On the Stylistic Evolution from Communism to Democracy: Solomon Marcus Study Case

Anca Dinu
Faculty of Foreign Languages and Literatures
University of Bucharest
anca_d_dinu@yahoo.com

Liviu P. Dinu
Faculty of Mathematics and Computer Science
University of Bucharest
ldinu@fmi.unibuc.ro

Bogdan C. Dumitru
Faculty of Mathematics and Computer Science
University of Bucharest
bogdan27182@gmail.com

Abstract

In this article we propose a stylistic analysis of Solomon Marcus' publicistics, gathered in six volumes, aiming to uncover some of his quantitative and qualitative fingerprints. Moreover, we compare and cluster two distinct periods of time in his writing style: 22 years of communist regime (1967-1989) and 27 years of democracy (1990-2016). The distributional analysis of Marcus' text reveals that the passing from the communist regime period to democracy is sharply marked by two complementary changes in Marcus' writing: in the pre-democracy period, the communist norms of writing style demanded on the one hand long phrases, long words and clichés, and on the other hand, a short list of preferred "official" topics; in democracy, the tendency was towards shorter phrases and words, while approaching a broader area of topics.

1 Introduction

Authorship attribution became an important topic in the last decade, not only in computational linguistics, but also in applied areas such as forensics, journalism, education, etc. The most common scenario is the following: given some known sample documents from a small set of candidate authors, which of them wrote a certain document with unknown authorship ? (Koppel et al., 2009; Luyckx and Daelemans, 2008) A related question is author clustering, where, given a document collection, the task is to group documents written by the same author, so that each cluster corresponds to a different author.

Marcus (1989) identifies the following four situations in which text authorship is disputed:

- A text attributed to one author seems non-homogeneous, lacking unity, which raises the suspicion that there may be more than one author. If the text was originally attributed to one author, one must establish which fragments, if any, do not belong to him, and who are their real authors.
- A text is anonymous. If the author of a text is unknown, then based on the location, time frame and cultural context, we can conjecture who the author may be and test this hypothesis.
- If based on certain circumstances, arising from literature history, the paternity is disputed between two possibilities, A and B, we have to decide if A is preferred to B, or the other way around.
- Based on literary history information, a text seems to be the result of the collaboration of two authors; an ulterior analysis should establish, for each of the two authors, their corresponding text fragments.

Authorship analysis deals with the classification of texts into classes, based on the stylistic choices of their authors. The problem of authorship identification is based on the assumption that there are stylistic features that help distinguish the real author from any other possibility. This set of stylistic features was recently defined as linguistic fingerprint (or stylome), which can be measured, is largely unconscious and is constant (van Halteren et al., 2005).

Beyond the author identification and author verification tasks, recently other tasks occurred, like author profiling, author diarization, gender and age prediction, or author masking (given a document, paraphrase it so that the original style

Proceedings of Recent Advances in Natural Language Processing, pages 201–207,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_028

does not match that of its original author, anymore.) The main conference dedicated to authorship problems (PAN) organizes yearly competitions dedicated to these tasks (Trinidad et al., 2006; Rocha et al., 2017).

In this paper, we are interested in a related topic, strongly related with the stylistic fingerprint of an author, namely if an author preserves his stylome in two different political periods of his life. To be more precise, we want to see if we can discriminate between the essays written by an author in the communist period, and the essays written by the same author in the post-communism period. As a study case, we have chosen Solomon Marcus (1925-2016), one of the most prominent scientist and man of culture of modern Romania. As a scientist, he published in an impressive range of different fields, as mathematics, computer science, mathematical and computational linguistics, semiotics, etc. Marcus published about 50 books in Romanian, English, French, German, Italian, Spanish, Russian, Greek, Hungarian, Czech, Serbo-Croatian, and about 400 research articles in specialized journals in almost all European countries. He is one of the initiators of mathematical linguistics (Marcus 1970) and of mathematical poetics (cf. Encyclopaedia Universalis (French), vol. 9, 1971, p. 1057-1059, and vol. 13, 1989, p. 837), and has been a member of the editorial board of tens of international scientific journals covering all his domains of interest. One of his most famous books is probably *Mathematical Poetics* (Marcus 1973), which pioneered the interdisciplinary field of poetics and mathematical linguistics. As a man of culture, he has written an equally impressive amount of texts, having as main topics mathematical education, culture, science, children, etc. His wide interests and complex personality have left deep imprints in Romanian scientific and cultural world. In this article we propose a stylistic analysis of his publicistics texts gathered in six volumes (Marcus, 2012-2017), aiming to uncover some of his quantitative and qualitative fingerprints, and we test if we can distinguish between the essays written in two distinct periods of time in his writing style: 22 years of communist regime and 27 years of democracy.

2 The Corpus

The whole collection of Marcus' publications is available on print, in six volumes entitled *"Răni deschise"* (Opened wounds), consisting of almost 6000 pages and 1.056.400 words and spanning over a period of a half of century. The first volume comprises 1256 pages of texts, conferences and interviews, from 2002 to 2011 (and a few published before 1990). The second volume *"Cultură sub dictatură"* (Culture under dictatorship), of 1088 pages, and the third „Depun Mărturie" (I testify), of 668 pages, are collections of texts published between 1967 and 1989. The fourth volume *"Dezmeticindu-ne"* (Awaking), of 1030 pages, contains texts from the period immediately following the Romanian Revolution (December 1989) to the year 2011. The fifth volume of the collection, entitled *"Focul și Oglinda"* (The fire and the mirror), contains 709 pages and represents texts written by Marcus in 2012. Finally, the sixth and last volume, *"Eu doar întreb"* (I just ask) contains 592 pages and is written in 2013. From its foreword by Mihai Dinu, it is apparent that the publisher house (Spandugino) is preparing a final volume, with Marcus' texts from 2014 to 2016.

The input texts where provided in pdf format. Extracting texts required manual intervention in every volume to remove noise generating areas like images, text under images, etc. For conversion, we have used pdf to text converter. The resulted text where further processed to ensure proper diacritics. Python with NLTK library was used to extract tokens, phrases and other necessary information. TreeTagger was used for POS extraction. Hyphens where extracted using Pyhpen (python module), but further refinements where necessary to ensure a proper list of hyphens. One CVS file was created for every processed volume. It contains all the raw information necessary to perform future analysis on the texts. Extracting articles written until 1989 and after was performed in a semi-automatic approach. Regular expressions where used to detect article boundaries and human input was requested for validation.

3 Stylistic Analysis

One of the first attempts to uncover the stylistic fingerprint of an author was (Mendenhall, 1901), who tried to establish the authorship of texts from Shakespeare. Mendenhall argues in favor of the fact that words distribution, by their length, represents an essential feature of the style of an author. Such measurements abounded later on, quantifying almost anything, from the length of syllables,

words and phrases, to the distribution of the part of speech. While this type of measurements has its merits, clues about the stylistic fingerprint of an author should be more about the unconscious text patterns of the author, which cannot be voluntarily controlled. From this category, stop words (or grammatical words, which do not have semantic content of their own, like pronouns, quantifiers or determiners, etc.) distribution is among the most popular features used to determine the stylome of an author (Mosteller and Wallace, 1964).

Rank	Freq.	Romanian Word	English Translation
1	2540	matematică	mathematics
2	1094	știință	science
3	941	problemă	problem
4	710	științific	scientific
5	695	teorie	theory
6	604	domeniu	field
7	594	fapt	fact
8	554	exemplu	example
9	538	cercetare	research
10	538	limbaj	language
11	534	activitate	activity
12	512	cultură	culture
13	512	matematician	mathematician
14	498	trebui	must
15	476	exista	exists
16	471	parte	part
17	450	vrea	want
18	433	privi	regards
19	392	rezultat	result
20	390	românesc	Romanian

Table 1: The most frequent 20 content words in communism

3.1 A Quantitative Analysis of Linguistic Objects

To analyze Marcus' style, we first gather quantitative evidence: the six volumes have all together 1.056.400 words, 50.596 sentences and 2.226.621 syllables. Thus, the texts yield an average of 20.8 words per sentence and of 2.1 syllables per word. Second, since the six volumes are chronologically ordered, we investigated the possibility that the texts differ in their style, before and after the fall of the communist regime. The two volumes written in the communist period (the second and the third volume) have a clear preference for longer constructions: the average of number of words per sentence is 23.54 for the second volume and 24.95 for the third. Expectedly, the average for the rest of the volumes, written in democracy is sensibly lower, of around 20 words per sentence (even 18.2 for the fifth volume). Also, the same tendency occurs for the length of words: the communist period reveals a mean of 2.23 syllables per word for the second volume and of 2.24 syllables per word for the third, whilst the post-communist texts present a lower average of 2.01 syllables per word. Thirdly, we had a look at the way Marcus made use of the parts of speech in his texts from *"Răni Deschise"*. Thus, the distribution of the main parts of speech (verbs, nouns and adjectives) in Marcus' texts is as follows: 157.008 verbs, 309.613 nouns and 97.935 adjectives in all of the six volumes. It follows that Marcus used in average almost 3 verbs per sentence, 6.11 nouns per sentence and a much lower number of adjectives, which have a mean of fewer than 2 per sentence (1.93). Did Marcus make different use of the three categories of parts of speech in communism and democracy? The verbs distribution is rather the same over the two periods: about 3 verbs per sentence. However, the nouns and the adjectives present significant fluctuations. Thus, the average of adjectives per sentence from the second and third volumes, written in the communist period, is 2.64 and 2.82, respectively and the average of nouns per sentence is 7.75 and 7.68, respectively. As one can see, all of these values are homogenous and over the mean of the whole texts in the six volumes by almost one unit. On the contrary, the means of the adjective and nouns per sentence from the volumes written after 1989 are visibly lower than both the general mean and the mean from the communist period, as it follows: the average of adjectives (and nouns) per sentence, for the volumes one, four, five and six are, in this order: 1.64 (5.59), 1.85 (5.96), 1.41 (4.91) and 1.67 (5.51). Clearly, Marcus' appetite for nouns and adjectives diminished after the fall of the communism. It is already transparent that there is an obvious difference between the way Marcus wrote before and after the fall of communism.

3.2 Functional Words

But the preference for long words and phrases, and for more nouns and adjectives during the communist regime, might well have been voluntary, ac-

Rank	Freq.	Romanian Word	English Translation
1	3679	matematică	mathematics
2	2341	lucru	thing
3	1937	avea	have
4	1911	trebui	must
5	1686	spune	say
6	1592	știință	science
7	1504	vedea	see
8	1493	profesor	professor
9	1425	fapt	fact
10	1412	lume	world
11	1379	școală	school
12	1258	problemă	problem
13	1251	elev	student
14	1224	exista	exist
15	1218	cultură	culture
16	1188	viață	life
17	1156	vrea	want
18	1106	parte	part
19	1103	om	human
20	1098	educație	education

Table 2: The most frequent 20 content words in democracy

cording to the need to conform to the regime's norms, as well as the choice of the content words or topics.

What about the involuntary fingerprint of the two distinct periods? Was Marcus' style affected in depth? To answer this question, we resort to one of the most commonly accepted indicators of personal style of an author, namely the distribution of grammatical (functional) words, such as pronouns, determiners, prepositions, etc. Such lists of stop words were used for authorship attribution of texts with controversial paternity. We use here the list of 120 functional Romanian words (Dinu et al., 2012).

The first 15 functional words used by Marcus (and their raw frequencies) in all of the six volumes are the following: *în* (in 31.160) , *un* (a 29.411), *avea* (have 29.261), *fi* (be 28.223), *și* (and 27.486), *al* (of 17.851), *care* (which 17.229), *el* (he 16.906), *la* (at 16.088), *să* (to 14.591), *nu* (no 13.397), *ca* (that 11.202), *cu* (with 10.405), *mai* (more 9.503), *pe* (on 9.442).

For the communist period the first functional words are in volume 2: *un* (6206), *în* (6027), *și* (5408), *fi* (4519), *al* (4504), *avea* (4426), *care* (2977), *el* (2906), *la* (2697), *cu* (2003), *mai* (1968), *nu* (1965), *să* (1834), *acest* (1553), *din* (from 1471); and in volume 3: *în* (3683), *un* (3610), *și* (3164), *fi* (2689), *al* (2542), *avea* (2304), *care* (1683), *el* (1578), *la* (1395), *mai* (1171), *cu* (1128), *nu* (1084), *să* (1024), *acest* (this 923), *că* (849).

One instantly notices that the second and the third volume, both written in the communist period, have almost identical stop word lists, with only two exceptions: in volume two the first two words are *un* (a) and *în* (in), while in volume three the order is reversed, but with a very small difference in frequency; also, the same thing happens with the words in positions 10 and 11, *cu* (with) and *mai* (more).

The post-communism volumes (1, 4, 5 and 6) are also very similar within the group with regard to the ranking of the stop words and they definitely differ in this respect from the other two volumes. We only give here the stop word lists for two of them, the rest being similar. The list of the first 15 stop words for the first volume is: *avea, fi, în, un, și, el, la, care, să, al, nu, că, pe, cu, mai; and for the fifth is: avea, în, fi, un, și, să, care, el, la, nu, al, că, pe, cu, acest*.

This natural grouping is confirmed by analyzing the similarity of the six rankings. The computation of the exact distance between any two of the stop words ranking revealed quite small values for the distance between the two volumes written before 1989, and a neat clustering in one group of the other volumes. To this end, it seems that even a scientist and a man of culture of Marcus' amplitude was subject to both voluntary and involuntary differences in writing style in communism versus democracy periods.

3.3 Clustering Experiments

We want to confirm the above global intuitions on randomly selected short texts. We have chosen 23 texts written before 1989 and an equal number of texts written between 2000 and 2013. We have extracted from each text a 120 functional words frequency list, ranked them by their frequency, and computed the distance between the obtained rankings. The distance we have chosen is rank distance (Dinu, 2003; Popescu and Dinu, 2008), an ordinal distance which was successfully used also in other problems of authorship (Dinu et al., 2008, 2012). One proper way to test the virtues of a dis-

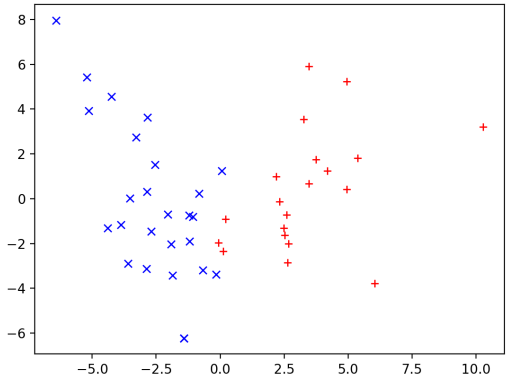

Figure 1: Clustering of random Marcus' texts from two distinct periods

tance measure is to use it as a base for a clustering algorithm (Duda et al., 2001). For our purposes, we applied a k-means clustering algorithm with k=2, corresponding to the two possible classes of texts: written before 1989 (in communism), and after 1989 (democracy). The results are plotted in figure 1, where by + we marked the texts written in communism period and by x the ones written in democracy.

One can clearly notice that out of the 46 texts, all 23 written after 1989 form a well separated group. For the other 23, 20 of them form a distinct cluster, while the remaining 3 marginally group with the texts written in democracy. Interestingly, those 3 texts were written immediately before the fall of the communism (one in 1988 and two in 1989). Could it be that professor Marcus understood it was the right time for the change?! This wouldn't be a surprise for those who knew him.

What is remarkable about this analysis is that the functional words were unconsciously used by the author in a different way in communism than in democracy period, and that we can distinguish the texts written in communism by an author on the base of such analysis.

4 Semantic Analysis

We also addressed some aspects of the distribution of the topics that Marcus tackled in his texts from *"Rani Deschise"*. To gain insight about the text content and his preferred topics, we employed the distribution of the content words. Since the texts are naturally divided in two periods (communism and democracy), an analysis of differences in semantic content and topics between the two periods is compelling.

Rank	Keyness Value	Romanian Word	English translation
1	1949	matematică	mathematics
2	1264	problemă	problem
3	1101	știință	science
4	1010	teorie	theory
5	980	științific	scientific
6	839	privi	look
7	780	vrea	want
8	746	exista	exist
9	701	diferit	different
10	694	domeniu	domain
11	666	carte	book
12	630	asa-zis	so called
13	624	trebui	must
14	619	limbaj	language
15	611	românesc	Romanian
16	541	fapt	fact
17	531	numara	count
18	518	cercetare	research
19	508	matematician	mathematician
20	501	situatie	situation

Table 3: Keywords from *"Răni Deschise"* in communism

We have seen that, in what the length of the words and phrases and the distribution of parts of speech are concerned, the two periods clearly differ from one another, the traces of the omnipresent plethoric style of the communist period being present in detectable quantities in Marcus' text.

We performed a statistical analysis on Marcus' texts, using AntConc, a concordance tool made by Laurence Anthony, available as open source at www.laurenceanthony.net.

In a pre-processing step, we first obtained a word frequency list for each of the two periods. Because the words we are interested in are content words (words that have a semantic content of their own), the stop words were being excluded from the counting. Also, we lemmatized the corpus and we collapsed all words which have the same lemma into the same word.

Since Romanian is a highly inflectional language, this had the effect of dramatically reducing the word types. Thus, a word family such as *matematică, matematici, matematica, matemati-*

Rank	Keyness Value	Romanian Word	English translation
1	2215	lucru	thing
2	1887	trebui	must
3	1828	avea	have
4	1686	spune	say
5	1401	profesor	teacher
6	1349	scoală	school
7	1348	matematică	mathematics
8	1312	asa-zis	so called
9	1305	vrea	want
10	1297	vedea	see
11	1282	exista	exist
12	1280	elev	student
13	1257	educatie	education
14	1180	lume	world
15	1160	stii	know
16	1155	viată	life
17	1150	copil	child
18	1089	tv. romana	romanian tv
19	1022	afla	find out
20	996	intelege	understand

Table 4: Keywords from *"Răni Deschise"* in democracy

cile, matematicii, matematicilor, etc. (mathematics singular, mathematics plural, the mathematics singular, the mathematics plural, mathematics singular genitive, mathematics plural genitive, etc.) was conflated into a single word, *matematică*.

In table 1 and table 2 we give the first 20 most frequent words from both periods (the complete list of words is at http://nlp.unibuc.ro/resources/sm.pdf). One can see that Marcus' favorite topics in the texts from the communist period revolve around mathematics, research, science and culture, while the texts written in democracy target more areas of education (teacher, school, education, etc.).

Such frequency content word lists are quite transparent w.r.t Marcus's favorite topics (mathematics, language, teaching, education, etc.). A more in depth analysis is needed to detect differences between the topics addressed during the two periods. Thus, in a preliminarry analyses we give here a comparative report of the word usage in the texts from the two periods. We generated a keyword list, based on the keyness values for each word. A keyword list is a list of words which are unusually frequent in the texts, as compared with a reference corpus. We have generated a keyword list for both possible cases: one for the communist period, with the democracy texts as reference corpus and one for the democracy period, with the communist period texts as reference corpus. This time, we did not exclude the stop words, since the difference in use of both stop and content words from the two distinct writing periods might prove to be of interest. The key word generation method we have used is log likelihood. The keywords list was sorted by the keyness value, which is an indication of a preference for using a word in the corpus, as compared to the reference corpus. A statistically significant value of keyness, taken from a table of statistical values is 3.9 for a 5use of a word, as compared to the use of that word in a reference corpus. The bigger the value, the more significant the preference is. The keyness values of the keywords from the communist period texts, considering the democracy texts as reference corpus, start from a very high value of 1949 and for the democracy text, considering the communism texts as reference corpus, from 2215. This is a clear indication of an important semantic difference between the content of the texts from the two periods. In table 3 and 4, we give some of the most relevant words that have significant keyness values for both cases. Interestingly enough, he only speaks about the communism during the democracy period.

5 Conclusions

The distributional analysis of Marcus' text uncovers that the passing from the communist regime period to democracy is sharply marked by two complementary changes in Marcus' texts.

In the pre-democracy period, the communist norms of writing style demanded on the one hand long phrases, long words and clichés, and on the other hand a short list of preferred 'official' topics. On the contrary, in democracy, Marcus shortened the phrases and words (naturally becoming more concise) and approached a broader area of topics.

The clustering approach based on the preference of the author regarding the functional words shows that the functional words were unconsciously used by the author in a different way in communism than in democracy period, and they can be used to discriminate between the communist and post-communist texts of a given author.

In future works we plan to investigate more Ro-

manian authors and to test the above hypothesis also for other languages.

6 Acknowledgements

Research supported by HerCoRe project, funded by Volkswagen Foundation (Project no. 91970).

References

Liviu P. Dinu. 2003. On the classification and aggregation of hierarchies with ifferent constitutive elements. *Fundamenta Informaticae 55.1* pages 39–50.

Liviu P. Dinu, Vlad Niculae, and Octavia-Maria Şulea. 2012. Pastiche detection based on stopword rankings: exposing impersonators of a romanian writer. *Proc. of the Workshop on Computational Approaches to Deception Detection. Association for Computational Linguistics* pages 72–77.

Liviu P Dinu, Marius Popescu, and Anca Dinu. 2008. Authorship identification of romanian texts with controversial paternity. *LREC 2008* pages 3392–3397.

R. O. Duda, P. E. Hart, and D. G. Stork. 2001. Wiley-Interscience Publication.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the Association for Information Science and Technology* 60(1):9–26.

Kim Luyckx and Walter Daelemans. 2008. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '08, pages 513–520. http://dl.acm.org/citation.cfm?id=1599081.1599146.

Solomon Marcus. 1973. *Mathematische Poetik*. Ed. Academiei, București-Athenaum Verlag, Frankfurt am Main.

Solomon Marcus. 1989. *Inventie si descoperire*. Ed. Cartea Romaneasca.

Solomon Marcus. 2012-2017. *Rani deschise (6 volumes)*. Ed. Spandugino.

Solomon Marcus, Ed. Nicolau, and S. Stati. 1970. *Introduzione alla linguistica matematica*. Casa editrice Riccardo Patron.

TC Mendenhall. 1901. *A mechanical solution of a literary problem*.

Frederick Mosteller and David Wallace. 1964. *Inference and disputed authorship: The Federalist*. Addison-Wesley, Massachusetts.

Marius Popescu and Liviu P. Dinu. 2008. Rank distance as a stylistic similarity. *COLING 2008* pages 91–94.

Anderson Rocha, Walter J. Scheirer, Christopher W. Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne Carvalho, and Efstathios Stamatatos. 2017. Authorship attribution for social media forensics. *IEEE Trans. Information Forensics and Security* 12(1):5–33. https://doi.org/10.1109/TIFS.2016.2603960.

José Francisco Martínez Trinidad, Jesús Ariel Carrasco-Ochoa, and Josef Kittler, editors. 2006. *Progress in Pattern Recognition, Image Analysis and Applications, 11th Iberoamerican Congress in Pattern Recognition, CIARP 2006, Cancun, Mexico, November 14-17, 2006, Proceedings*, volume 4225 of *Lecture Notes in Computer Science*. Springer. https://doi.org/10.1007/11892755.

Hans van Halteren, R. Harald Baayen, Fiona J. Tweedie, Marco Haverkort, and Anneke Neijt. 2005. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics* pages 65—77.