ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume **21**, Number 2, 2018, 99–113

Stylometric and Topic Analysis of a Historical Text

A Computerized Study of General Averescu's 'War Memoirs' (1916–1918)

Horia Nicolai L. TEODORESCU[1, 2] and Speranta Cecilia BOLEA[3]

[1] *Gheorghe Asachi* Technical University of Iaşi, ETTI, Iaşi, Romania
[2]Romanian Academy, Iaşi Branch, Iaşi, Romania
[2]Institute of Computer Science, Romanian Academy, Iasi Branch, Iaşi, Romania
E-mail: hteodor@etti.tuiasi.ro,
cecilia.bolea@iit.academiaromana-is.ro

Abstract. The study applies stylometric tools to analyze the relationship between topic and style in a WWI work, General Averescu's *War Memoirs*. New stylometric indices are suggested and applied.

Key-words: autobiographic work, statistics, computational linguistics, vocabulary, POS statistics, text color index, personalization index, literary style, stylometry.

1. Introduction

Stylometric, sentiment, and topic analyses have been applied mostly if not only in a disjoint manner, and mostly to contemporary texts. We posit that these methods, especially when used together, may help historians in the identification of the social contexts and general sentiments at various epochs based on writings describing events of the time; in addition, these methods of analysis may help discerning the degree of objectivity of those texts. Also, we suggest that stylometric and topic analysis should be combined in order to improve the segmentation of the texts into meaningful sections for further analyses.

We have proposed in [1] the application of the polarity and sentiment analysis to the *War Memoir*s of General Averescu ("Notie Zilnice din Rzboiu", written in Romanian) [2] with the aim of improving the understanding of the attitude of the writer and, through his eyes, of the Romanian military during WWI. This paper aims to show how stylometric analysis can help determining the degree of objectivity of historical texts and their splitting in sections related to major historical periods and facts and then we applied the analysis to the same text [2]. This

article also complements with technical details the article [1]; moreover, it relates to and has some parallel ideas with [3].

The computer-based analysis of the historical texts related to WWI is scarce. Berindei [4] discusses the speeches and declarations at the time of WWI, but without doing a language analysis. In [5–8] the authors analyze the speeches, declarations and memoirs from WWI, regarding the nationalistic sentiment and spirit from the central and Eastern Europe.

In this study, we apply stylometric, topic, and personalization analysis technics to study the 100 year-old text [2]. We are especially interested in vocabulary aspects, including unusual words that occur with higher frequencies in these texts, and in the discourse analysis in general. The analysis is first directed toward the vocabulary and basic sentence analysis: the number of words per phrase, the structure of the phrases, and the frequency of various *parts of speech* (POS) – expressly of the pronouns. Several stylometric indices are applied and a few new ones are introduced for refining the analysis. The topic-, color-, personalization-, and sentiment-study show, among others, that the book [2] is more an autobiography than a chronicle. On the other hand, topic analysis, especially the change in the military vocabulary used, the average sentence length, and partly the frequency of pronouns discriminate between major sections of the text. The color index, Zipf's law, hapax legomena, Sichel's and Honoré's measures, color indices, and the negativity index are found almost impractical in the discrimination. Because the topic and personalization relate to the witnessed war events and to the author's participation in them, historians may derive that there is a significant chance that the first two parts of the book correspond to two different periods of the war.

The stylometric analysis is a well-established method in forensic analysis [9–13]. Zu Eissen and Stein [10] have proposed "to operationalize plagiarism detection by dividing a document into "natural" parts, which may be sentences, paragraphs, or sections, and analyzing the variance of certain style features." We suggest, among others, a reversed method for use in historic and linguistic studies, namely the use of stylometric analysis to determine sections of writings that are related to specific epochs and events. This method may also be useful in the search of "temporal tracks" and temporal patterns that [14] among others advocate as a tool "for establishing the real chronological order" of the narrated events.

The organization of the paper is as follows. Section 2 describes the pre-processing method. Section 3 presents the results of the statistical analysis. The final section discusses the results and concludes the study. Throughout the paper, the translations of the Romanian words into English are given into square brackets and the Romanian words in the text are given in italics. The sign '+' after *a, s, t* marks diacritics (ă, ş, ţ).

2. Pre-processing method

In order to detail the exploration, we empirically divided this work into three sections (suggested by us, based on an empirical topic assessment) [1]: the first section, denoted by S1, starts from 21st August and ends on 4th Dec. 1916, pages 13–112, and regards the Romanian's army defense and the loss of the capital, Bucharest. The second Section (S2) lasts from 6th Dec. 1916 to 22nd Nov. 1917, pages 112–253; it regards the defensive war in 1917, and the fights that have been given to preserve the independence of the Romanian territories which were still in the possession of the Romanian government. The third section (S3) corresponds to the period that lasts from 26th Nov. 1917 to 5th Mar. 1918; this part extends on pages 253–311, and refers to political events after the military operations ceased as a consequence of the revolution in Russia.

Averescu's *War Memoirs* [2] was saved in a DOC file and manually corrected; namely,

- The preface, the annexes (written in italics in the original), the tables and the images with their explanations were removed; The missing diacritics were introduced;
- The short sentences written in French were replaced with their Romanian translation;
- Several words have been re-written for compatibility with the current orthography and with the lemmatizer; for example *inimic* is today *inamic*. A partial list of the old and new orthography for words is given in [15] (Annex 2 of [15]), for exemplification;
- We kept the words that changed their meaning or are infrequent today [16].

After making these amendments, the file was saved in TXT format; then it was parsed to lemma and part of speech, using the TTL parser developed by RACAI [17]. The results from the first parsing were not so good. Some abbreviations are not correctly recognized, while the period "." used in the format of the dates (see the examples given in the Annex) has been considered a punctuation sign by the lemmatizer. For solving these problems, a list of abbreviations was created [16]; we substituted the period from all abbreviations with "*x*" (for example, "*Arm.*" was modified in "*Armx*"), and then we solved them as abbreviations. (Only a few neologisms, such as 'complex' and 'duplex' end in 'x' in Romanian; therefore, it was easy to sort out the abbreviations coded with an x at the end.) In situations like ".VIII." we deleted the periods altogether. Next, we parsed the text again. The result, better than the first one, was input in a Visual C program; the output (TXT file) contains all the lemmas and their count.

3. Statistical analysis

3.1. Basic statistics

The first two sections of the volume, denoted by S1 and S2, which refer to the period of the military operations of the Romanian army, are studied here. The vocabulary analysis produced a number of 2,583 lemmas and 20,266 words for S1, and 3,203 lemmas, 25,924 words for S2 (we also took into account compound words). In S1, there are 805 paragraphs and 1,274 sentences, with an average of 15.88 words per sentence, while S2 has 1,157 paragraphs, and 1,773 sentences, with an average of 14.63 words/sentence (where the sentence is delimited by two successive full stops). While the difference of the averages for S1 and S2 is less than 10%, it is significant, with 99.999%, confidence interval of difference –2.1112 $<$ –1.3 $<$ –0.4888 (Wald), t-difference: –7.079, *df-t*= 3044.5; *p*= 0. The third part has 2,243 lemmas, 15,305 words, and 1,049 (composed) sentences; the average sentence length is 14.59 words.

Zipf's distributions for the two parts are shown in Figs. 1 and 2. Both graphs (Zipf's) have slope –1.1 (with a small difference of 0.07) and the vertical axis intersection at about 8.5 (8.475 and 8.67), with $R^2 > 0.97$. Zipf's distributions ($y = -1.112x + 8.47$ for S1, double logarithmic scale, $y = -1.105x + 8.67$ for S2) do not differentiate between the two parts, but the slope is slightly larger than for most of the texts in Romanian, see [18], and larger than the one for large texts in English ([19] provides values of 0.96, 0.89, and 0.82 for three texts in English, see Fig. 4 in that paper).

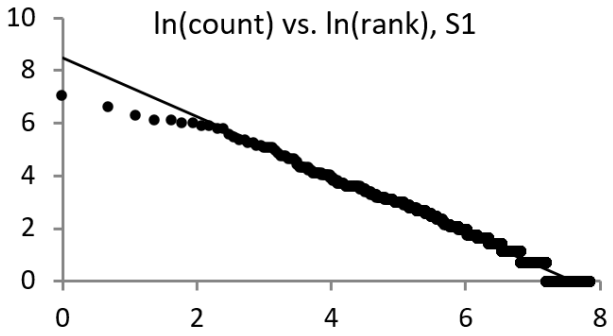

Fig. 1. Rank distribution of the words in the first Section of the War Memoirs.

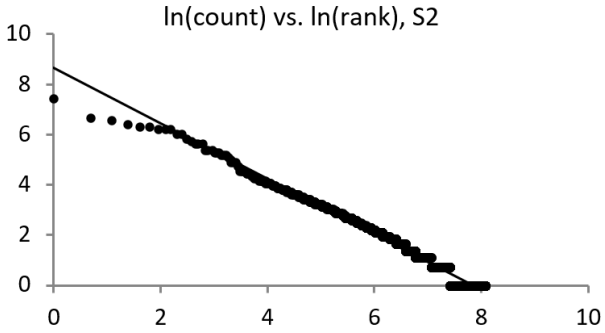

Fig. 2. Rank distribution of the words in the second Section of the *War Memoirs*.

3.2. Vocabulary, topic analysis and the frequency of military words

The sum of the numbers of lemmas in the two parts, 4,186, and the ratios of lemmas to words of 0.127 for S1 and 0.126 for S2 are indicative of a restrained vocabulary, which is more specific to technical writings than to literature [1]. This view is also supported by the large frequencies of several words related to the military profession among the first 100 words. Excluding prepositions, conjunctions and other stop words (so called 'function words', but not including pronouns and auxiliary verbs), the first 100 content words [2] and retained function words in S1 are { *eu, el, general, armata+, vrea, divizie, meu, putea, situat+ie, face, spune, mare, ordin, tot, cartier, foarte, zi, trupa+, inimic, retragere, commandant*} *[I, he, general, army, to will, division, my, being able, situation, do, say, large, order, all, quarter, very, day, troop, enemy, retreat, commander]*. Two topics emerge: the relationship (*I, my he, his*) and the military topic. The same is true for S2, but not in the third section of the text, where the military topic almost vanishes. A short list of military terms and of their count in the first two sections of the text is telling:

[1] This assertion should be considered with care: First, the vocabulary increases with the number of words, with a tendency of saturation for large numbers of words. The ratio of lemmas to words is a more appropriate indicator in this respect. Second, some major literature writers have used a rather low vocabulary compared with others.

[2] Words are usually classified in 'function words' and 'content words', the last category carrying meanings. But we fully agree that "actually there are not any crisp boundary lines between", e.g. function words (synsemantics) and content words (autosemantics). See [20].

- *"general" [general]*, meaning a rank in the army, or with the adjectival meaning, as in *cartier general*, [headquarter'] appears 211 times in S1 and 201 times in S2;
- *"armat" [army]*, with the meaning: unit of the armed forces, occurs 168 times in S1 and 212 times in S2;
- *"divizie" [division]* , unit of the armed forces, occurs 152 times in S1 and 129 times in S2;
- *"inimic" [enemy]* appears 63 times in S1 and 36 times in S2.

While in most studies the aim is to find texts with the same topic, we need a different approach here, because we search for parts of the same text that deal with different topics. Therefore, we need a fine grain study of the vocabulary changes along the text.

The military words occupying the first three positions in S1 also occupy the first three positions in S2, namely *general, armata+, divizie*. The list of the military terms among the first words in S1 and their frequencies are given in Table A1 in the Annex, along with the frequencies in S2. The correlation of the series represented by the absolute frequencies in Table A1 is large (0.64) and the regression line of the ranks has a coefficient of determination $R^2 = 0.41$. The correlation increases when using the relative frequencies (percentage) to $C = 0.876$ and the linear regression between the frequencies is $y = 0.792x - 0.239, R^2 = 0.767$. However, this good correlation is largely due to the first three terms. The χ^2 test shows that the two distributions are not similar, see the Annex. The correlation for the last ten tokens in the Table A1 is, for absolute frequencies, 0.282 and the respective regression line has $R^2 = 0.0797$. Even this correlation value is deceptive; a mixture of terms that occur among the most frequent 100 words in S1 or in S2, but not in the top 5 positions, see Table 1, proves to have a negligible correlation (correlation coefficient –0.010; slope of the regression line –0.005, coefficient of determination $R^2 = 0.0001$); for the last ten terms in Table A1, $\chi^2 = 84.30621$, df = 9, p-value $< 10^{-5}$. Therefore, the military events recounted have required different uses of some of the military terms, whose relative frequency may change up to 5-fold, but for others the use remains almost equally frequent in the two parts. The change in the frequency of some terms, but not all, is another reason justifying the division of the volume in three parts, based on topic considerations.

If we focus on the frequent words depicting activities, a different picture emerges. The 'activity words' frequent in S1 (see Table A1) are *retragere, ofensiva+, operat+ie*, and *retrage*. The positive one, *ofensiva+*, occurs 37 times, while the negative ones, *retrage, retragere*, occur 92 times; this might justify for S1 the title "The Retreat". In S2, the 'activity words' frequently occurring are *atac* (45 times) and *ofensiva+* (31 times, not among the first 100). The negative ones, *retrage* and *retragere* occur together 49 times. Interestingly, the word *dus+man* (meaning *foe* and having a much more personal connotation than *enemy*) occurs only 5 times in S1, but 49 times in S2, while *inimic [enemy]* occurs 63 times in S1 and 36 times in S2, indicating a much more personal involvement of the author in the actions in S2. All these topic- and polarity-related elements support the notion of different parts of the memoirs, for S1 and S2.

Also, expectedly, several words related to the war conditions may help recover the image of the events. For example, the fact that *aliat [ally/ies]*, appears once in S1 and 12 times in S2 indicates Averescu's perception that the allies have played a much lesser role in the first part of the war in Romania, which is historically true. The similarity of S2 and S3 (the final part) is negligible, χ^2= 30.566, df = 10, p-value = 0.0007, confirming the distinction between the three parts.

Concluding, while both S1 and S2 have a military topic, there are notable differences in the frequencies of the main military terms, especially those related to military activities. This points toward different military sub-topics in S1 and S2.

Table 1. LExamples of military terms in the two parts of Averescu's *War Memoirs*

Term	S1	S2	S3	% S1	% S2	% S3
trupa+ [troop]	64	69	28	0.32	0.27	0.18
retragere	62	25	10	0.31	0.10	0.07
comandant	55	40	10	0.27	0.15	0.07
cartier	71	42	2	0.35	0.16	0.01
colonel	39	53	18	0.19	0.20	0.12
dus+man	6	49	16	0.03	0.19	0.10
corp	34	48	5	0.17	0.19	0.03
atac	19	45	5	0.09	0.17	0.03
regiment	41	44	4	0.20	0.17	0.03
ofit+er	25	42	13	0.12	0.16	0.08
act+iune	15	40	4	0.07	0.15	0.03
			Sum %	2.13	1.92	0.75

3.3. The frequencies of the parts of speech

We only consider here nouns, verbs, adjectives, and adverbs (see Table 2). Stop words, such as prepositions and conjunctions, are neglected. The statistics of the POS is as follows. In the section S1 of the memoirs there are 4,935 nouns, 4,496 main and auxiliary verbs, 1,034 adjectives, 1,557 adverbs, 1,609 pronouns and 171 abbreviations, while in S2 there are 6,236 nouns, 5,453 verbs, 1,324 adjectives; 1,771 adverbs, 1,866 pronouns, and 144 abbreviations, see Table 2. (We also determined the list of abbreviation [16], where these are recognized by the parser as nouns; there are 50 of these in S1 and 14 in S2). Table 2 seems to show that the two statistics are different; however, when weighted by the number of words in the two parts, the statistics is almost the same. Similar text features, based on the frequencies of parts of speech, were used by other authors, e.g. [13] in fraud detection [3].

Table 2. The count of parts of speech

Parts of speech	No. of occurrences			No. of occurrences weighted by the number of words	
	S1	S2	S3	S1 %	S2 %
Nouns*	4,936*	6,546**	3,687	0.244	0.253
Main verbs	3,049	3,745	2,566	0.150	0.144
Auxiliary verbs	1,447	1,950	1,246	0.071	0.075
Adjectives	1,034	1,383	688	0.051	0.053
Adverbs	1,557	1,855	963	0.077	0.072
Pronouns	1,609	1,956	1,286	0.079	0.075
Abbreviations**	160	144	44	0.008	0.006

* excluding the 50 abbreviations initially provided by the parser.
** excluding the 14 abbreviations initially provided by the parser.

[3] During the revision of the paper after receiving the referees comments, we found that Pearl and Steyvers also use as text features ratios of the frequencies of parts of speech, as well as ratios of pronouns frequencies, but in a way different from ours.

There is an almost perfect correlation between the use of verbs, nouns, adjectives and adverbs in the two parts (without abbreviations as in Table 2), with the correlation value 0.997. The regression line further validates this matching, with the coefficient of determination of R = 0.994. The *chi square* test (excluding abbreviations) provides χ^2= 16.44, df = 5, p-value = 0.0057. However, this result may be due to the sensitivity of the chi test to large numbers. With the artifice of counting the tens of words (by removing the last figure in each number), one obtains χ^2=1.632, df = 5, p-value = 0.897. The ratio of the relative frequencies of the POS to the number of the nouns, see Table 4, shows that the statistical similarity of the sections S1 and S2 is preserved. This matching indicates that Averescu's use of POS was not affected by the topic of the first two sections of the *War Memoirs*. For the second and the third parts, S2 and S3, the result is less clear-cut, χ^2= 75.61, df = 5, p-value $<$ 10-5; after removing the last figure (counting the tens of words), χ^2=7.595, df = 5, p-value = 0.1799. This points to some dissimilarity between S2 and S3.

However, if only the main verbs and adverbs are considered, for S1 and S2, χ^2= 0.532, df=1, p-value = 0.466, showing that there is doubt that the two populations are identical. We checked the effect of the scale on this result, asking the question if the distributions of "tens of units" (verbs, adverbs) are similar; using 305, 375, 156, 185 instead of 3049, 3745, 1557, 1855, we obtained χ^2= 0.07344, df=1, p-value = 0.786, which preserves the conclusion.

Table 3. The count of parts of speech divided by the number of nouns

Parts of speech	No. of occurrences relative to the number of nouns	
	S1	S2
Nouns	1.000	1.000
Main verbs	0.618	0.572
Auxiliary verbs	0.293	0.298
Adjectives	0.209	0.211
Adverbs	0.315	0.283
Pronouns	0.326	0.299
Abbreviations	0.032	0.022

3.4. How personal, self-reflecting are the War Memoirs? Personalization analysis

There are a few features worth noticing in the Memoirs, namely the heavy use of the "I" and "me" and the complete lack of the pronouns "they", "their", "you", and "yours". The abundance of 'I' and 'me' could be expected for a work titled Memoirs; yet, this is not without interest, because this abundance shows that the *War Memoirs* are largely a self-biographic work rather than an objective recount of events that the author witnessed. This is somewhat troubling because it indicates a very subjective description of the war. The lack of "you" and "they" is even more troubling, showing that only personal relations are emphasized, with no neutral description to groups or populations. As a matter of fact,

- "*eu*" [*I*] is used 402 times in S1 and 481 times in S2 where "*eu*" refers to the author, general Averescu (these are his memoirs); the relative frequencies are respectively 1.98% and 1.85%;

- "*el*" [*he*] occurs 239 times in S1 and 291 times in S2, ("*he*" refers to various commanders of the army, the King, or important political leaders);

Further details are shown in Table 4.

Table 4. Frequencies of the Pronouns

Pronouns	S1	S2	S1 relative	S2 relative	In [12]**
eu [I]	402	479	1.98	1.85	-
meu [my]	125	198	0.62	0.76	-
mie [to me]	1	1	0.005	0.004	-
noi [us]	0	0	0	0	0,11
noua+ [to us]	0	1	0	0.004	-
nostru [ours]	0	0	0	0	-
el [he]	239	291	1.18	1.12	0,34
lui [his]	54	61	0.27	0.235	0.49
sa+u [his, variants]	24	19	0.12	0.07	-
voi* [you]	0	0	0	0	-
voua [you], [to you]	0	0	0	0	-
vostru [yours]	0	0	0	0	-
ei [they]	0	0	0	0	0,29
lor [their]	0	0	0	0	

* In Romanian, 'voi' stands for 'you' as well as for 'will' in the future tense. All 'voi' occurrences are instances of the verb: there are 11 uses of 'voi' in S1 and 21 occurrences in S2, all verbs.
** Collection ≠CMG of 93 volumes in Romanian, mixt corpus, 8,806,433 words [18]. Only for the first 100 words the frequencies are provided in [18].

The count of the pronouns in [2] is extremely different compared with other works. Pronoun analysis reveals that much of the volume is a first-person assertion of the authors viewpoint on the war and a monologue. Therefore, the volume is as much a documentary work and the recount of a witness and actor as an apologia, a self-praise work. While historical dates and war events are probably fully correct, the interpretations of the military actions and especially political decisions of the time are probably quite subjective and biased. For the Table 4, using only the data for the *eu, meu, el, lui, sa+u* (that occur more than once), one obtains χ^2= 7.428, df = 4, p-value = 0.115, which points to a chance that the use of pronouns is different in the two parts. However, the correlation between the counts of pronouns expressed by percentages of the total number of words in the two parts S1 and S2 is very high, 0.99, and the linear regression between them has the coefficient of determination R^2 = 0.98. The only visible difference in the percentage of pronouns is for the *he+his* (el+lui+(sa+u)), with a total percentage of 1.57% in S1 and 1.425% in S2, a relative difference larger than $10\%(100 \times \frac{1.57-1.425}{1.425} \approx 13\%)$.

We suggest that an index of personalization, defined as $I_p = \frac{(numberof\text{“}I\text{”}+numberof\text{“}my\text{”})}{(numberof\text{“}he\text{”}+\text{“}she\text{”}+\text{“}his\text{”}+\text{“}her\text{”})}$ could be a good synthetic measure of the degree of subjectivity of the text. This index has the value 1.8 for S1 and 1.92 for S2. The value of this index is much lower than 1 when determined for the collection ≠CMG of 93 volumes in Romanian reported in [18], which indicates that the War Memoirs represent a very personal writing.

We conclude that the use of pronouns is consistent in S1 and S2, with the possible exceptions for *meu* and *sa+u*, which have palpable differences. In addition, the employment of the pronouns is very different in this text compared with other works and points toward a very subjective text.

3.5. The negativity index

Table 5 illustrates the use in *War Memoirs* of several negating or restricting words such as *nu* (equivalent of the determiner/adverb *no*, or adverb *not*), *niciodata+* [never], *niciun* [none], *fa+ra+ [without], deloc [not at all]*, and *nimic* [nothing].

Table 5. Frequency of negation and restriction words

Word, Romanian	Word, English	S1	S2	S3	S1 %	S2 %	S3 %
deloc	not at all	6	5	5	0.03	0.02	0.03
fa+ra+	without	21	33	25	0.10	0.13	0.16
nici	neither	24	35	16	0.12	0.14	0.10
nici_un	neither one	20	31	17	0.10	0.12	0.11
nimic	nimic	13	21	5	0.06	0.08	0.03
nu	no, not	269	349	240	1.33	1.35	1.57
numai	only	33	37	20	0.16	0.14	0.13

The parts S1 and S2 have a significant chance to be different with respect of the occurrence of the negative words, S1-S2: (χ^2= 2.273, df = 6, p-value = 0.893. The same is true for the comparison of the parts S1 and S3 (χ^2= 6.006, df = 6, p-value = 0.422) and S2-S3 (χ^2= 7.809, df = 6, p-value = 0.252). The correlation of the series for S1 and S2 is 0.999 and the linear regression S2(S1) has R^2 = 0.999. Similar results are obtained for the couple S2 and S3. Therefore, the use of the negative words in Table 5 does not differ for S1 and S2 or for S2 and S3.

The negation index is defined by the number of negation words n_{neg}such as 'no' (nu), 'without' (*fără*), 'none' (*nimeni, niciunul*) etc., divided by the total number of words, n_w,

$$I_{neq} = n_{neq}/n_w.$$

Surprisingly, the negation index of the first part (expressed here for convenience in %), with $I_{neg} = 1.98\%$, is slightly lower than for the second part, $I_{neg} = 2.03\%$, where Averescu recounts his own victories as a commander. The third part is even more negative, with $I_{neg} = 2.25\%$. Compared to other works reported in [18], having the frequency of no equal with 1.33 % in the corpus CLG of literary works and 1.28 in the corpus of general texts in [18], the frequency of no is almost the same in the two parts of the discussed text (1.33% in S1, 1.35% in S2).

3.6. Newer indices for style analysis and text section finding - Adjectival and adverbial coloring

0.5cmWe suggest that the style is characterization could be improved using a set of features suitable for statistical analysis, namely (i) length of sentences and phrases; (ii) "verbal color index", represented by the average number of adverbs and adverbial constructions vs. number of verbs and verbal phrases, (iii) "nominal color index", represented by the number of adjective and attributes vs. the number of nouns and nominal constructions, (iv) the negation index, and (v) the index of personalization. These features are intuitively connected with the degree of subjectivity and affective self- involvement of the writer. The negativity index may also have connections with the mood and feelings of the writer. Such features may be telling to the historians, possibly indicating an undesired entanglement of the writer interests or feelings in the description of the facts.

The verbal color index is defined as

$$I_{vc} = n_{adv}/n_{verb}.$$

where n_{adv} is the number of adverbs and adverbial constructions and n_{verb} is the number of verbal constructions. Similarly the nominal color index is defined as

$$I_{nc} = n_{adj} n_{noun}.$$

where n_{adj} is the number of adjectival constructions and n_{noun} is the number of nominal constructions. The total color index I_c is then $I_c = (I_{vc} + I_{nc})/2$.

3.7. Basic stylometric measures

Sichel's stylometry measure (see [21]), which is a vocabulary novelty measure applied to lemmas, is defined as

$$S = V(2, N)/V(N)$$

where V(2,N) is the number of types (lemmas) that occur twice in the text (*dis-legomena*) and V(N) is the total number of types (lemmas) in the text of N tokens. For the first part, $V_{S1}(2, N) = 413, V_{S1}(N) = 2,583$ (here, including numerals and abbreviations.) For the second part, $V_{S1}(2, N) = 502, V_{S2}(N) = 3,263$. The values of Sichel measure are shown in Table. 6. For ease, both V(1,N) and V(2,N) have been approximated by the number of tokens that appear once, respectively twice; we verified that the approximation is good enough for V(2,N). Notice that tokens with different meanings but with the same root (or etymology from the same word), such as *t+inere, t+inut, t+inuta+*, or *Zaharia* and *Zaharow* have been considered as different types.

Similarly, for describing the low frequency end of the tokens distribution, Honoré's measure is used [21]; it is defined based on the hapax legomena (the terms occurring only once in the text) as

$$H_o = \frac{100 \log N}{1 - V(1, N)/V(N)}.$$

The values of Honoré's measure H_o are also shown in Table 6. The values of the color indices in the two parts are given in Table 7.

Table 6. The frequency of Hapax legomena and stylometric measure values

	Hapax legomena % (H)	Sichel (Si)	Honoré (Ho)
S1	0.482	0.160	658.2
S2	0.492	0.154	691.9

Table 7. The values of the color indices

Index	S1	S2	Whole text
I_{ve} Verbal color index*	0.346	0.326	0.335
I_{ne}, Nominal color index	0.209	0.210	0.210
I_e, Total color index	0.278	0.268	0.275

* using both main and auxiliary verb counts.

The proportion of dis-legomena in The *War Memoirs* is consistent with the Sichel statistic prediction, which puts the percentage of dis-legomena at about 15% for a wide range of texts, from about 1,000 to $4 \cdot 10^5$ words [22]. The proportion of hapax legomena (the relative frequencies) in the two parts is quite high compared with typical texts; [23] indicates for texts between $2 \cdot 10^5$ and 10^7 words values between 0.4359 and 0.4472. However, this should be expected for small texts with total word numbers much less than 10^5[23]. Therefore, the difference in the relative frequency hapax legomena between S1 and S2 is not indicative of a difference in style, because the number of words is substantially larger in S2; moreover, the frequency is higher in S2, as expected [23].

Using the features as elements of the style feature vector, one obtains the feature vector of the text, $F = (I_{vc}, I_{nc}, I_c, H, Si, Ho)$. One may add the personalization index and the negation index to the vector, but we do not use them here. Then, two texts or parts of texts have different styles when at least one of the components of their feature vectors significantly differ statistically.

The feature vectors F1 and F2 of the text sections S1 and S2 are:

F1=(0.346,0.209,0.278,0.482,0.160,658.2), respectively

F2=(0.326,0.210,0.268,0.492,0.154,691.9).

The correlation of the first five elements of the vector elements (except Honor) is 0.997. The regression line is $y = 1.027x - 0.0131, R^2 = 0.99$. Therefore, from this stylometric point of view, the sections S1 and S2 are quasi-identical. Including the index of personalization and the index of negativity in the feature vector does not change this conclusion.

Noticeable, the number of abbreviations is high for typical memoirs; this feature typically designates a rather technical work. In fact, the large number of abbreviations points to the conclusion that the *Memoirs* (mainly the first two parts) were written in the "technical" style of a military professional.

3.8. Variation of the stylometric measures with respect to the boundaries of the sections

The division of the text into three major sections has been made, as already specified, by the authors, based on the subjective assessment of the historic events narrated. It is therefore useful to check if changes in the partition produce significant variations of the values of the stylometric measures. The principle we put forward is that an optimal (stylometric) partition produces the largest distances between the parts; in addition, we assume that the optimal stylometric partition corresponds to the historic partition.

We repeated the analysis with the first part reduced by the last 10 pages and the second part increased with the 10 pages removed from the initial first part. Them we repeated the procedure by increasing the first part by 10 pages (the first 10 pages of the second part), and decreasing correspondingly the second part. Then, we computed for the three versions of the partition, (S1, S2, S3), (S1+10, S2-10, S3), and (S1-10, S2+10, S3), the Euclidean distance between the corresponding stylometric vectors. The vectors used in this case comprised (i) the frequencies of the pronouns; (ii) the verbal color index, the nominal color index, and the total color index, and (iii) the relative frequencies of the parts of speech (normalized by the total number of words). For the color index vectors, the results are as in Table 8, with the lowest distances obtained for the partition (S1+10, S2-10, S3). However, for the frequencies of the pronouns and for the the relative frequencies of the parts of speech, the largest distances are obtained for the partition (S1-10, S2+10, S3), see Tables 9 and 10.

Table 8. Squared distances for three partitions, for the vectors of color indices

d(S1-S2)	0.001	d(S2-S3)	0.0078	S1+10 p
d(S1-S2)	0.0005	d(S2-S3)	**0.0093**	S1-10 p
d(S1-S2)	0.0005	d(S2-S3)	0.0093	S1 & S2

Table 9. Squared distances for three partitions, for the vectors of the absolute frequencies of the pronouns

d(S1-S2)	1961	d(S2-S3)	16271	S1+10 p
d(S1-S2)	**40193**	d(S2-S3)	**4611**8	S1-10 p
d(S1-S2)	14189	d(S2-S3)	30911	S1 & S2

Table 10. Squared distances for three partitions, for the vectors of he relative frequencies of the parts of speech

d(S1-S2)	0.000132	d(S2-S3)	0.000986	S1+10 p
d(S1-S2)	**0.000833**	d(S2-S3)	**0.001187**	S1-10 p
d(S1-S2)	0.000182	d(S2-S3)	0.000991	S1 & S2

These results show that there is not a perfect equivalence between the partitioning based on stylometric measures and the historical events (as subjectively assessed, in this case); yet, the partitioning based on stylometric measures reasonably coincides with the sections derived based on historic considerations. This is, however, a preliminary conclusion.

4. Discussion and conclusions

The approach in this study has been to determine various lexical and stylistic features of the three parts of the *Memoirs* that correspond to the three phases of the war described by Averescu; the first aim was to statistically describe the three parts and to determine what features are statistically relevant in the discrimination of the three parts. The second, more ambitious aim was to show by an example that the statistics of the features might be useful as a tool for automatically differentiating major military and political events and settings in writings. In the analyzed case, the results support our hypothesis that the applied method may be useful in historical studies; however, much more extensive studies should be conducted on large corpora of texts of historical interest in order to prove the validity of the method.

A substantial difficulty encountered was the selection of the pertinent features of the text. We completely agree with [24] that "*Function words* [stop words], *type/tokens, word lengths, hapax legomena, and other specific style markers may not in themselves be an indicator of a unique style*"; we found that some of these statistical indicators may correlate poorly with the topic and polarity of the texts. Yet, other indicators, such as the frequencies of specific POS, mainly the specific uses of pronouns, changes in the length of the sentences, and the variation in the frequency of the professional terms may help meaningfully characterizing parts of a text in relation with the specific topic dealt with, and possibly with the sentiments carried by those parts. Among the new stylometric measures we suggested, the color indices and the negativity index have not been useful in distinguishing the first two sections of the text; yet, these indices differ more for the third part; moreover, and they may be worthwhile in other analyses.

The study avoided the use of features based on n-grams; although methods based on n-grams have been proved to be highly efficient, they may have low intuitive appeal for linguists and historians. In contrast, the use of specific POS, the use of rare words (hapax legomena), and the richness of the vocabulary, and features derived from these may be easier to connect to other knowledge pieces the human scientists use.

Averescu's *War Memoirs*, a text of significant historical interest, proved to be separable into parts based on the events recounted by the author, *i.e.*, based largely on topics; in addition, the separability is marginally supported by a few of the stylometric measures discussed, especially for the last (third) part from the others. In many respects, the third part is so different from the previous ones that one might believe it is written by another author, or at a different stylistic period of the author. However, we can not elaborate on the last issue; in fact, that would be a different line of research, requiring new tools, as no statistical studies on discriminating between writings of the same authors at different epochs of their life have been published at our best knowledge.

Acknowledgements. The authors are grateful to the anonymous referees for their very constructive remarks and suggestions.

Authors' contributions. Conceived and designed the analysis: HNT. Processed the primary data (parsing, POS, vocabulary): CSB. Further processed and analyzed the data: HNT. Wrote the paper: HNT. Corrected and agreed with the paper: CSB, HNT.

References

[1] TEODORESCU H.N., BOLEA C., *Polarity and sentiment analysis of General Averescu's War Memoirs*, http://iit.academiaromana-is.ro/centenar/paperPSAen.html.

[2] AVERESCU A., *Notiţe Zilnice Din Războiu (1916-1918)*, Editura "Cultura Naională Bucureşti", 1935.

[3] GOGALNICEANU P., TEODORESCU H.N., *Polarity and Sentiment Analysis of Medical Transplant Literature*. Proc. COMM-2018, 12th Int. Conference on Communications COMM2018, 2018, Bucharest, România.

[4] BERINDEI D., *Independenţa în contextul istoriei României*, Academica, **5-6**, 2017, Anul XXVII, pp. 33–35, 2017.

[5] IORDACHI C., *The Unyielding Boundaries of Citizenship: The Emancipation of 'Non-Citizens' in Romania, 18661918*, European Review of History: Revue européenne d'histoire, **8** (2) pp. 157–186, 2001.

[6] CUSCO A., GROM O., SOLOMON Fl., *Discourses of Empire and Nation in Early Twentieth-Century Bessarabia: RussianRomanian Symbolic Competition and the 1912 Anniversary*, Ab Imperio. Theory and History of Nationalism and Empire in the Post-Soviet Space, **4**, pp. 91–129, 2015.

[7] TURDA M., *Conservative Palingenesis and Cultural Modernism, Early Twentieth-century Romania*, Totalitarian Movements and Political Religions, **9**(4) pp. 437–453, 2008.

[8] ERSOY A., GÓRNY M., KECHRIOTIS V., *Modernism: Representations of National Culture*, Discourses of Collective Identity in Central and Southeast Europe 17701945: Texts and Commentaries, **3** (2) Central European University Press, 2010.

[9] ZU EISSEN S.M. and STEIN B., Intrinsic Plagiarism Detection, in (Eds. Lalmas, Mounia et al.), Advances in Information Retrieval, 2006, Springer Berlin Heidelberg, pp. 565–569.

[10] OREBAUGH A. and ALLNUTT J., Classification of Instant Messaging Communications for Forensics Analysis, The International Journal of FORENSIC COMPUTER SCIENCE, IJoFCS (2009) 1, 22–28.

[11] IQBAL F., BINSALLEEH H., FUNG B.C.M., DEBBABI M., Mining writeprints from anonymous e-mails for forensic investigation. Digital Investigation, **7**, (1-2), October 2010, pp. 56–64, https://doi.org/10.1016/j.diin.2010.03.003.

[12] HARPALANI M., M. HART, S. SINGH, R. JOHNSON, Y. CHOI, Language of Vandalism: Improving Wikipedia Vandalism Detection via Stylometric Analysis. Proceeding HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - **2**, pp. 83–88, Portland, Oregon, 2011.

[13] PEARL L., STEYVERS M., Detecting authorship deception: a supervised machine learning approach using author writeprints. Literary and Linguistic Computing, **27**(2), 2012, pp. 183-196.

[14] MACOVEI A., An Introduction to Time Tracks, Proceedings of the 12^{TH} International Conference *Linguistic Resources and Tools for Processing the Romanian Language* Mlini, 27-29 Oct 2016, Eds. M. Mitrofan et al., ISSN 1843-911X, pp. 19–27.

[15] Supporting material and annexes to [1], http://iit.academiaromana-is.ro/centenar/paper/Support%20material%20and%20anexes_Polarity&Sentiment%20Analysis.pdf.

[16] TEODORESCU H.N., BOLEA C., *Tiny dictionary of words and expressions in Romanian used during the WWI period Alexandru Averescu "Notie Zilnice din Rzboiu"* (1916-1918), http://iit.academiaromana-is.ro/centenar/dict_en.html.

[17] TTL Parser – Racai http://www.racai.ro/en/tools/text/.

[18] VLAD A., MITREA A., *Contribuii privind structura statistic de cuvinte n limba romn scris*, Dan Tufis, Florin Gh. Filip (coordonatori), Limba Română în Societatea Informatională - Societatea Cunoasterii. Editura Expert, Bucuresti, România, pp. 209–236, 2002. available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.145.3285&rep=rep1&type=pdf

[19] MORENO-SÁNCHEZ I., FONT-CLOS F., CORRAL A., *Large-Scale Analysis of Zipf's Law in English Texts*, 2016. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0147073.

[20] POPESCU I.I., ALTMANN G., KÖHLER R., *Zipf's LawAnother View*, Quality & Quantity, **44**(4), pp. 713–731, 2010.

[21] BAAYEN H., Van HALTEREN H., TWEEDIE F., *Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution*, Literary and Linguistic Computing, **11**(3), pp. 121–132, 1996.

[22] GANI J., *Characterizing an Author's Vocabulary*, South African Statistical Journal, **31**(1), pp. 1–11, 1997.

[23] HOLMES D. I., *A Stylometric Analysis of Mormon Scripture and Related Texts*, Journal of the Royal Statistical Society. Series A (Statistics in Society), **155**(1), pp. 91–120, 1992.

[24] RUDMAN J., *The State of Authorship Attribution Studies: Some Problems and Solutions*, Computers and the Humanities, **31**, pp. 351–365, 1998.

Annex

Remark 1. The 26 archaic words listed in [16 0], are found in both sections (S1-S2) of the *War Memoirs* 157 (97+60) times.

Remark 2. Two examples of issues produced by the punctuation signs are shown below

For *.VIII.*, the result from the parser is

<seg . . . ><s> . . .
<c>.</c>
</s></seg>
<seg lang="ro"><s id="id_temp_aiurea.2">
<w lemma="VIII" ana="Mc" chunk="Np≠1">**VIII**</w>
<c>.</c>
</s></seg>

"*Arm.*", the result from the parser is:

<w lemma="arm" ana="Ncms-n" chunk= "N≠3">**Arm**</w>
<c>.</c>

Table 11. Military terms among the most frequent 100 tokens in S1 and their frequency in S2.

Military term	S1	S2
general	211	201
armata+	168	212
divizie	152	129
situat+ie	105	56
ordin	77	59
cartier	71	42
trupa+	64	69
inimic	63	*36
retragere	62	*25
comandant	55	40
front	42	110
regiment	41	110
grup	40	*5
colonel	39	53
ofensiva+	37	*31
rezerva+	36	*21
corp	34	48
operat+ie	34	*9
retrage	30	*24

*not in the first 100 in S2

For the data in this Table, χ^2= 131.599, df = 18, p-value $<$ 10-5; after excluding 'grup', which has only 5 occurrences in S2, χ^2= 107.59, df = 17, p-value $<$ 10-5; after excluding the first three terms, χ^2= 95.416, df = 14, p-value $<$ 10-5. Therefore, the two sets differ almost sure.