Similarity Computation based on the Tails of the Rank Distributions and the Related Graphs

Cecilia Bolea
Institute of Computer Science
Romanian Academy
Iasi, Romania
cecilia.bolea@iit.academiaromana-is.ro

Mike HM Teodorescu
Information School
University of Washington
Seattle, WA, US
https://orcid.org/0000-0001-7330-832X; miketeod@uw.edu

Silviu Bejinariu
Institute of Computer Science
Romanian Academy
Iasi, Romania

Daniela Gîfu
Institute of Computer Science
Romanian Academy
Iasi, Romania

Horia-Nicolai Teodorescu
"Gheorghe Asachi" Technical
University of Iasi
Iasi, Romania
hteodor@etti.tuiasi.ro

Vasile Apopei
Institute of Computer Science
Romanian Academy
Iasi, Romania
vasile.apopei@iit.academiaromana-is.ro

***Abstract*—Tools for document analysis, characterization, and retrieval are introduced based on a rigorous framework. The procedure is based on an accurate alignment method for text of different lengths. The alignment refers to connected rare words in short texts with more frequent words in longer texts, where the connection is performed according to Zipf's law. A discussion of the algorithmic approach is presented. An algorithm suitable for parallelization is presented.**

Keywords — corpus analysis, rank distribution, power law, NLP, graph, similarity, style analysis, authorship analysis

I. Introduction

In many applications there is a need to determine the specificities and the novelty of a document or to extract relevant information. Such applications include authorship analysis, keyword extraction, historical analysis to date woks (Smith & Kelly, 2002) [1], information retrieval, automatic classification of patents, and style analysis [2], [3], [4], [5], [6], [7]. Some approaches are based on rare words and style. However, some authors, including Hoover (2003) [8] contested the validity of the vocabulary richness for authorship attribution and consequently in some other domains, such as style analysis and information retrieval. In fact, the TF-IDF retrieval method does not account for the size of the document, a limit that is overcome by the presented method. For a good example of a comparison of Zipf's laws occurring in various types of texts (Brown Corpus, the patent titles corpus, the patent abstracts corpus, and the patent claims corpus), see [9], Figure 3, page 76.

Hoover [8] has been highly critical of methods of authorship attribution based on vocabulary, saying that "vocabulary richness is ineffective for large groups of texts … vocabulary richness is of marginal value in stylistic and authorship studies because the basic assumption that it constitutes a word print for authors is false." In line with authors who question some of the methods of authorship attribution, we start from the power law distribution association with large corpora to derive corrections to style analysis based on rare words. However, the method is computationally intensive and may require parallelization.

Specifically, we propose a method of text comparison in large corpuses based on the relationships between hapax legomena, dis- and tris-legomena, where the relationships are given by the distance between words (or lemmas) in these categories. Recognizing the arguments of Hoover [8], we suggest a method or wrapping these categories based on a 'normalization' method accounting for the number of words in the texts. The method is based on [10].

In Section II we recall elements on Zipf's law and power distributions. Section III recalls the notions of hapax legomena, dis- and tris-legomena and intuitively presents the main ideas of the paper. Section IV analyzes some properties induced by the integer counts and ranks; these represent the foundation of the algorithm described in the same Section. The last section of the paper is a discussion and conclusion.

II. Context: Power Law Distribution

Power law distributions, where the probability (occurrence frequency) of objects is proportional with a negative power of the rank, are present in numerous natural, socio-economic, linguistic, and network-related processes. City sizes, incomes, and words in natural language texts exhibit such distributions. In case of text corpora, Zipf ranked the words according to their frequencies and has shown that the frequency of occurrence of a word is inversely correlated with the rank of that word [11], $f(r, \alpha, C) = C \,/\, r^{a}$, where f is the frequency of occurrence of the word, r is the rank, C is a constant and a is the exponent that characterizes the distribution. With small variations, the Zipf's law is valid for all languages [12]. Thurner [13] determined that it is also applicable to the spoken language; it applies not only at the level of words but also to a range of linguistic units. [12] have shown that for 50 languages Zipf's laws share a similar 3-segment structural pattern. It is unclear if Zipf's law in languages has roots in human cognitive mechanisms.

III. Foundations and Main Ideas

Throughout the paper we assume a corpus containing numerous texts of various numbers of words in a specified

language, where the corpus and the included texts may be assumed to have approximately the same rank distribution. Consider a text where the distributions of the less frequent words (lemmas) follow a power law (Zipf's law). It is well known that the distant part (high ranks / lower probabilities) of these graphs of these distributions resemble a "broom tail". An example is shown, for a theoretical power law with $a = 1.7$, in Fig. 1, where the meaning of the constant $H = 2.042889$ is explained later. Actually, the problem of "broom tail" is not specific to the power law distribution, but to all non-uniform distributions of discrete variables with countable number of types [10]; in all these cases, it is a result of quantization. For example, Fig. 2 shows the graph of the empirical rank distribution $ln(n_w) \sim ln(r_w)$, n_w denoting the number of occurrences of the word w in the text and r_w denoting its rank, for the autobiographic volume "Under Three Kings" (UTK), ("Supt Trei Regi", in original Romanian text) by Nicolae Iorga (1932). Computing specificities for the tail for any type of distribution with quantized values is similar to the computations for Zipf's law, which will be shown in the next Section.

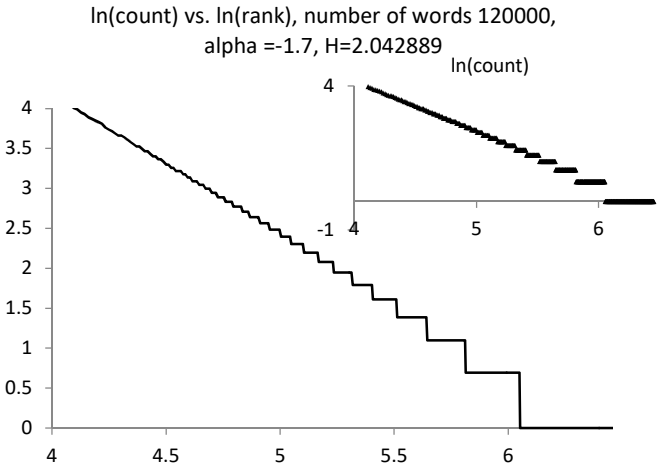

Fig. 1. An ideal (theoretical) model for a text of 120'000 words, with $\alpha = -1.7$, under the convention $f(r, \alpha, C) = C \cdot r^{\alpha}$. Notice in the inlay, where each datapoint is represented individually, whitout connecting lines, the broom-like end.

Stylistic analysis, authorship attribution, and text segmentation, among others, use various lexically based measures, lexically based style markers / features, syntax-based markers [14], [15], [16], [17], [18], [19], words / part of speech n-grams [21], [20], or dependency relationships [21]. Among these markers are words with a single occurrence in the text (hapax legomena) and words with two or three occurrences (dis- and trislegomena) [22], [15]. Madden, Storey, and Baskerville [19] intuitively noticed that Zipf's law imposes some form of "regularization", "taking into account the total number of words" in the text; they used division by the number of words as a regularization method.

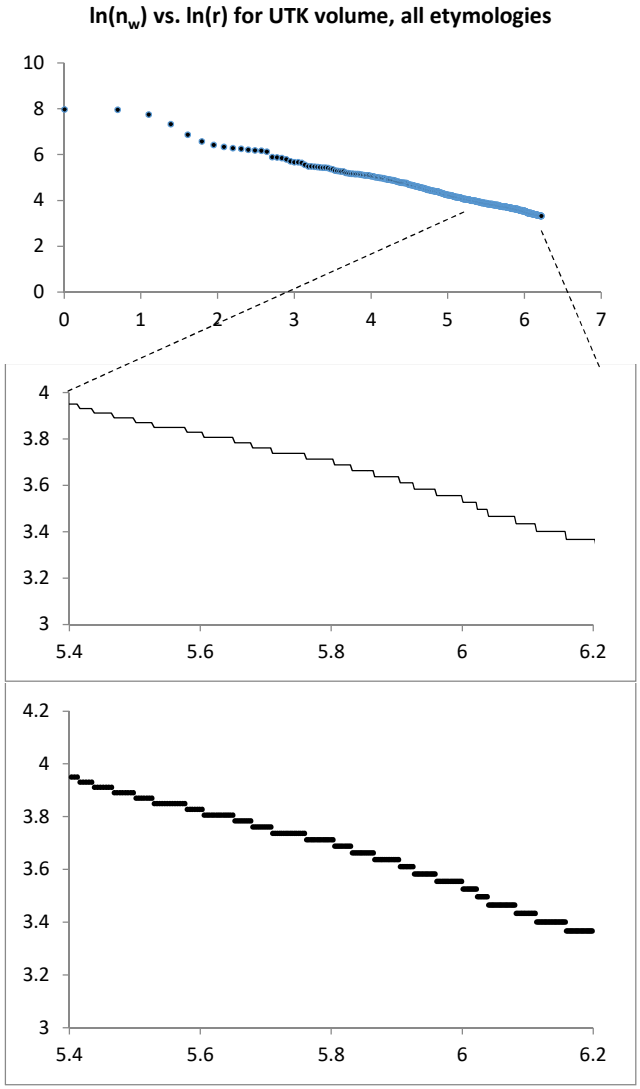

Fig. 2. Graph of the logarithm of count of lemmas versus the logarithm of rank, with details in continuous lines and bullets (markers) showing several lemmas / words with the same count but different, successive ranks in the text UTK (Under three kings, by N. Iorga) selfbiography.

It may be interesting to find departures from the expected "broom"-like shape of the distant end of the distribution and to determine differences and similarities of these departures between different populations (i.e., texts, text corpora etc.). Such departures are easy to find for individual texts and may be related with the author's style, hence to authorship. Hapaxes, dis- and tris-legomena are at the last few levels of the distributions. The sets of these words may be similar in different texts, pointing to same authorship or style similarities. We suggest that, beyond their numbers and set similarities between different texts, the dependency relationships between hapaxes, dis- and tris-legomena may show similarities among texts, thus bringing more information. However, numerical considerations indicate that a direct comparison is not possible for texts of different lengths and that some form of alignment is needed, see below and Section IV.

Denote by H the hapaxes, by D the dis-legomena, and by T the tris-legomena. We use the distance between words defined by the number of words, in the sequence represented by the text, between the words. Denote by T_1', T_1'', T_1''' the set of three instances representing the first tris-legomenon (T_1') appearing in the text, by T_2', T_2'', T_2''' the second a.s.o. Denote by D_1', D_1'' the set of two instances of the first dis legomenon in the text and by $H_1, H_2 \ldots$ the hapaxes. Start with T_1' and determine the closest dis legomenon to it, D_{11}; denote the distance between them by $d(T_1, D_{11})$. Determine the closest hapax to D_{11}, H_{111}; denote the distance between them by $d(H_{111}, D_{11})$. The described operation corresponds to a tripartite graph and to a method of building edges between the tris-legomena nodes and the dis-legomena, then from dis-legomena to hapaxes, see Fig. 1, where $D_{11} = D_1'$ (necessarily), $H_{111} = H_1$ etc.

It is possible that a dis-legomenon is the closest neighbor to two or more tris-legomena. This count of edges is of interest and will be denoted by $\nu(D_1), \nu(D_2), \ldots$ Similarly, a hapax may be the closest to several dis legomena; the count of the corresponding edges is denoted as $\nu(H_1), \nu(H_2), \ldots$ Then, a text is defined by the vectors $\{\nu(D_1), \nu(D_2), \ldots\}$ and $\{\nu(H_1), \nu(H_2), \ldots\}$. By definition, $\nu(D_i) = \nu(D_i') + \nu(D_i'')$. Two texts will be considered similar (in content, meaning) when the corresponding two vectors are similar, where the vector similarity is defined in a typical way, for example, by cosine similarity. However, the comparison of texts of different sizes requires one more step, explained later.

A further example of usefulness of the method described is as follows: The novelty degree of a work (patent) – compared with the domain is given by the number of hapaxes in the domain that are included in the work times their count in the work + ½ times the number of dis-legomena in the domain that are included in the work times their count in the work + 1/3 times number of before-before-hapaxes (tris-legomena) in the domain that are included in the work times their count in the work. One can imagine various approaches of comparing works based largely on the same principles.

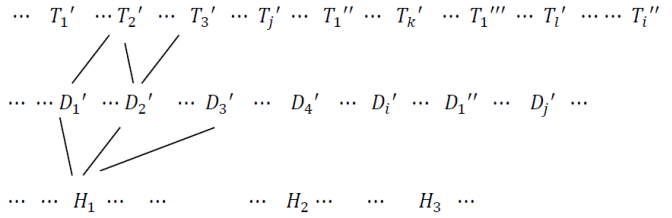

Fig. 3. Explanation for the graph and relational counts for the hapaxes, dis-, and tris-legomena in a text: $\nu(D_1') = 1, \nu(D_2') = 2, \nu(D_1) = \nu(D_1') + \nu(D_1'')$, $\nu(H_1) = 3, \ldots$

The graph in Fig. 3 and the related counts ν establish a set of proximity relations between the most infrequent words in a text.

The issue with the method described is anticipated by Hoover [8]: the results are dependent on the dimensions of the texts. The first solution proposed is to determine foremost the relations ν for the shortest text, then to search in the longer text the words representing the tris and dis legomena and the hapaxes in the shortest text and to determine for them the relations, then to scale to the sum of the counts ν in the text. Only after normalization for both texts one should make the comparison, for it is meaningful. The second solution proposed is to consider that, when the test has a larger number of words, hapaxes 'migrate' to a higher level (dis or tris legomena, or even higher). Therefore, for comparing texts of different sizes, one needs to align the multi-legomena in the larger text to the hapaxes, dis-, and tris-legomena of the shorter texts, considering the relationship established by the respective text dimensions. The next section shows the fundamentals of the procedure.

IV. Properties Induced by the Integer Counts and Ranks

The discussion in this Section tightly reproduces [10].

When a large text with thousands of different lemmas (or distinct words, for the matter) is analyzed and the graph of the logarithm of lemma counts versus logarithm of the rank of the lemmas is drawn, one obtains a picture as in Fig. 2, upper panel.

For two populations of dimensions N_1 and N_2, the first having m_1 types and the second having m_2 types, because of the different population counts and different types, the comparison is not trivial.

We count the number of elements of type t in the population, $n(t)$, and sort these counts descendingly; the type with the largest count is ranked first, and the count will be denoted $n(r = 1)$ or $n(1)$. We admit that the studied population may be a sub-multiset of a larger multiset U and that some types in U may lack in the studied population; however, we are interested in ranks up to $n(r) \geq 1$, i.e., ranks of the elements of the population, not of U.

One is interested in counts that are natural numbers and ranks are natural numbers. Therefore, the power laws have the form

$$n(r) = \frac{A}{r^a}, a \in \mathbf{R}, a > 0, A \in \mathbf{R}, \;\; A > 1. \qquad (1)$$

Because the counts $n(r)$ are integers, the equality in (1) is not generally possible; therefore, one should interpret (1) as the rounded $n(r) = \left[\frac{A}{r^a}\right]$, or the transformation of the right side to an integer can be done using the floor or the ceiling functions. Given the concept of rank and from the rank ordering it results that, for any $h < j, n(h) \geq n(j)$.

The definition (1) is reminiscent of Riemann function $\zeta(s) = \sum_{r=1}^{\infty} \frac{1}{r^s}$, $s \in \mathbf{C}$. We will use several elementary properties of this function, among others the values of $\zeta(2)$ and $\zeta(3)$. Because the populations we are concerned with are finite, the sums of counts for all ranks are partial sums of ζ up to the maximal rank r_M, the counts (number of elements) in the populations are given by

$$S_{r_M}(a) = \sum_{r=1}^{r_M} n(r) = A \sum_{r=1}^{r_M} \frac{1}{r^a} \qquad (2)$$

and are finite; we are not concerned with the convergence of S_{r_M}, although we use the asymptotic approximation according to $\zeta(s)$. For larger even values of the exponent, the partial series S_{r_M} tend to the values computed as in [22]. Notice that

$S_{r_M}(a) = A \cdot H_{r_M}^{(a)}$, where here $H_m^{(a)}$ denotes the harmonic number with exponent a; standard harmonic numbers are $H_m^{(1)}$.

The actual counts may be either modeled as rounded numbers, as discussed in [10] for $a = 1$, or as variables with attached probabilities. For example, the count for rank r may have two values, $\left\lfloor \frac{A}{r^a} \right\rfloor$ with probability p_1, or $\left\lceil \frac{A}{r^a} \right\rceil$ with probability p_2. For applications, one may choose $p_1 = p_2 = 0.5$, or one may chose $p_1 = \int_r^{r+0.5} \frac{B}{r^a} dr$, $p_2 = \int_{r+0.5}^{r+1} \frac{B}{r^a} dr$, $B = 1/\int_r^{r+1} \frac{1}{r^a} dr$. For now, we will use the adjusted model based on rounding,

$$n(r) = \left[\frac{A}{r^a} \right]. \tag{5}$$

The next problem is that, for large enough values of r, one obtains

$$n(r) = \left[\frac{A}{r^a} \right] = \left[\frac{A}{(r+1)^a} \right] = n(r+1). \tag{6}$$

But this equality contradicts the unicity of ranks in statistics where only "entire" elements are counted, such as entire words, not fractions of them. Therefore, two or several successive ranks have to be allowed to correspond to the same count. It is why in the Preliminaries section we allowed for the condition that for any two ranks $h < j$, $n(h) \geq n(j)$ instead of using the strict condition $n(h) > n(j)$.

We may be interested if there is a threshold rank, r_0, such that for ranks $r < r_0$ the model predicts a single rank for a given count, while for ranks $r > r_0$ the model predicts the possibility of two or more ranks with the same count. Equivalently, the conditions defining r_0 are

$$r < r_0 \rightarrow \frac{A}{r^a} \geq \frac{A}{(r+1)^a} + 1, \tag{7}$$

$$\forall r \geq r_0, j \geq 1: \ \frac{A}{(r+j)^a} - \frac{A}{(r+j+1)^a} < 1. \tag{8}$$

The first condition, (7), leads to the equation ($A, r \in \mathrm{N}$)

$$A(r+1)^a \geq r^a (A + (r+1)^a), \tag{9}$$

with r_{01} the smallest value satisfying the above. The minimal value of r in the second condition (8) leads to r_{02}, if $r_{02} < r_M$. However, it is possible that the two conditions produce different values, $r_{01} \neq r_{02}$; we chose $r_0 = r_{01}$. Both conditions have to be satisfied for r_0 has the sense of the maximal range with a count guaranteed different from the next one.

To simplify the discussion, for the count at rank r we use in the remaining part of the paper the definition based on the floor function,

$$n(r) = \left\lfloor \frac{A}{r^a} \right\rfloor. \tag{10}$$

Equation (10) allows us to find the corresponding ranks in two texts of different sizes and to align them, based on a rigorous relationship, in view of comparing the vectors of hapaxes and dislegomena.

Consider two texts V_1 and V_2 of different lengths (i.e., number of words), with maximal ranks r_{M1} and r_{M2}; these ranks are derivable from (2), knowing their total numbers of words, $N_1 < N_2$ (assuming V_1 has fewer words than V_2). In V_1, all words with rank r_{M1} are hapaxes, those with rank $r_{M1} - 1$ are dislegomena, and those with rank $r_{M1} - 2$ are tris-legomena. However, hapaxes in V_1 are not hapaxes in V_2. Ideally, assuming both texts have precisely the same distributions, hapaxes in V_1 will preserve the same rank in V_2, but their number will be larger than 1 and given by $n(r_{M2}|V_2) = \frac{A_2}{r_{M2}^a}$. Therefore, the "alignment" between the two texts requires that we divide the actual number of words of rank r_{M2} in V_2 by A_2/r_{M2}^a. In applications, the texts V_1 and V_2 will be similar when the words that are hapaxes in V_1 have, in V_2, frequencies $n(r_{M2}|V_2)/(A_2/r_{M2}^a)$ close to 1. Applied to the problem of authorship attribution, being given a text V_1 with unknown author, we will look for all the texts in the corpora for those having the hapaxes in V_1 on the same (or close) rank r_{M1}, moreover having the set of values $n(r_{M2}|V_2)/(A_2/r_{M2}^a)$ closest to 1.

Further, for using relationship information where relationship is defined by distance (vicinity), we first establish for the text V_1 the distances between hapaxes, dis- and tris-legomena (Fig. 3), retaining only those pairs that are at a distance lower than a specified one (e.g., immediate vicinity). Then we search the other texts for the same couples of neighbors, again taking into account their regularized frequency (as explained for words, above). Texts that have similar neighbors, with similar frequencies, are considered related (by authorship and/or style).

Clearly, the algorithm suggested by Fig. 3 and described above is easily parallelizable, where the parallelization is performed over subsets of the corpus, or even at the level of documents.

V. Discussion and Conclusions

The proposed method overcomes the deficiencies of the brute force comparison of the set of hapaxes, dis-legomena etc.

While the computational effort required by the approach presented here and by others derivable from the same principles is much more important, we believe that it is worth doing, even necessary in some application areas. We argue that, because authorship attribution is often related to ethical, legal, and social repercussions for the analyzed authors and thus, errors in results can cause a number of negative consequences, efforts are justified for avoiding errors in results. Correcting simplistic methods and algorithms that may have severe negative socio-economic effects is an issue of fairness and has been analyzed in the machine learning fairness literature, such as in human-ML augmentation [23], and intensive statistical analysis of the results [24].

There are various ways of extending the principles suggested in this paper. For example, one method is to compare the

empirical distributions of both texts with the distribution of the theoretical law and generate the respective vectors of differences. The number of differences and the rank of the differences times the differences values are indicators of the difference of the compared populations.

Authors' Contributions. CB and SB contributed Section II (with editing by HNT); CB also collected and generated data used in Fig. 2. DG provided input on corpora and their applications and the impetus for part of [10]. MT and HNT conceived the method and applications and contributed the material in Sections I, III-V. HNT wrote the paper with input primarily from MT and SB. VA supervised part of the work.

REFERENCES

[1] J.A. Smith, Kelly, C., "Stylistic Constancy and Change Across Literary Corpora: Using Measures of Lexical Richness to Date Works." Computers and the Humanities 36, 411–430 (2002). https://doi.org/10.1023/A:1020201615753.

[2] H. Baayen, "Statistical models for word frequency distributions: A linguistic evaluation." Comput Hum 26, 347–363 (1992). https://doi.org/10.1007/BF00136980.

[3] H. Baayen, (1996). "The effects of lexical specialization on the growth curve of the vocabulary." Computational Linguistics, 22(4), 455-480.

[4] R.L. Brooks, "In your own words, please: using authorship attribution to identify cheating on translation tests." In: Lynda Taylor and Cyril J Weir (Eds.), Language Testing Matters - Investigating the wider social and educational impact of assessment. Proc. ALTE Cambridge Conf., Apr 2008. Cambridge University Press, Cambridge, UK.

[5] C.P. Chandrika, Kallimani JS. "Authorship Attribution for Kannada Text Using Profile Based Approach." In Proceedings of the 2nd International Conf. Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021 2022 (pp. 679-688). Springer Singapore.

[6] A. Corral, G. Boleda, Ferrer-i-Cancho R (2015) "Zipf's Law for Word Frequencies: Word Forms versus Lemmas in Long Texts." PLoS ONE 10(7): e0129031. https://doi.org/10.1371/journal.pone.0129031.

[7] M. Coulthard, "Author Identification, Idiolect, and Linguistic Uniqueness," Applied Linguistics, Vol. 25, Issue 4, Dec 2004, pp. 431–447, https://doi.org/10.1093/applin/25.4.431.

[8] D.L. Hoover, "Another Perspective on Vocabulary Richness." Computers & Humanities, 37, 151–178 (2003). doi.org/10.1023/A:1022673822140.

[9] M. Teodorescu, 2018. "Knowledge Flows and IP Within and Across Firms – Economics and Machine Learning Approaches." Doctoral dissertation, Harvard Business School. Permanent link http://nrs.harvard.edu/urn-3:HUL.InstRepos:41940978.

[10] HNL Teodorescu, "Big Data and Large Numbers. Interpreting Zipf's Law." https://arxiv.org/abs/2305.02687.

[11] G.M. Linders, Louwerse, M.M., "Zipf's law revisited: Spoken dialog, linguistic units, parameters, and the principle of least effort". Psychon Bull Rev 30, 77-101, 2023, https://doi.org/10.3758/s13423-022-02142-9.

[12] S. Yu, C. Xu, H. Liu, "Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation", arXiv:1807.01855v1 [cs.CL], 2018, https://doi.org/10.48550/arXiv.1807.01855.

[13] S. Thurner, Hanel R, Liu B, Corominas-Murtra B., "Understanding Zipf's law of word frequencies through sample-space collapse in sentence formation", J. R. Soc. Interface 12: 20150330, 2015, http://dx.doi.org/10.1098/rsif.2015.0330.

[14] H.-N. Teodorescu and S. C. Bolea, "Automatic Segmentation of Texts based on Stylistic Features," 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 2021, pp. 1-6, doi: 10.1109/SpeD53181.2021.9587362.

[15] E. Stamatatos, N. Fakotakis and G. Kokkinakis, Automatic Authorship Attribution. Proceedings of EACL '99, pp. 158-164.

[16] H.N.L. Teodorescu, S.C. Bolea, Stylometric and Topic Analysis of a Historical Text. Romanian Journal of Information Science and Technology – ROMJIST, Vol. 21, No. 2, 2018, pp. 99-113.

[17] S.C. Bolea, "Implementation of an Algorithm for Automatic Segmentation of Texts based on Stylometric Analysis," 2021 7th International Symposium on Electrical and Electronics Engineering (ISEEE), Galati, Romania, 2021, pp. 1-6, doi: 10.1109/ISEEE53383.2021.9628868.

[18] S.C. Bolea, H.N.L. Teodorescu, Applying a Delta-type measure to text coherence analysis and text segmentation. Proc. Ro. Acad., Series A, Vol. 21, No. 3, 2020, pp. 283-292.

[19] J. Madden, V. Storey, R. Baskerville, Identifying Authorship from Linguistic Text Patterns. Proceedings of the 52nd Hawaii International Conference on System Sciences, HICSS, 2019, pp. 5745-54.

[20] Y. Zhang, W. Wu, How effective are lexical richness measures for differentiations of vocabulary proficiency? A comprehensive examination with clustering analysis. Language Testing in Asia (2021), https://doi.org/10.1186/s40468-021-00133-6.

[21] H. Wu, Z. Zhang, Q. Wu, Exploring syntactic and semantic features for authorship attribution. Applied Soft Computing, Vol. 111, Nov. 2021, 107815.

[22] D. Faltýnek, V. Matlach, Hapax remains: Regularity of low-frequency words in authorial texts, Digital Scholarship in the Humanities, Vol. 37, Issue 3, Sep 2022, pp. 693–715, https://doi.org/10.1093/llc/fqab077.

[23] Ó. Ciaurri, LM. Navas, FJ. Ruiz & JL. Varona (2015) A Simple Computation of ζ (2k), The American Mathematical Monthly, 122:5, 444-451, DOI: 10.4169/amer.math.monthly.122.5.444

[24] M.H.M. Teodorescu, L Morse, Y Awwad, G Kane, Failures of Fairness in Automation Require a Deeper Understanding of Human–ML Augmentation. MIS Quarterly 45 (3), DOI: 10.25300/MISQ/2021/16535.

[25] M.H.M. Teodorescu and X. Yao, "Machine Learning Fairness is Computationally Difficult and Algorithmically Unsatisfactorily Solved," 2021 IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, USA, 2021, doi: 10.1109/HPEC49654.2021.9622861.