THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 21, Number 3/2020, pp. 283–292

APPLYING A DELTA-TYPE MEASURE TO TEXT COHERENCE ANALYSIS AND TEXT SEGMENTATION

Speranta Cecilia BOLEA[1], Horia-Nicolai TEODORESCU [1,2]

[1] Institute of Computer Science of the Romanian Academy, Iași, Romania
[2] "Gheorghe Asachi" Technical University of Iasi, Iași, Romania

Corresponding author: Horia Nicolai TEODORESCU, E-mail: hteodor@etti.tuiasi.ro

Abstract. Burrow's Delta measure is based on the assumption that the most frequent words differentiate in a corpus between authors represented by sub-corpuses. The assumption was empirically confirmed on various corpuses and several languages. We extend the assumption to parts of a single work, by replacing the corpus with the respective work and the sub-corpuses by segments of the work. Highest similarity between successive parts of the work is assumed to indicate stylistic coherence between those parts. In this way, we introduce a variant of Delta measure that we name coherence Delta measure and that can serve in text segmentation based on stylistic features. We expose an algorithm and apply the method to two works. The results support the applicability of the method for text segmentation.

***Key words*:** stylometry, Burrow's Delta measure, text segmentation, coherence Delta measure.

1. INTRODUCTION

The principal aim of this study is to apply stylometric tools to the segmentation of literary works, especially memoirs and self-biographies, according to criteria related to chronologic events and attitude of the writer toward those events, where the writer's attitude is reflected in the style of the sections of the texts. For this purpose, we introduce a new stylometric measure rooted in the Burrows' Delta measure (see [1-3]), which we modify to correspond to the goal of text segmentation based on stylistic coherence.

Stylometric analysis has many settings and several applications, including authorship analysis [1-3], online identity verification [4], correcting authorship by de-attribution [5], detecting the gender of the authors [6], finding fake news [7] and threats in cyber space [8] with potential positive and negative (censuring) implication in policing the cyber space, analysis of legal corpuses [9], warding off and combating spam messages [10], analysis of the scientific literature [11], plagiarism detection [12], and even determining the age of unknown writers [13]. Stylometric tools were also used in narrowly specialized applications, including combating censorship by stylistic anonymization [14], finding effects of aphasias on the style of conversational speech [15], and analysis of chronology of the works [16] – the latter topic being the closest to the one dealt in this study. Several of these studies suggest the use of Burrows' measure as a suitable measure for the applications related with authorship [1-3].

While the use of Burrows' stylometric measure in authorship attribution is based on the implicit assumption of stylistic coherence of the entire work of an author, we suggest that the stylistic coherence of a text is imperfect and may be applied in segmenting that text. We suppose that the stylistic traits are related to the attitude of the author with respect of the storyline and so, *ceteris paribus*, the style changes along the tale depending what the author narrates. Thus, we question the stylistic coherence of the texts and use the change in stylistic traits to detect its main parts.

The second section of the paper reviews Burrows' distance and its limits and then presents the new stylometric measure that adapts Burrows' one to the inner stylistic coherence of a text. The algorithm for applying the suggested measure is described in detail. The third part presents results of applying the inner coherence measure for the segmentation task of a text of memoirs. The last section discusses the potential

extension of the method and its limits. This study is in line with and continues [17–19], but departs from them in the method developed, which has stronger foundations and better suited to automatic analysis.

2. THE ALGORITHM

2.1. Burrows' Delta measure and its limitations

The so-called Delta measures are typically used for authorship attribution by comparing one text with a sub-corpus, and for comparing two texts between them [20]. We extend the use of these distances by comparing segments of a given text with the purpose of automatic segmentation of the text in its most "natural" parts, where "natural" means higher cohesions of the parts at the vocabulary level. In fact, the proposed method substantially differs from that used by Burrows in determining authorship, but the reasoning that is founding both methods is the same.

Burrows' Delta stylometric distance is based on the assumptions that, in a corpus composed of several sub-corpuses, each with single and known authorship, each sub-corpus has a unique statistics of the most frequent words, see [20], where the respective statistics is unique to the corresponding author in general, not only for the works in the sub-corpus. Then, the statistics derived from the sub-corpuses can be used as a pattern to identify other works by authors included in the corpus simply by the "distance-to-prototype" method applied in various fields: a work with unknown author is attributed to the sub-corpus to which it is closest. Some details are worth noting. First, there is a new, hidden assumption that the work to attribute was written by one of the authors with works in the corpus. Next, the prototype method works well only when the classes (here, set of works) are well separated, which is a strong condition. In addition, it is known that the choice of the distance function may be essential to find a good separation between classes. There is also an apparently technical detail that is also a strong hypothesis, namely that there is a (relatively small) set of words, which happens to be the most frequent in the entire corpus, that can be used to represent all sub-corpuses. That is, instead of considering the most frequent n words in each sub-corpus, one uses the set of n most frequent words over all works. Obviously, by doing so one introduces the relativity of the n words and of their relative frequencies: whenever a work or a sub-corpus is added, the set and the frequencies changes.

Despite these strong assumptions and limitations of the method, the Delta distance proved to be a strong and useful tool in authorship attribution; for definition and properties of Burrows' distance see [20]. We acknowledge that at least some of the limits of the Delta measure also are limits of the c-Delta measure defined in the subsection 2.3 and of the related method described in this article.

2.2. Description of the basic text segmentation method

The method as described in this subsection is applicable to texts that are already sectioned in parts or chapters by the author, or for texts that have a proposed sectioning by the critics or by a human expert, as in [18, 19]. However, the method is expandable to the case when a preliminary sectioning is not available, by dividing the works in 3 to 5 equal parts and taking the initial boundary parts described in 2.3 of a tenth of each part; in this case, a random sectioning can be performed and then iteratively corrected. In case of texts already sectioned, the purpose of applying the method is to determine if the sectioning choice due to the author is reasonable from the stylistic coherence point of view, or if it can be automatically improved.

Since the hypothesis underlying the use of the proposed method is that the text is not stylistically uniform (coherent), the preliminary phase of the method consists in checking that a perceptible variation in style occurs between its parts. Because the suggested method, similarly to Burrows' one, relies on lexical features, the rank statistics of the words in the different parts is determined and compared. Assume that the text is divided in ν parts by the author. Also assume that w_h is the h^{th} most frequent word (actually, lemma) *in the entire text* and $p_{h,j}$ is the relative frequency of w_h in the j^{th} part. Then, each part j of the text, $j = 1,\ldots,\nu$, as segmented by the author, is characterized by a vector of n probabilities $V_j = \left(p_{1,j}, p_{2,j}, \cdots, p_{h,j}, \cdots, p_{n,j} \right)$. In this way, one creates a "vector space model" for each part of the text. The stylistic coherence of the work is defined in terms of similarity of segments of texts, as used by [21] for comparing claims of a patent. We follow Burrows' method of computing distances between the vectors V_j.

2.3. Description of the applied method and algorithm

Various versions of the algorithm are conceivable to fit the above general discussion. We explore here one of the simplest variants. For performing text segments comparisons, we propose that the role of the corpus, as used in the literature when applying Burrows' Delta, is taken by the entire text to be segmented for the coherence Delta (c-Delta) measure. This implies that the analyzed text is ample enough (at least 100 pages). Also, we assume that the number of sections is known or pre-determined and larger than two. For each initially chosen section, two sub-sections (chunks of text) are chosen, one at the beginning and one at the end of the sections. These subsections may correspond to actual subchapters or chapters determined by the author, or may be a determined number of pages or paragraphs. A sketch of the method, in algorithmic form, is given below.

Inputs:

(large) Text of N_w words (lemmas) on N_p pages

$\nu > 2$ number of sections

Initial sections of $N_{10}, N_{20}, \ldots, N_{\nu 0}$ pages, as proposed by the author, by a linguist, or randomly chosen; $N_{10} + N_{20} + \ldots + N_{\nu 0} = N_p$. (Instead of pages, one may use as unit the paragraph, or sentences, or even words.)

Δ_p choice of the increment of pages (as we used in case of Averescu's text [18, 19]), or

$\Delta_{1,2}, \Delta_{2,1}, \Delta_{2,3}, \ldots, \Delta_{\nu-1,\nu-2}, \Delta_{\nu-1,\nu}, \Delta_{\nu,\nu-1}$ choice of the ending and starting subsections in the initial sections that might better belong stylistically to the next ($\Delta_{j-1,j}$), or to the previous sections ($\Delta_{j,j-1}$) – according to the method we used in case of Iorga's text, see also [19]. See Fig. 1 for a graphical representation of the parts and sub-sections used in the algorithm. The notation $\Delta_{h,h+1}$ means the ending sub-section of the initial section *h*, which might better align at lexical level with the next section; similarly, $\Delta_{h,h-1}$ denotes the starting sub-section of the section h that might better match at the vocabulary level the previous $(h-1)$ section.

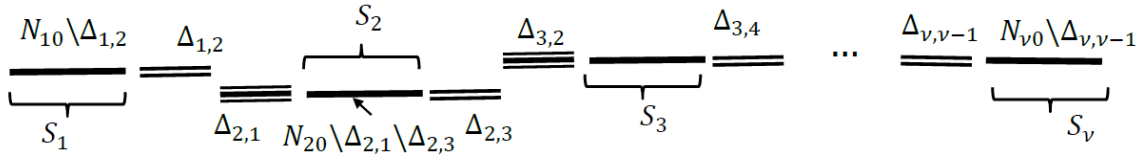

Fig. 1 – Representation of the sections and subsections.

Step 1: For the entire text of N_w words, compute the absolute frequencies of all the words w_j over the entire text, f_{Aw_j}, and the relative frequencies, $f_{rw_j} = f_{Aw_j} / N_w$. Select the first q most frequent words, $w_1, \ldots, w_q$, for example $q = 30 \ldots 50$. Because this is not a problem of authorship, we believe that the choice of q is less important; because the problem is the segmentation of a work, we err toward larger values of q for obtaining stronger distinction between parts.

Step 2. Define the "main sections" as the sections $S_1 = N_{10} \setminus \Delta_{1,2}$ (that is, the initial first section of N_{10} pages less than the ending section $\Delta_{1,2}$), $S_2 = \left(N_{20} \setminus \Delta_{2,1}\right) \setminus \Delta_{2,3}$, …, $S_h = \left(N_{h0} \setminus \Delta_{h,h+1}\right) \setminus \Delta_{h,h-1} \ldots$, $S_\nu = N_{\nu 0} \setminus \Delta_{\nu,\nu-1}$.

Step 3. For all the "main" sections S_h, determine the relative frequencies of the most frequent words determined at Step 1 in each section S_h; denote these frequencies by $f_{j,h}$.

Step 4. Determine the average relative frequencies of the q words, $\overline{f_j} = \overline{f_{j,h}} = \overline{f_{rw_j}^h}$ taking the average over all values $h = 1, \ldots, \nu$, and the standard deviations $\sigma_j = \mathrm{STDEV}_h\left(f_{rw_j}^h\right)$, summing over h.

Step 5. Define the vectors of the "main" sections S_h as $\mathbf{u_h} = \left(u_{1h},\ldots,u_{jh},\ldots,u_{qh}\right)$ by

$$u_{jh} = \frac{f_{j,h} - \overline{f_j}}{\sigma_j} .$$

Step 6. Determine the relative frequencies of the most frequent words (as determined at Step 1) in each sub-section $\Delta_{h,h-1}$ and $\Delta_{h,h+1}$; denote them as $f_{j,h,h-1}$ and $f_{j,h,h+1}$.

Step 7. Define the vectors of the subsections $\Delta_{h,h-1}$ and $\Delta_{h,h+1}$ as $\mathbf{v_{h,h-1}} = \left(v_{1h,h-1},\ldots,v_{jh,h-1},\ldots,v_{qh,h-1}\right)$ by $v_{jh,h-1} = \frac{f_{j,h,h-1} - \overline{f_j}}{\sigma_j}$ and similarly for $\mathbf{v_{h,h+1}}$.

Step 8. Compute the distances (Euclidean distance in this paper) between the subsection $\Delta_{1,2}$ and the sections S_1 and S_2. Denote these distances by $\delta_{1,2\leftrightarrow 1}$ and $\delta_{1,2\leftrightarrow 2}$. Compute the distances between the subsection $\Delta_{2,1}$ and the sections S_1 and S_2. Denote these distances as $\delta_{2,1\leftrightarrow 1}$ and $\delta_{2,1\leftrightarrow 2}$.

Step 9. [N.B.: initial step exemplified in detail; can be replaced using h instead of 1 and $h+1$ instead of 2] Assign $\Delta_{1,2}$ and $\Delta_{2,1}$ as follows

If $\delta_{1,2\leftrightarrow 1} \le \delta_{1,2\leftrightarrow 2}$, then $S_1 \leftarrow S_1 \cup \Delta_{1,2}$,

else $S_2 \leftarrow \Delta_{1,2} \cup S_2$.

If $\delta_{2,1\leftrightarrow 1} \le \delta_{2,1\leftrightarrow 2}$, then $S_1 \leftarrow S_1 \cup \Delta_{2,1}$,

else $S_2 \leftarrow \Delta_{2,1} \cup S_2$.

...

[General case:] Repeat for $h < \nu$:

Compute the distances $\delta_{h,h+1\leftrightarrow h}$ between $\Delta_{h,h+1}$ and S_h, $\delta_{h+1,h\leftrightarrow h}$ between $\Delta_{h+1,h}$ and S_h and so on, for $h < \nu$. Then,

If $\delta_{h,h+1\leftrightarrow h} \le \delta_{h,h+1\leftrightarrow h+1}$, then $S_h \leftarrow S_h \cup \Delta_{h,h+1}$,

else $S_{h+1} \leftarrow \Delta_{h,h+1} \cup S_{h+1}$.

If $\delta_{h+1,h\leftrightarrow h} \le \delta_{h+1,h\leftrightarrow h+1}$, then $S_h \leftarrow \Delta_{h+1,h} \cup S_h$,

else $S_{h+1} \leftarrow \Delta_{h+1,h} \cup S_{h+1}$.

Step 10. Output the (new) sections obtained, $S_1,\ldots,S_\nu$.

Variations of the algorithm can be imagined; for example, distances $\delta_{h+1,h\leftrightarrow h}$ are computed only after determining if $\Delta_{h,h+1}$ remains with S_h, and $\delta_{h+1,h\leftrightarrow h}$ is defined as the distance between $\Delta_{h+1,h}$ and $S_h \cup \Delta_{h,h+1}$ etc.

3. RESULTS

We applied the described c-Delta method on two self-biographic volumes in Romanian, both published between World War I (WWI) and WWII. The first text, titled *Supt trei regi* ("Under Three Kings"; second edition, published 1932), is a historical narrative and political analysis based on the personal experience of the author, Nicolae Iorga, who was a famous European historian and Romanian politician and also served as a Prime Minister of Romania [22, 23]. The volume is divided in three parts (named "books") by the author, each section comprising several chapters (Roman numbering is used for the chapters), and each chapter having subchapters (in Arab numbers). In addition, the first section ("Book") has two parts. The total number of pages is 464.

We manually pre-processed the text by correcting older versions of words and older lexical forms or writing rules, e.g. "Franciei" was replaced by "Franței" (of France), "supt" by "sub", "s'o' by "să o", "s'a"

by “s-a”, “reunia” by “reunea”, “găsia” by “găsea”, “refus” by “refuz” etc., and removed the footnotes. Then, the text was divided according to the sections proposed by the author (initially sections), as following:

$N_{10} = text(Book\ 1)$, 124 pages split into two parts, each with six sub-chapters;

$N_{20} = text(Book\ 2)$, eight sub-chapters – 169 pages in total;

$N_{30} = text(Book\ 3)$, 167 pages split into six sub-chapters,

and then lemmatized using RACAI TTL parser [24]. Finally, the results of the TTL parser were saved in XML format.

Words written in other languages: French, German, English, have been parsed correctly only for the proper nouns. In all the other cases, we replaced the value of the *"ana"* attribute with *"X"* and then we deleted the *chunk* attribute, see Fig. 2. After these corrections, we applied the algorithm described in 2.3 as:

```
<w lemma="un" ana="Timsr" chunk="Np#14">un</w>
<w lemma="certain" ana="Ncms-n" chunk="Np#14">certain</w>
<w lemma="Iorga" ana="Np" chunk="Np#14">Iorga</w>
```

a)

```
<w lemma="un" ana="X">un</w>
<w lemma="certain" ana="X">certain</w>
<w lemma="Iorga" ana="Np" chunk="Np#14">Iorga</w>
```

b)

Fig. 2 – Example of parsing result for words written in other languages:
a) from TTL parser, b) after our modifications.

We chose sub-sections $S_{12} = \text{text}(\text{last 37 pages of } N_{10})$, $S_{21} = \text{text}(\text{first 35 pages of } N_{20})$; $S_{23} = \text{text}(\text{last 36 pages of } N_{20})$; $S_{32} = \text{text}(\text{first 34 pages of } N_{30})$; and the main sections $S_1 = N_{10} \setminus S_{12}$, $S_2 = N_{20} \setminus S_{21} \setminus S_{23}$ and $S_3 = N_{30} \setminus S_{32}$.

The XML files have been processed as follows: we extracted the lemmas and the number of their occurrences in descending order, then we computed the relative frequencies, selected the first 50 most frequent lemmas, and afterwards we applied the algorithm.

The main results of applying the method to Iorga’s volume (N. Iorga, *Supt trei regi. Istorie a unei lupte pentru un ideal moral si national*, 2nd ed., București, 1932) are summarized in Table 1; the results show that the main sections of the work (the “books”) as sectioned by the author have a good stylistic cohesion, except the beginning of last Books, which better connects with the second Book.

Table 1

The Euclidean distances between the sub-sections in Iorga’s text; method based on c-Delta measure

Distances	**Values**	**Conclusion(s)**
$d(S_{12}, S_1)$	14.983	$d(S_{12}, S_1) < d(S_{12}, S_2)$ thus S_{12} remains with S_1; $S_1 = (N_{10} \setminus S_{12}) \cup S_{12} = N_{10}$
$d(S_{12}, S_2)$	16.266	
$d(S_{21}, S_1)$	21.768	$d(S_{21}, S_1) > d(S_{21}, S_2)$ thus S_{21} remains with S_2; $S_2 = (N_{20} \setminus S_{21} \setminus S_{23}) \cup S_{21} = N_{20} \setminus S_{23}$
$d(S_{21}, S_2)$	19.584	
$d(S_{23}, S_2)$	12.473	$d(S_{23}, S_2) < d(S_{23}, S_3)$ thus S_{23} remains with S_2; $S_2 = (N_{20} \setminus S_{23}) \cup S_{23} = N_{20}$
$d(S_{23}, S_3)$	12.684	
$d(S_{32}, S_2)$	16.247	$d(S_{32}, S_2) < d(S_{32}, S_3)$ thus S_{32} should be adjoined with S_2; $S_2 = N_{20} \cup S_{32}$ and $S_3 = N_{30} \setminus S_{32}$.
$d(S_{32}, S_3)$	19.535	

On the other hand, some starting or ending sub-parts have a large distance to the proximal main parts. Remarkably, S_{21} is distant from both S_1 and S_2, S_{32} is quite distant from its parent part, S_3, while S_{23} is very close with both its adjacent parts (S_2 and S_3), indicating a very smooth stylistic transition from the second to

the third part. The resulted sections in Iorga's text are $S_1 = N_{10}$; $S_2 = N_{20} \cup S_{32}$; $S_3 = N_{30} \setminus S_{32}$. Based on the values in Table 1 and the above remarks, we are encouraged to say that the third Book should have started after the section S_{32} for more coherence at the stylistic level, as assessed by Delta distances between sub-sections.

The algorithm was also tested on the volume "War Memories, 1916-1918" (394 pages) by Marshal Alexandru Averescu (A. Averescu, *Notițe Zilnice din Războiu (1916-1918)*, Cultura Națională, București, 1935). Because the author has not split the volume into chapters, we divided it in three main sections (books), (N_{10}, N_{20}, N_{30}) [18, 19]. We removed the footnotes, tables, figures and their explanations, the preface and the annexes from the text. Because the author has not provided chapters or sub-sections of the book, sub-sections S_{12}, S_{21}, S_{23} and S_{32} were chosen to comprise twenty pages at the beginning and the end of the three parts we established. The results of applying the algorithm are given in Table 2 and show that, from the point of view of the c-Delta stylistic coherence, the first section is similar with the first book, the second one is stylistically twenty pages (S_{23}) smaller than the second book, and the last section stylistically should start with S_{23} and continue with the third book.

Table 2

The Euclidean distances between the sub-sections in Averescu's text; method based on c-Delta measure

Distances	**Values**	**Conclusions**
$d(S_{12}, S_1)$	12.105	$d(S_{12}, S_1) < d(S_{12}, S_2)$ thus S_{12} remains with S_1; $S_1 = \left(N_{10} \setminus S_{12}\right) \cup S_{12} = N_{10}$
$d(S_{12}, S_2)$	13.188	
$d(S_{21}, S_1)$	13.442	$d(S_{21}, S_1) > d(S_{21}, S_2)$ thus S_{21} remains with S_2; $S_2 = \left(N_{20} \setminus S_{21} \setminus S_{23}\right) \cup S_{21} = N_{20} \setminus S_{23}$
$d(S_{21}, S_2)$	12.362	
$d(S_{23}, S_2)$	14.233	$d(S_{23}, S_3) < d(S_{23}, S_2)$ thus S_{23} remains with S_3; $S_3 = S_{23} \cup \left(N_{30} \setminus S_{32}\right)$
$d(S_{23}, S_3)$	13.981	
$d(S_{32}, S_2)$	11.808	$d(S_{32}, S_2) > d(S_{32}, S_3)$, thus S_{32} remains with S_3; $S_3 = \left[S_{23} \cup \left(N_{30} \setminus S_{32}\right)\right] \cup S_{32} = S_{23} \cup N_{30}$
$d(S_{32}, S_3)$	10.779	

The sections in Averescu's text are therefore, according to the c-Delta criterion, $S_1 = N_{10}$, $S_2 = N_{20} \setminus S_{23}$, and $S_3 = S_{23} \cup N_{30}$.

Table 3

The Euclidean distances in the feature space of the parts of speech of sub-sections in Iorga's text and Averescu's text

Distances	**Values,** Iorga	**Values,** Averescu	**Conclusion(s)**, Iorga	**Conclusion(s)**, Averescu
$d_{POS}(S_{12}, S_1)$	4,194.448	2,971.734	$d_{POS}(S_{12}, S_1) < d_{POS}(S_{12}, S_2)$, thus S_{12} remains with S_1	$d_{POS}(S_{12}, S_1) < d_{POS}(S_{12}, S_2)$ thus S_{12} remains with S_1
$d_{POS}(S_{12}, S_2)$	5,510.559	3,762.207		
$d_{POS}(S_{21}, S_1)$	4,978.224	3,423.675	$d_{POS}(S_{21}, S_1) < d_{POS}(S_{21}, S_2)$, thus S_{21} remains with S_1	$d_{POS}(S_{21}, S_1) < d_{POS}(S_{21}, S_2)$ thus S_{21} remains with S_1
$d_{POS}(S_{21}, S_2)$	6,242.750	4,212.662		
$d_{POS}(S_{23}, S_2)$	5,592.755	3,952.637	$d_{POS}(S_{23}, S_2) < d_{POS}(S_{23}, S_3)$, thus S_{23} remains with S_2	$d_{POS}(S_{23}, S_2) > d_{POS}(S_{23}, S_3)$ thus S_{23} remains with S_3
$d_{POS}(S_{23}, S_3)$	9,408.720	1,810.492		
$d_{POS}(S_{32}, S_2)$	5,921.468	3,909.026	$d_{POS}(S_{32}, S_2) < d_{POS}(S_{32}, S_3)$, thus S_{32} should be adjoined with S_2	$d_{POS}(S_{32}, S_2) > d_{POS}(S_{32}, S_3)$ thus S_{32} should be adjoined with S_3
$d_{POS}(S_{32}, S_3)$	9,737.648	1,750.320		

For comparison of the results obtained with the c-Delta method with other stylistic descriptions, as in [18], the results of the TTL parser have been analyzed and then we extracted the total number of all parts of speech (POS). In the space determined by the POS features [18], we applied a similar algorithm (see 2.3). When calculating the distances (Table 3), we were interested in the following parts of speech: nouns, main

verbs, auxiliary verbs, adjectives, adverbs, pronouns, abbreviations, and residuals. Notice that we neither have normalized the vector of features, nor took the logarithm of the feature values; this is a limit when one of the features has a value much larger than the values of the other features, masking their contributions to the distance. Thus, these results could be less useful than others reported in this article could. The results based on distances in the POS feature space are given in Table 3 for Iorga's and Averescu's texts.

The sub-sections as resulted from the analysis of the Euclidean distances in the POS space are identical with those obtained with the c-Delta methods, substantiating the conclusion that there is a better sectioning of the text than the one proposed by the author, for both texts, see Tables 1, 2 and 3.

4. COMPARISON OF THE RESULTS OBTAINED IN SEVERAL FEATURE SPACES

Formerly, we proposed several color indices for stylistic characterization and segmentation of texts [18, 19]. While color indices also operate at the level of the lexicon, they address the functional lexicon, that is, the use (ratios of frequencies) of several POS. Therefore, the two characterizations, one based on the c-Delta distance and the other on color indices address different aspects of the text. Consequently, the obtained results of text segmentation may differ. However, when both representations of the text produce the same segmentation, there is reason to be more confident that the segmentation is meaningful.

We are also interested in the values of several color indices [18, 19], namely: verbal color index $I_{VC} = Adv / \left(Main_{verbs} + Aux_{verbs} \right)$, nominal color index $I_{NC} = Adj / Nouns$, and total color index I_C defined as the average of the first two ones. The Euclidean distances in the color indices space, for Iorga's and Averescu's texts are shown in Table 4.

Table 4

The Euclidean distances in the space of the color indices between the sub-sections in Iorga's and Averescu's texts

Distances	**Values,** Iorga	**Values,** Averescu	**Conclusions**, Iorga's text	**Conclusions**, Averescu's text
$d_{CI}(S_{12}, S_1)$	0.03766	0.01166	$d_{CI}(S_{12}, S_1) < d_{CI}(S_{12}, S_2)$; S_{12} remains with S_1	$d_{CI}(S_{12}, S_1) < d_{CI}(S_{12}, S_2)$; S_{12} remains with S_1
$d_{CI}(S_{12}, S_2)$	0.12901	0.03708		
$d_{CI}(S_{21}, S_1)$	0.15942	0.01702	$d_{CI}(S_{21}, S_1) > d_{CI}(S_{21}, S_2)$; S_{21} remains with S_2	$d_{CI}(S_{21}, S_1) < d_{CI}(S_{21}, S_2)$; S_{21} should be adjoined with S_1
$d_{CI}(S_{21}, S_2)$	0.06178	0.04209		
$d_{CI}(S_{23}, S_2)$	0.00787	0.05003	$d_{CI}(S_{23}, S_2) < d_{CI}(S_{23}, S_3)$; S_{23} remains with S_2	$d_{CI}(S_{23}, S_2) < d_{CI}(S_{23}, S_3)$; S_{23} remains with S_2
$d_{CI}(S_{23}, S_3)$	0.00988	0.09346		
$d_{CI}(S_{32}, S_2)$	0.12142	0.08691	$d_{CI}(S_{32}, S_2) < d_{CI}(S_{32}, S_3)$; S_{32} should be adjoined with S_2	$d_{CI}(S_{32}, S_2) > d_{CI}(S_{32}, S_3)$; S_{32} remains with S_3
$d_{CI}(S_{32}, S_3)$	0.12643	0.02171		

Notice that the results according this criterion (stylistic color feature space) differ from the ones obtained by c-Distance. The sections resulted from the analysis of the Euclidean distances between color indices are, for Iorga's text: $S_1 = N_{10}$, $S_2 = N_{20} \cup S_{32}$, and $S_3 = N_{30} \setminus S_{32}$, while the sections in Averescu's text are $S_1 = N_{10} \cup S_{21}$, $S_2 = N_{20} \setminus S_{21}$, and $S_3 = N_{30}$.

We also calculated the stylometric indices (Hapax-Legomena, Sichel, Honore and Theta, (see [19]) and then the Euclidean distances between the vectors with the above components. We recall the definitions of several indices: Hapax-Legomena $\left(Hapax = V(1,N)/N \right)$; Sichel's measure $\left(S = V(2,N)/V(N) \right)$; Honore's measure $\left(H_o = \frac{100 \log N}{1 - V(1,N)/V(N)} \right)$; Theta measure $\left(\theta = N/V(N) \right)$, where $\{V(i,N) | i = 1,2\}$ = the number of lemmas occurring once $(i = 1)$, twice $(i = 2)$ in the text; N = number of lemmas and $V(N)$ = number of co-occurrences of lemmas. The sections resulted in Iorga's text (Table 5) are then $S_1 = N_{10}$; $S_2 = N_{20} \cup S_{32}$; $S_3 = N_{30} \setminus S_{32}$.

Table 5

Euclidean distances in the space of the stylometric indices between the sub-sections in Iorga's and Averescu's texts

Distances	**Values,** Iorga	**Values,** Averescu	**Conclusions,** Iorga's text	**Conclusions** Averescu's text
$d_{\text{Stilo}}(S_{12}, S_1)$	70.358	6.0294	$d_{\text{Stilo}}(S_{12}, S_1) < d_{\text{Stilo}}(S_{12}, S_2)$, thus S_{12} remains with S_1	$d_{\text{Stilo}}(S_{12}, S_1) < d_{\text{Stilo}}(S_{12}, S_2)$, thus S_{12} remains with S_1
$d_{\text{Stilo}}(S_{12}, S_2)$	146.357	22.0890		
$d_{\text{Stilo}}(S_{21}, S_1)$	38.226	23.7575	$d_{\text{Stilo}}(S_{21}, S_1) > d_{\text{Stilo}}(S_{21}, S_2)$, thus S_{21} remains with S_2	$d_{\text{Stilo}}(S_{21}, S_1) > d_{\text{Stilo}}(S_{21}, S_2)$, thus S_{21} remains with S_2
$d_{\text{Stilo}}(S_{21}, S_2)$	37.772	4.3626		
$d_{\text{Stilo}}(S_{23}, S_2)$	67.308	28.4333	$d_{\text{Stilo}}(S_{23}, S_2) < d_{\text{Stilo}}(S_{23}, S_3)$ thus S_{23} remains with S_2	$d_{\text{Stilo}}(S_{23}, S_2) > d_{\text{Stilo}}(S_{23}, S_3)$, thus S_{23} remains with S_3
$d_{\text{Stilo}}(S_{23}, S_3)$	72.894	16.1325		
$d_{\text{Stilo}}(S_{32}, S_2)$	117.667	23.3404	$d_{\text{Stilo}}(S_{32}, S_2) < d_{\text{Stilo}}(S_{32}, S_3)$, thus S_{32} remains with S_2	$d_{\text{Stilo}}(S_{32}, S_2) > d_{\text{Stilo}}(S_{32}, S_3)$, thus S_{32} remains with S_3
$d_{\text{Stilo}}(S_{32}, S_3)$	123.254	11.0397		

The results for Iorga's text are the same for the c-Delta (feature space of lemmas: the first 50 lemmas in descending order – Table 1), color indices (Table 4) and stylometrc indices (Table 5). In Averescu's text the sections are identically only for the feature space of lemmas and stylometric indices. The aggregated results are shown in Table 6. The results are interesting from two points of views: first, they show that, at least based on a style feature space, the volumes discussed could have been segmented in different manners than proposed by the authors (or by the readers); second, they show that various style feature spaces carry different information. The second conclusion requires further research for an explanation.

Table 6

Overview of the sub-sections resulted from the analysis in several spaces with Euclidean distances

Feature space: vectors of	**Resulted text segments**	
	Iorga's text	**Averescu's** text
Lemmas	$S_1 = N_{10}$; $S_2 = N_{20} \cup S_{32}$; $S_3 = N_{30} \setminus S_{32}$	$S_1 = N_{10}$; $S_2 = N_{20} \setminus S_{23}$; $S_3 = S_{23} \cup N_{30}$
Parts of speech	$S_1 = N_{10} \cup S_{21}$; $S_2 = (N_{20} \setminus S_{21}) \cup S_{32}$; $S_3 = N_{30} \setminus S_{32}$	$S_1 = N_{10} \cup S_{21}$; $S_2 = (N_{20} \setminus S_{21}) \setminus S_{23}$; $S_3 = N_{30} \cup S_{23}$.
Color indices	$S_1 = N_{10}$; $S_2 = N_{20} \cup S_{32}$; $S_3 = N_{30} \setminus S_{32}$	$S_1 = N_{10} \cup S_{21}$; $S_2 = N_{20} \setminus S_{21}$; $S_3 = N_{30}$.
Stylometric indices	$S_1 = N_{10}$; $S_2 = N_{20} \cup S_{32}$; $S_3 = N_{30} \setminus S_{32}$	$S_1 = N_{10}$; $S_2 = N_{20} \setminus S_{23}$; $S_3 = S_{23} \cup N_{30}$

5. DISCUSSION AND CONCLUSIONS

The notion of stylistic coherence and the related partitioning method extending Burrows' Delta distance proved to work in case of two large texts (Averescu's and Iorga's memoirs) that were otherwise sectioned in detail, on three levels, by the author (Iorga's text) or preliminary divided by us based on the phases of WWI recounted (Averescu's text). The method revealed a potentially better (stylistically more uniform) way of sectioning the text in the three major parts and evidences a subsection (S_{21}) which is stylistically perturbing the transitions between sections, as it is very different from both adjacent parts.

There are various ways for defining coherence measures based on the vectors $\mathbf{V}_j$. For example, $k = \min_j \left(1 - \frac{\text{STDEV}(|\mathbf{V}_j|)}{\text{average}(|\mathbf{V}_j|)}\right)$ could be a good candidate [25]. For stylistically uniform texts, k will be close to 1 (low spreading compared with the average value). If the text has a high coherence, there is no justification to apply the proposed method, except when there is one or a few parts of the text that are

significantly different from the others, yet their contribution to the spreading is low because of their reduced number. To check if this particular case arises, the distances between the vectors $\mathbf{V}_j$ are computed; if there is at least one pair of vectors, $\mathbf{V}_j$, $\mathbf{V}_k$ such that $d(\mathbf{V}_j, \mathbf{V}_k) >> \text{average}_{(h,i)}\, d(\mathbf{V}_i, \mathbf{V}_h)$, then the use of the proposed method is justified.

The proposed method can easily be refined, for example by variably selecting subparts or even by continuing the stylometric analysis at the level of starting and ending paragraphs at the level of subsections. We considered sections between 5-10%, preferably 10% of the total number of pages, to produce a relevant statistics. The granularity of the division cannot be extra fine, because else it would not be a reasonable statistical base for the lexicon. We believe the text segmentation method based on c-Delta and the proposed related algorithm(s) may find applications in the media and in classes on writing and literature. However, we are aware of the various limitations [26] of the stylistic analysis.

ACKNOWLEDGMENTS

Our thanks to the reviewers for their very useful comments. HNT thanks Prof. M.H. Teodorescu for several ideas on text partitioning and stylometric coherence.

Authors' contributions. C.B. performed the parsing and all computations, derived the tables and contributed writing the Results section. HNT proposed the research, the principles and definition of the cohesion-based segmentation based on the proposed cohesion Delta measure, the algorithm, and wrote most of the paper. Both authors agreed with the final form of the paper.

REFERENCES

1. S.H.H. DING, B.C.M. FUNG, F. IQBAL, W.K. CHEUNG, *Learning stylometric representations for authorship analysis*, IEEE Transactions on Cybernetics, **49**, *1*, pp. 107-121, 2019.
2. M.L. BROCARDO, I. TRAORE, S. SAAD, I. WOUNGANG, *Authorship verification for short messages using stylometry*, Int. Conf. Computer, Information and Telecommunication Systems (CITS), pp. 1-6, 2013.
3. D.I. HOLMES, J. KARDOS, *Who was the author? An introduction to stylometry*, Journal CHANCE, **16**, *2*, pp. 5-8, 2003.
4. M.L. BROCARDO, I. TRAORE, S. SAAD, I. WOUNGANG, *Verifying online user identity using stylometric analysis for short messages*, Journal of Networks, **9**, *12*, pp. 3347-3355, 2014.
5. I.N. ROTHMAN, *Defoe de-attributions scrutinized under Hargevik criteria: Applying stylometrics to the Canon*, The Papers of the Bibliographical Society of America, **94**, *3*, pp. 375-398, 2000.
6. R. SARAWGI, K. GAJULAPALLI, Y. CHOI, *Gender attribution: tracing stylometric evidence beyond topic and genre*, Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pp. 78-86, 2011.
7. M. POTTHAST, J. KIESEL, K. REINARTZ, J. BEVENDORFF, B. STEIN, *A stylometric inquiry into hyperpartisan and Fake News*, Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, **1**, pp. 231-240, 2018.
8. J. SUN, Z. YANG, S. LIU, P. WANG, *Applying stylometric analysis techniques to counter anonymity in cyberspace*, Journal of Networks, **7**, *2*, pp. 259-266, 2012.
9. G. BROUSALIS, G. MIKROS, *Stylometric profiling of the Greek Legal Corpus*, Selected papers of the 10th ICGL, pp. 167-176, 2012.
10. R. SHAMS, R. E. MERCER, *Supervised classification of spam emails with natural language stylometry*, Neural Computing and Applications, **27**, *8*, pp. 2315-2331, 2016.
11. S. BERGSMA, M. POST, D. YAROWSKY, *Stylometric analysis of scientific articles*, Proc. of the Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2012, pp. 327-337.
12. M. ALSALLAL, R. IQBAL, S. AMIN, A. JAMES, *Intrinsic plagiarism detection using latent semantic indexing and stylometry*, Sixth International Conference on Developments in eSystems Engineering, 2013, pp. 145-150.
13. S. GOSWAMI, S. SARKAR, M. RUSTAGI, *Stylometric analysis of bloggers' age and gender*, Proc. of the Third Int. Conf. on Weblogs and Social Media, 2009, pp. 214-217.
14. M. BRENNAN, S. AFROZ, R. GREENSTADT, *Adversarial stylometry: circumventing authorship recognition to preserve privacy and anonymity*, ACM Transactions on Information and System Security, **15**, *3*, article 12, 2012.
15. D. I. HOLMES, S SINGH, *A stylometric analysis of conversational speech of aphasic patients*, Literary and Linguistic Computing, **11**, *3*, pp. 133-140, 1996.
16. L. BRANDWOOD, *Stylometry and chronology*, The Cambridge Companion to Plato, 1992, pp. 90-120.
17. H.N. TEODORESCU, *The dynamics of the words*, Abstract, The 11th Conference on Applied and Industrial Math., 2003.
18. H.N.L. TEODORESCU, S.C. BOLEA, *Stylometric and topic analysis of a historical text a computerized study of general Averescu's 'War Memoirs'* (1916–1918), Romanian Journal Information Science and Technology, **21**, *2*, pp. 99-113, 2018.

19. H.N. TEODORESCU, S.C. BOLEA, *Text sectioning based on stylometric distances*, Int. Conf. Speech Technology and Human-Computer Dialogue (SpeD), 2019.
20. S. EVERT, T. PROISL, F. JANNIDIS, I. REGER, S. PIELSTRÖM, C. SCHÖCH, T. VITT, *Understanding and explaining Delta measures for authorship attribution*, Digital Scholarship in the Humanities, **32**, issue suppl_*2*, pp. ii4-ii16, 2017.
21. C. DEGRAZIA, N. PAIROLERO, M. TEODORESCU, *Shorter patent pendency without sacrificing quality: The use of examiner's amendments at the USPTO*, USPTO Economic Working Paper No. 2019-03, available at SSRN: https://ssrn.com/abstract=3416891 or http://dx.doi.org/10.2139/ssrn.3416891, 2019.
22. M. PEARTON, *Nicolae Iorga as historian and politician*, in: "Historians as Nation-Builders, Central and South-East Europe" (eds. D. Deletant, H. Hanak), 1988, pp. 157-173.
23. W.O. OLDSON, *The historical and nationalistic thought of Nicolae Iorga*, East European Quarterly, Columbia Univ. Press, 1974.
24. TTL Parser, http://www.racai.ro/tools/text/.
25. M.H. TEODORESCU, personal communication, 2020.
26. T. SCHUSTER, R. SCHUSTER, D. J SHAH, R. BARZILAY, *The limitations of stylometry for detecting machine-generated fake news*, Computational Linguistics Accepted for publication, pp. 1-18, 2020.

Received March 10, 2020