, ne, ni</i>)	3	14	30
II, pl. (<i>voi, vouă, vă, vi, v</i>)	2	25	5
Sentences	22	7	43
Exclamatory	22	18	99
Interrogative	50	34	62

- the second journalist investigated, MT, is very expeditious (approx. 7 sentences/article), but often the receptor must find the answers to his interrogations (34). The reader can discover also some aggressive exclamatory sentences (18).

From the perspective of pragmalinguistics boundaries, this journalist uses:

1. personal pronouns, first-person, singular - in contexts with nostalgic tint, sometimes anxious (e.g. *I love break-ups over the shoulder*).
2. personal pronoun, first-person, plural - emphasizes, easy to recreate, in this sense, the name of the daily newspaper that he leads and his quality in the editorial. (e.g. *Contacted by the National Journal, Alina Alexandra Mihai, winner of the Giumbix contest, told us how her idea came out...*).
3. personal pronoun, second-person, plural - the clear delimitation, even pornographic, from others (e.g. *It fills your screen with boobs and butts / sex is on money, do not hurry,/don't masturbate, because you will use the credit card.*)

- the third type of journalist, IC, writes very long articles (approx. 43 sentences/article), because he insists in details in any political subject (demonstrating to be a mature political analyst). It is noted that he uses the personal pronoun "I" more often than the other two journalists, denoting a higher concern towards himself. Here are a few examples:

1. the personal pronoun, first person, singular – emphasizing his experience as a journalist, but also his ideological position: *In 2003, I was, as a director, at the forefront of the Realitatea TV, shepherded by Silviu Prigoană.*
2. the personal pronoun, first person, plural – underlines the verticality (positioning himself in the good camp) in relation to the wicked. Here's a snippet of speech: *The distinguished are left with the millions, and we, the rest, with the honour.*
3. the personal pronoun, second person, plural – clear induction in the eyes of public opinion to the camp of morality he makes part of. For instance: *Put yourselves in our job and you are going to be fine!*

4.4 Lexical-semantic analysis

The corpus was processed with the DAT⁷ tool (a platform for lexical-semantic analysis of public discourse). To identify the predominant tonalities in the discourses of each journalist, we included in this case study only 15th semantic classes, arranged hierarchically: positive (with 3 subclasses: spectacular, firmness and moderation), rational (with 5 subclasses: uncertain, inhibition, intuition, certain, and determine), and negative (with 3 subclasses: anger, anxiety, sadness), and sexual.

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 will be incremented.

Bellow, we have the results outputted by DAT (Fig. 2), when analysing the streams of textual data for each semantic class. The 3 profiles analysed (CTP, MT and IC) can be interpreted as follows:

- the discourse of the first type of journalist, CTP, is predominantly negative in emotional tonality (class negative), in two different intensities (classes anger and sadness). In general, he prefers ironical expresses (for instance: *His work is a systematic and tenacious huge collection of kitsch and clichés*). Although he is a well known journalist, sometimes he mentions that he is a specialist in the cinema art. (e.g. *I do it as a cinema-goer*). Also, he prefers to use expressions in the English language (for instance, *Who's that stumblin' around in the dark? State your business or prepare to get winged! or Auf wiedersehen, Bullseye! or Alexandre Dumas is black and so on*), but in an ironic way. We will call this type, *egocentric (self-)ironic*.
- the second type, MT, is a dynamic guy (class positive) and prefers metaphorical languages (class spectacular) but, often, in a pornographic tone (*don't masturbate, because you will use the credit card, / Look only at the nipples, like some teenagers*). Most of the times, he has in attention the themes presented in his show, "The MT Show". He takes every opportunity to promote their own journalistic press trust (for instance: *a new TV rubric in the MT Show*). He writes short, confidently in himself (class certain), often in verse, because he enjoys being a bohemian wistful (e.g. *How many question marks/sending a riot lifestyle/of a desertion from proper*

⁷ DAT (Discourse Analysis Tool) has some similarities with LIWC (Linguistic Inquire and Word Count), used during the American presidential elections in 2008 [Pennebaker, J. W.]. The Romanian lexicon resourcing DAT contains a collection of over 9,500 entries (lemmas).

decorum/things/people,/ to be able to respond.). We will call this type, *egocentric puffy*.

- the third type, IC, has a rational discourse (class *rational*) is very convinced of his ideas. He prefers long texts to explain in a determined way (class *determine*) the chosen subject (especially, with political flavour). He is an old fashioned journalist; his articles have a title, a short resume and a long body). Actually, all his texts appear under the heading *Romania's C*.

Fig. 2. The comparative semantic analysis for journalistic articles

He has an indulgent point of view over the President (for instance, *these days that were spent, a few things have made me realize that the president was right*), even if not long time ago he had a different opinion. Now he attacks the Ponta's Parliament, situated on the opposite side of the president (e.g. *I noticed first of all Victor Ponta's inability to overcome the posture of politician opposition, essentially babble*). We will call this type *egocentric all-knowing*.

5 Conclusions

The discursive-*egocentric* differences, found in this study, can be attributed partly to idiosyncratic rhetorical styles, and partially, to the ethics the authored adhered to (editorial policy applied in the public space). Of course, in the current context, ethics is confused with a specific ideology that belongs to a specific editorial group. No less important in identifying the pragmatolinguistic characteristics is the cultural universe and the generation to which belongs the signatory of a press article.

We believe that the findings revealed in the present study may lay the basis for the delineation of a journalistic identity that brings in the space of the Romanian

journalism critics an expansion of the possibilities of public discourse analysis by computer mediated techniques.

We are aware that the corpus of manually annotated texts is still in an early phase and this study should be understood only as allowing to perform a pilot study towards a statistical investigation on a larger corpus, that would be used in a process of automatic learning, such that, in the future, the machine be capable of efficient automatic annotation. Right now we are testing the feasibility of using our natural language processing instruments in the automatic analysis of journalistic texts. Not less important if to reveal new significant features of discourse which, on one hand, could be automatically extracted from the text and, on the other, are useful for our goals. It is also important to perform tests that would reveal to what extend the authorship types identified (3 – after the present study, maybe more – after a more rigorous one) could be clearly delimited by statistical means.

Another focus of attention in our future research is towards supervising the public, as commenter of the journalistic texts, an aspect which has not received much attention to date. In doing all these, we look also to adapt our instruments to other languages as well, which, among other things, would allow us to compare the results obtained in the Department against those published elsewhere.

Acknowledgments: In order to perform this research the first author received financial support from the POSDRU/89/1.5/S/63663 grant. We are grateful to Radu Simionescu, from the NLP-Group@UAIC-FII, for POS-tagging the Romanian corpus.

References

1. Barth, F.: *Ethnic Group and Boundaries*, Bergen, Oslo, Universitetsforlaget (1969).
2. Denis, Al., Quignard, M., Freard, D., Detienne, F., Baker, M. and Barcellini, F.: *Détection de conflits dans les communautés épistémiques en ligne?* TALN 2012, Grenoble, France.
3. Ducrot, O. et. Anscombre, J.-C.: *L'argumentation dans la langue*, Mardaga (1983).
4. Forsyth, E, Martell, C.: *Lexical and Discourse Analysis of Online Chat Dialog*. In: *International Conference on Semantic Computing* (2009).
5. Eensoo, E. and Valette, M.: *Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments*. In: *Proceedings of TALN 2012, Grenoble, France*.
6. Garera, N., Yarowsky, D.: *Modelling Latent Biographic Attributes in Conversational Genres*. In: *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, Suntec, Singapore, 2-7 August 2009, pp. 710–718 (2009).
7. Gifu, D., Cristea, D.: *Multi-dimensional analysis of political language*. In: J.J. (J. Hyuk) Park, V. Leung, T. Shon, C. Wang (Eds.), *Future Information Technology, Application, and Service: FutureTech2012* (volume 1) Springer, Netherlands, pp. 213-221 (2012).

8. Grivel, L., Bousquet, O.: A discourse analysis methodology based on semantic principles - an application to brands, journalists and consumers discourses. In: *Journal of Intelligence Studies in Business* 1, pp. 76-86 (2011).
9. Iftene, A., Rotaru, A.: User Profile Modelling in eLearning using Sentiment Extraction from Text. In: *Research in Computing Science, Special issue: Natural Language Processing and its Applications*, vol. 46, pp. 267-278, Mexico (2010).
10. Kerbrat-Orecchioni, C.: Analyse des conversations et négociations conversationnelles . In M. Grosjean et. L. Mondada (éds.) *La négociation au travail*, Lyon, PUL/ARCI, pp. 17-41 (2004).
11. Lin, J.: Automatic Author Profiling of Online Chat Logs, M.S. Thesis, Naval Postgraduate School, Monterey (2007).
12. Lortal, G., Todirascu-Courtier, A., Lewkowicz, M: AnT&CoW:Share, Classify and Elaborate Documents by means of Annotation. In *Journal of Digital Information Management*, (eds. Richard Chbeir, Ajith Abraham, Pit Pichappan), no 6(1), pp. 61-70 (2008).
13. Pennebaker, J. W., Francis, Martha E., Booth, R. J.: *Linguistic Inquiry and Word Count – LIWC2001*, Mahwah, NJ, Erlbaum Publishers (2001).
14. Portele, T.: Data-driven Classification of Linguistic Styles. In *Spoken Dialogues. COLING (The 19th International Conference on Computational Linguistics)* (2002).
15. Schiaffino, S., Amandi, A.: Intelligent user profiling. In *Artificial intelligence, Max Bramer (Ed.). Lecture Notes In Computer Science*, Vol. 5640. Springer-Verlag, Berlin, Heidelberg, pp. 193-216 (2009).
16. Simionescu, R.: POS-tagger hibrid. Dissertation at the “Alexandru Ioan Cuza” Universitatea of Iași (2011).
17. Stark, A., Dürscheid, Christa: SMS4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland. In Crispin Thurlow/Kristine Mroczek (Hrsg.): *Digital Discourse. Language in the New Media*. Oxford: Oxford University Press, pp. 299-320 (2011).
18. Tufiş D., Ştefănescu D.: A Differential Semantics Approach to the Annotation of Synsets in WordNet. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, Malta, May 2010, ELRA.
19. Wimmer, A.: The Making and Unmaking of Ethnic Boundaries: A Multilevel Process Theory. In *American Journal of Sociology*, 113 (4), pp. 970-1022, (2008).
20. White, B.Y., Frederiksen, J.R.: Inquiry, modelling, and meta-cognition: Making science accessible to students. In *Cognition and Instruction*, 16, pp. 3-118 (1998).
21. Wittgenstein, L.: *Cercetări filozofice*, trad. de Mircea Dumitru şi Mircea Flonta Ed. Humanitas, 117 (2004).
22. Zukerman, I. and Albrecht, D.: Predictive Statistical Models for User Modeling. In *User Modeling and User-Adapted Interaction*, 11(1-2), pp. 5-18 (2001).