Research Article Open Access

Gabriela Andrejková* and Abdulwahed Almarimi

Dissimilarities Detections in Texts Using Symbol n-grams and Word Histograms

DOI 10.1515/comp-2016-0014
Received June 17, 2016; accepted July 20, 2016

Abstract: Texts (books, novels, papers, short messages) are sequences of sentences, words or symbols. Each author has an unique writing style. It can be characterized by some collection of attributes obtained from texts. The text verification is the case of an authorship verification where we have some text and we analyze if all parts of this text were written by the same (unknown or known) author. In this paper, there are analyzed and compared results of two developed methods for a text verification based on n-grams of symbols and on local histograms of words. The results of a symbol n-gram method and a method of word histograms for a dissimilarities searching in text parts of each text are analyzed and evaluated. The searched dissimilarities call for an attention to the text (or not) if the text parts were written by the same author or not. The attention depends on selected parameters prepared in experiments. Results illustrate usability of the methods to dissimilarities searching in text parts.

Keywords: n-grams of symbols, histograms of words, bag of words, stylistic measure, text dissimilarity

1 Introduction

In a text processing, there are solved problems like an intrinsic plagiarism, an external plagiarism, an authorship attribution, an authorship verification and a text verification. The above named problems are very important in time of Internet. Texts should be copied, combine and so on and in many cases it is necessary to cover these activities, for example in student's solutions of homeworks to cover copied parts of the solutions.

Some researchers contributed to the area by a developing of methods to cover an intrinsic and an external plagiarism [1–5]. The methods are based on n-grams characteristics. The other authors [6, 7] used local histograms methods to cover similarities of texts. The problems are followed by series of scientific events and shared tasks on digital text forensics - called PAN. Information and benchmarked texts of PAN are on web page [8]. Basic overviews of methods applied in solving of the problems are described in [9]. The used methods work with a set of texts mainly.

In our goal "to get information about each text (as much as possible)" we concentrate to a problem of the text verification. The text verification problem compares the similarity of a text of unknown authorship with all the text of a set of authors and search the most similar one. The most similar text and the text of unknown authorship are probably written by the same author. In the same way it is possible to compare parts of the given text and to find similar or dissimilar parts in this text. The problem *text part dissimilarities (TPD) problem* is the special case of the text verification where we have some given text and this text is split to coherent parts. We analyze if these parts are similar/dissimilar. A dissimilarity between two text parts can give information that text parts are written by different authors.

In the paper, we follow an information given by each text using some statistics on words, and on small substrings of texts (n-grams). Then the text is split into smaller parts and we compare the statistical values of these parts to the complete text and among them too. Our main idea is that "some characteristics of text parts follow characteristics of the complete text" (but it is necessary to cover these characteristics).

The paper has the following structure: In the second section, basic notions are defined and according to them the statistical analysis of some Arabic and English texts is described. The third section contains a description of symbol n-grams profile method and its application to some Arabic and English texts. In the fourth section, the analysis is done using histograms on words applied to occurrences of words in the text. The fifth section describes an evalua-

***Corresponding Author: Gabriela Andrejková:** Institute of Computer Sciencel, P. J. Šafárik University in Košice, E-mail: gabriela.andrejkoval@upjs.sk
Abdulwahed Almarimi: Institute of Computer Sciencel, P. J. Šafárik University in Košice, E-mail: abdoalmarimi@gmail.com

tion of both methods and in the conclusion, a summary of results and some plan for a future work is written.

2 Background and text statistics

In our text analysis we used English recommended texts from benchmark [8] and Arabic texts from [10, 11].
We use the following symbols and definitions:

- Γ - a finite alphabet of symbols; $|\Gamma|$ is the number of symbols in Γ; in our texts, Γ_A will be Arabic and Γ_E English alphabet;
- V - a finite vocabulary of words in the alphabet Γ presented in the alphabetic order; $|V|$ - the numbers of words in the vocabulary V;
- T - a text; $T = \langle w_1, \dots, w_N \rangle$; $w_i \in V$ - a finite sequence of words, N - the number of words in T;
- $T = \langle s_1 s_2 \dots s_M \rangle$, $s_i \in \Gamma$; M – the number of symbols in the text T;
- n-gram - the substring $s_{l+1}s_{l+2} \dots s_{l+n}$, $l = 1, \dots, M - n + 1$, $n = 1, 2, \dots, n$ - the length of n- grams;

2.1 English texts

The statistical analysis of 4 English texts is described in Table 1. The longest/shortest analyzed text E1/E4 has 176598/106359 words and 874761/471566 symbols. In both texts, 3-grams have highest frequency and 3-gram – the word *the* is the real word in English language. In the text E1, 4-gram – the word *that* has the highest frequency. In the text E4, the word *the* has 9.1% occurrences in the full text and 24.91% occurrences among all words of the length 3. The frequency analysis of all 4 texts is drawn in the Fig. 1, the right panel. The top of all texts is in the length 3, the 3 symbols long words have the highest frequency. The majority of words has the lengths 1 – 15. If texts are analyzed using 3-grams then it is analyzed 25% of real words. The number of all 3-grams in English alphabet (independent on capital symbols) is $|\Gamma_E|^3 = 26^3 = 17576$. It means, in the case of English language 3-grams are very usable.

Table 1: Statistics of 4 English texts, the number of words by length for 1-10 and maximal frequencies of 3 and 4-grams.

	Name of texts			
	E1	E2	E3	E4
Total-words	176598	132020	125487	106359
Total-symbols	874761	607783	498696	471566
Total diff. words	22954	19268	15853	15321
# words				
by length 1	5111	4642	4642	3474
by length 2	27250* 15.43%	17048	20418* 16.27%	17285
by length 3	**38733** 21.93%	**27599** 20.90%	**23780** 18.95%	**26497** 24.91%
by length 4	26321	20909* 15.83%	18429	19224* 18.07%
by length 5	16646	13328	15587	11459
by length 6	14343	10125	9929	8372
by length 7	12341	9130	9519	6514
by length 8	10599	6686	6040	4020
by length 9	7866	4534	5430	3039
by length 10	4957	3335	3485	2039
Max frq. 3-grams	the 18790	the 12808	the 14829	the 9742
Max frq. 4-grams	that 3026	nthe 1821	ofth 2582	ther 1774

2.2 Arabic texts

The statistical analysis of 4 Arabic texts is described in Table 2. The longest/shortest analyzed text A1/A4 has 94197/31656 words and 395065/135573 symbols. In the text A1, the real words of the length 3 symbols have highest frequency 23287 occurrences and 4-gram – the word *allh* is the real word in Arabic language. In A4, 3-gram – the word *nal* has the highest frequency.
In the texts (A2, A4), 4-gram – the word *fiyal* has the highest frequency. The frequency analysis of all 4 texts is drawn in the Fig. 1, the left panel.

If we compare the graphs in the left panel (Arabic texts) and the right panel (English texts) in Fig. 1, we can see the following differences in the Arabic texts:

- the highest percentage of occurrences is for words of the length 3 and 4,
- the majority of words has the length 1 – 10, shorter than English texts.

The number of 3-grams in Arabic language is $|\Gamma_A|^3 = 28^3$ and 4-grams is 28^4. The decision to use 3- or 4-grams depends now on time complexity only. The analysis what is better for Arabic language is open for us.

In the Fig. 2, the left panel, there is drawn percentage of word occurrences in Arabic text A4 and occurrences of its two disjunctive parts covering the full text (the text was split into two parts). The similar percentage frequencies are in the right panel for English text E3. We illustrate

Table 2: Statistics of 4 Arabic texts, the numbers of words by length for 1-10 and maximal frequencies of 3 and 4-grams.

	Name of texts			
	A1	A2	A3	A4
Total-words	94197	48358	51938	31656
Total-symbols	395065	198019	247448	135573
Total diff. words	14110	9061	25755	10098
# words				
by length 1	6	48	183	81
by length 2	13217	7358	5365	4816
by length 3	**23287**	**12130**	9353*	**7795**
	24.72%	25.08%	18.00%	24.62%
by length 4	22426*	11653*	**9779**	6324 *
	23.80%	24.09%	18.82%	19.97%
by length 5	15887	8336	9175	5417
by length 6	9459	4931	6605	3364
by length 7	5559	2368	4187	2004
by length 8	1978	891	2460	802
by length 9	921	334	973	349
by length 10	336	91	376	182
Arabic	الم	واا	واا	نال
Latin	alm	waa	waa	nal
Max frq.				
3-grams	3027	1797	4242	789
Arabic	الله	فيال	هوهو	فيال
Latin	allh	fiyal	huhu	fiya
Max frq.				
4-grams	1479	525	797	346

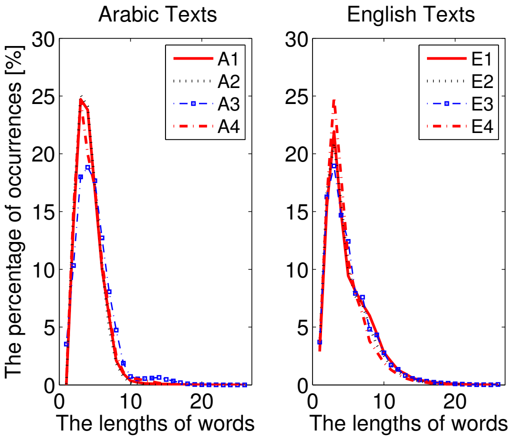

Figure 1: Arabic and English texts. The graphs of normalized word frequencies for 4 texts, the top is for the texts in the length 3. The left panel shows results of Arabic texts. The right panel shows results of English texts.

here that if the text is divided into some parts the percentage of word occurrences should be very similar to the percentage in the full text. It means, the property *frequency of word occurrences* can be used to an evaluation of a similarity/dissimilarity of long text parts.

3 Symbol *n*-gram profile method

The profile of a text will be computed using a dissimilarity measure applied systematically to text parts. Let

- ${}^nT$ – a finite sequence of all n-grams in T prepared according to the text; $|{}^nT| = M - n + 1$ – the number of all n-grams in T;
- $\#o_T^n(g)$ – *the number of occurrences* of n-gram g in the text T;
- $f_T^n(g)$ – *the frequency* of g in the text T defined as
$$f_T^n(g) = \frac{\#o_T^n(g)}{|{}^nT|};$$
- $|{}^nA| = |A| - n + 1$ – the number of all n-grams in the coherent text part A;
- $\#o_A^n(g)$ – *the number of occurrences* of n-gram g in the coherent text part A;
- $f_A^n(g)$ – *the frequency* of g in the coherent text part A defined as
$$f_A^n(g) = \frac{\#o_A^n(g)}{|{}^nA|};$$
- *dissimilarity measure of two text parts A and B* based on n-grams of a text T (not necessary disjunctive parts)
$$d\,(A,B) = \sum_{g\in P(A)} \left[\frac{2\,(f_A^n(g) - f_B^n(g))}{f_A^n(g) + f_B^n(g)}\right]^2 \tag{1}$$
where $f_A^n(g)$ and $f_B^n(g)$ are the frequencies of n-gram g in the text part A and B, respectively, $P(A)$ is the set of all different n-grams in the part A;
- if the numbers of occurrences of g in the two text parts A and B of text are known, a function on n-grams can be defined
$$k_{A,B}^n(g) = \frac{\#o_B^n(g)}{\#o_A^n(g)}, \tag{2}$$
and the formula (1) can be modified as (3)
$$d(A,B) = \sum_{g\in P(A)} \left[\frac{2(|{}^nB| - |{}^nA| * k_{A,B}^n(g))}{|{}^nB| + |{}^nA| * k_{A,B}^n(g)}\right]^2 \tag{3}$$

The method is based on similarity/dissimilarity of the text parts and their occurrences of n-grams in comparison to

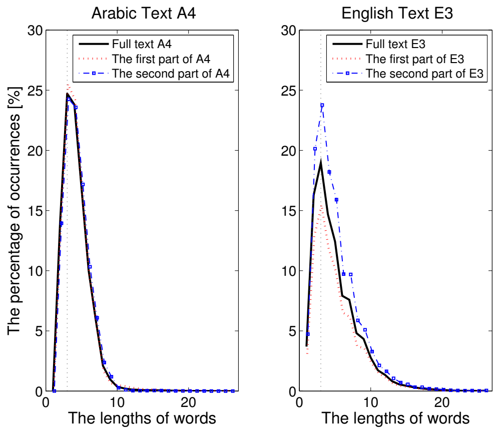

Figure 2: The graphs of normalized word frequencies of one Arabic and English text, the top is for the texts is in the length 3. Both panels show the results of the full text and of the first and the second part. The left panel contains results of Arabic text, the right panel contains results of English text.

the full text T. We modify the dissimilarity measure defined by (3) using (2) to *normalized dissimilarity measure nd* as follows:

$$nd\,(A,T) = \frac{1}{|^nA|} \star \sum_{g\in P(A)} \left[\frac{|^nT| - |^nA| \star k^n_{A,T}(g)}{|^nT| + |^nA| \star k^n_{A,T}(g)}\right]^2 \tag{4}$$

where T is the full text, and $P(A)$ is the set of all n-grams in the text part A, $k^n_{A,T}(g) \geq 1$, A is some coherent part of T. The denominator $|^nA|$ ensures that the values of the dissimilarity function lie between 0 (highest similarity) and 1.

3.1 *n*-gram profile and a style function

We will evaluate a dissimilarity of text parts to the full text using (4). The text part will be a window W moving through the text.

Let W be a sliding window of the length w (in symbols) moving through the text by a step s (in symbols). The window represents a text part and will be moved in each time to the right by s symbols. The *profile of the window W* is defined by the value $nd(W,T)$ (4).

It is possible to define the *style function* of a text T using n-gram profiles of the moving windows as follows:

$$sf(i,T) = nd(W_i,T), i = 1 \ldots \lceil M/s \rceil,$$

where W_i is a window, $\lceil M/s \rceil$ is the total number of windows (it depends on a text length). If $w > s$ the windows are overlapping.

It means, a text part in each window will be evaluated in a comparison to the full text. The size of the window and the step moving should have some influence for a stability of the style function nd. The different results are illustrated in Fig. 3. The figure shows sf function of Arabic text A4, for 4-grams, the length of the moving window was 2000 symbols. In the above panel the moving step was 100 symbols and in the down panel 500 symbols. In Fig. 3, there is drawn the line representing the mean of all $sf(i,T)$ values and the lines of the mean +/- standard deviation.

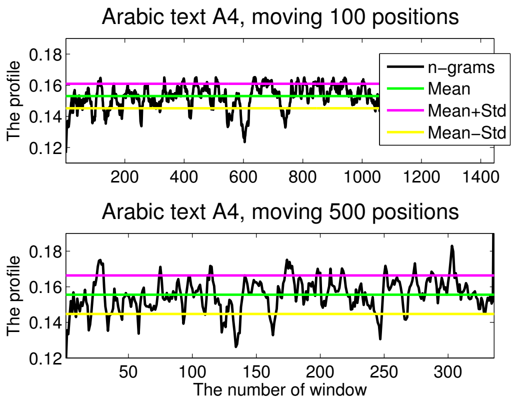

Figure 3: The style function of Arabic text A4, the window of the length 2000 symbols moving by 100 positions using 4-grams (up) and moving by 500 positions (down).

3.2 Algorithm for dissimilarities in a text

We expect that the style function is relatively stable (it does not change value dramatically) if the text is written by the same author. If the style function has very different values (some peaks [5]) for different windows, it is necessary to analyze the covered parts.

Let μ be a mean value of $sf(i,T)$ function values. The existence of some peaks can be indicated by the standard deviation. Let S denote the standard deviation of the style function. If S is lower than a predefined threshold and profile values in the interval $\langle \mu - S, \mu + S \rangle$, then the text parts (windows) look like consistent text of one author. The windows with the profile out of the interval $\langle \mu - S, \mu + S \rangle$ will be analyzed again using the algorithm NGRAM.

The algorithm NGRAM:

1. To compute values of sf function of the text T using setting parameters;
2. To remove all the text windows with the profile out of the interval $\langle \mu - S, \mu + S \rangle$ from sf. These windows correspond probably to the same author. The reduced text in windows is T'.

3. Let $sf(i', T')$ denote the style function after removing above described windows. Let μ' and S' be the mean and standard deviation of $sf(i', T')$.
4. The criterion (5) defines a dissimilarity windows i' as follows [4]:
$$sf(i', T') > \mu' + a * S'. \qquad (5)$$
where parameter a determines the sensitivity of the dissimilarities detection method. For the higher value a, the less number (and more likely problematic) dissimilar windows are detected. The value a is determined empirically (2.0 to attain a good combination of precision and recall [4]). We used $a = 1$.
5. Let $\#dsf$ be the number of windows constructed according to the condition (5). The percentage of dissimilarities can be evaluated by (6)
$$P_{disc} = \frac{100 * \#dsf}{\lceil M/s \rceil} \qquad (6)$$

All steps of the algorithm are illustrated in Fig. 4.

4 Histograms of Words

The text will be analyzed using histograms of words. An introduction to study the method we did in [12], here we show a new final evaluation using distances. Let

- T is divided into a sequence of coherent text parts, $T = \langle T_1, T_2 \dots, T_r \rangle$; $r \geq 1$; $T_i, 1 \leq i \leq r$;
- f^j_T - a frequency of the word $w_j \in V$ in the text T.

If the vocabulary V is alphabetically ordered, then it is possible to do its mapping to integer numbers $1, \dots, |V|$. The texts should be considered as a integer valued time series (the sequences of the numbers of words in the vocabulary V). The analysis using n-gram profiles method do not keep the sequence of the words, it follows occurrences of the subwords in texts. According to [7] the sequences of words can be followed in time using **weighted bag of words (lowbow)**. It can follow a track of changes in histograms connected to words through all text.

The text T has r coherent parts. It is possible to suppose, that each part was written by one author, not necessary different authors. In the method it is analyzed how different the parts are using histograms of words and their dissimilarity.

Histograms will be done in the interval $\langle 0, 1 \rangle$ and it is necessary to map text parts into the interval $\langle 0, 1 \rangle$. *A length-normalized text part* x_{T_i} of the text part $T_i, 1 \leq i \leq r$ is a function $x_{T_i} : \langle 0, 1 \rangle \times V \to \langle 0, 1 \rangle$ such that
$$\sum_{j \in V} x_{T_i}(t, j) = 1, \forall t \in \langle 0, 1 \rangle.$$

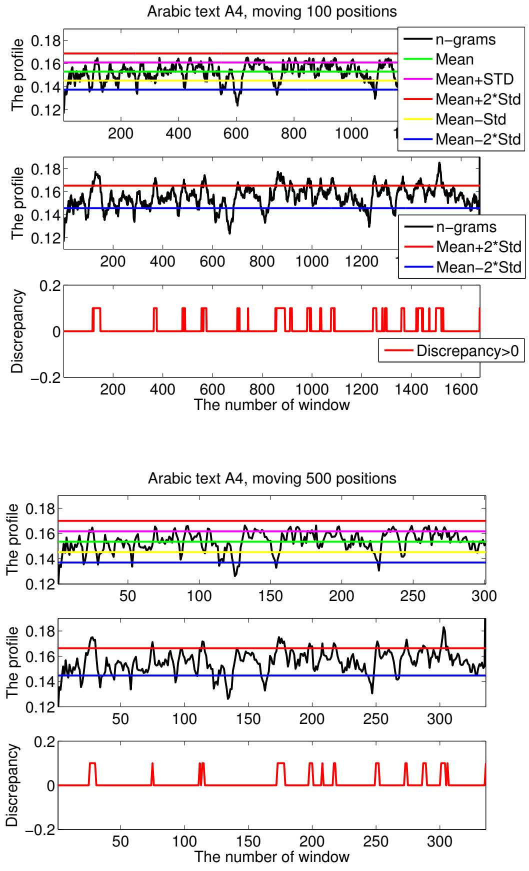

Figure 4: The style function of Arabic text A4, the window of the length 2000 symbols moving by 100 positions using 4-grams (up) and moving by 500 positions (down). The binary function in the down panels indicates windows with a probable dissimilarity (values greater than 0).

If $f^j_{T_i}$ is the frequency of $j \in V$ in the text part T_i then $x_{T_i}(t, j) = f^j_{T_i}/N$ of j in the mapping position t. The mapping is important because of the different lengths of text parts.

The main idea behind the locally weighted bag of words framework [6] is to use a local smoothing kernel to smooth the original word sequence temporally. Our modified algorithm is formulated in the following steps.

The algorithm HISTO:

1. To split the text T into r coherent text parts $T_1, T_2, \dots, T_r, r \geq 1$, N_i is the length of T_i, $1 \leq i \leq r$.
2. To map a text part T_i to the interval $\langle 0, 1 \rangle$.
Let $\mathbf{t} = (t_1, t_2, \dots, t_N) = (1/N, 2/N, \dots, N/N)$ be the vector of values from $\langle 0, 1 \rangle$, $\sum_{j=1}^{N-1}(t_{j+1} - t_j) = 1$. Each t_j will be associated to the word in the position j in the text.

Mapping text of the text part T_i is

$$MT_i(\mathbf{t}) = \langle md_{t_1}, md_{t_2}, \dots md_{t_N} \rangle,$$

where $md_{t_s} = f^j_{T_i}/N$, s is a word in the text part T_i and j is index of word s in the vocabulary V.

3. Let $K^s_{\mu,\sigma}(x) : \langle 0, 1 \rangle \to \mathcal{R}$ be some kernel smoothing function with location parameter $\mu, \mu \in \langle 0, 1 \rangle$ and scale parameter σ. We take k positions of parameter μ, $(\mu_1, \mu_2, \dots, \mu_k)$, such that $\sum_{j=1}^{k} K^s_{\mu,\sigma}(t_j) = 1$.
   It is possible to use Gaussian Probability Density Function (PDF) (7) restricted to the interval $\langle 0, 1 \rangle$ and renormalized

$$\mathcal{N}(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{(2\pi)}} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \qquad (7)$$

   We will use a modification of the function $\mathcal{N}$, the function (8)

$$K^s_{\mu,\sigma}(x) = \begin{cases} \frac{\mathcal{N}(x;\mu,\sigma)}{\phi(1,\mu,\sigma)-\phi(0,\mu,\sigma)}, & \text{if } x \in \langle 0, 1 \rangle, \\ 0, & \text{otherwise.} \end{cases} \qquad (8)$$

   where $\phi(x)$ is a Cumulative Distribution Function

$$\phi(x, \mu, \sigma) = (1 + erf\left(\frac{x-\mu}{\sigma\sqrt{(2)}}\right)),$$

   where $erf(x) = \frac{1}{\sqrt{(pi)}} \int_{-x}^{x} exp(-t^2)dt$. Compute vectors $K^s_{\mu,\sigma}(\mathbf{t})$ for each position $\mu_j, j = 1, \dots, k$ and chosen σ.

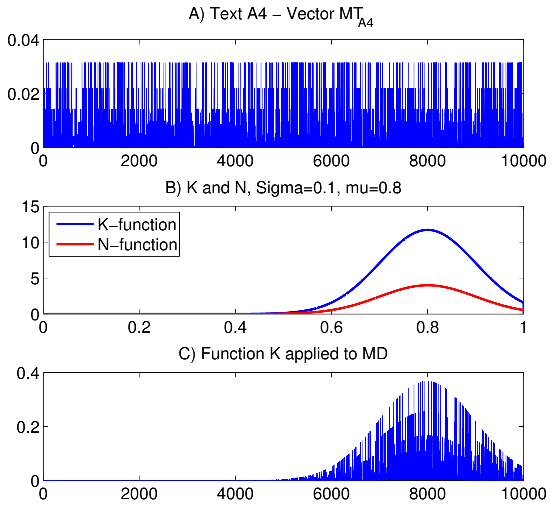

Figure 5: The illustration of the algorithm HISTO on Arabic text A4 (31656 words). In the panel A, there is build the vector MT_{A4} for words in the text. The prepared smooth functions $K^s_{\mu,\sigma}$ and $\mathcal{N}$ are shown in the panel B. The panel C shows the application of function K to vector MT_{A4}.

4. For each text part T_i compute local modified vectors $LH^j_{T_i}$ for each position $\mu_j, j = 1, \dots, k$ as follows:

$$LH^j_{T_i}(\mathbf{t}) = MD(\mathbf{t}) \times K^s_{\mu_j,\sigma}(\mathbf{t}) \qquad (9)$$

   $LH^j_{T_i}$ represents the sequence of histograms usable for some possible analysis of the text part T_i. The steps 1-4 of the algorithm are shown in the Fig. 5.

5. To analyze dissimilarities of the text parts T_{i1} and $T_{i2}, i1 \neq i2, 1 \leq i1, i2 \leq r$ using distances of histograms on words. If a distance of two texts is less than some border parameter $BP_\star$, the texts are similar, a higher value of the distance express a dissimilarity of texts. The values of parameters were prepared in experiments
   Let $H^1 = \{h^1_i\}$ and $H^2 = \{h^2_i\}$ be two histograms of the same length. The used distance functions:
   - *Euclidean distance function*, a border parameter BP_{ED}

$$Dist_{Euc}(H^1, H^2) = \left(\sum_i (h^1_i - h^2_i)^2\right)^{1/2} \qquad (10)$$

   - *Histogram intersection function* - its ability is to handle partial matches when the areas of two histograms are different. A border parameter $BP_{\cap D}$.

$$Dist_{\cap}(H^1, H^2) = 1 - \frac{\sum_i \min\{h^1_i, h^2_i\}}{\sum_i h^2_i} \qquad (11)$$

     If $h^2_i < h^1_i, 1 \leq i \leq |H^1| = |H^2|$ then $Dist_{\cap}(H^1, H^2) = 0$. If differences between histogram values H^1 and H^2 are small then $Dist_{\cap}(H^1, H^2)$ is closed to 0.
   - χ^2 *statistics function* - it measures how unlikely it is that one distribution was drawn from the population represented by the other. A border parameter BP_{SD}.

$$Dist_{\chi^2}(H^1, H^2) = \sum_i \frac{(h^1_i - h^2_i)^2}{2(h^1_i + h^2_i)} \qquad (12)$$

5 Evaluation

We applied both methods - (1) Character n-gram profiles method and (2) Histograms on words - on 40 Arabic and 40 English texts.

We verified usability of both methods to cover dissimilarities in texts. Each text should be divided into some parts and each part can be evaluated in the same way. The

Table 3: Parameter settings used in the experiments.

Description	Symbol	value
Character n-gram length	English	3
Character n-gram length	Arabic	4
Sliding window length	w	1000 / 2000
Sliding window moving	*s*	100 / 500
Sensitivity of detection	a	1

analyzed texts were chosen from benchmark [8] of English texts and benchmark [10, 11] of Arabic texts. In the experiments, it was necessary to prepare parameters. We decided to use the values of parameters given in the Table 3.

5.1 Symbol *n*-gram profiles method

We prepared combined texts from parts of two different texts. It means, the texts were really written by two or more different authors. Our method covered the dissimilarities in such texts. In Fig. 6, it is shown the application of 4-gram profile method on combined Arabic text. The sf' function of the combined text has different shape in the first part and in the first part dissimilarities were identified. The percentage of dissimilarities is P_{diss} = 43.1734%.

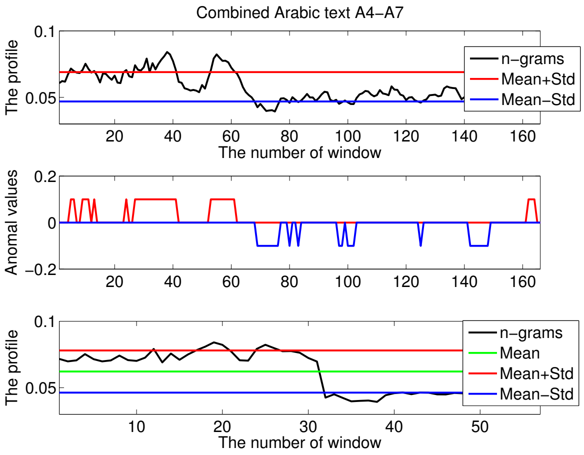

Figure 6: The style function of combined Arabic text A7A4 (75021+37095 symbols), the moving step 100 positions using 4-grams. The binary function (the middle panel) indicates problematic windows. They are analyzed in the down panel. The percentage of dissimilarities is P_{diss} = 43.1734%.

Using parameters in the Table 3 we analyzed the combined texts (it is clear that the texts were written by more authors) given in Table 4, columns 1 and 3. According to the results we recommend to use 35% – 40% borders for the dissimilarity percentage. It means the texts with higher dissimilarity percentage values need to be analyzed by some another method.

Table 4: Results of symbol *n*-gram method for 3 combined Arabic and 3 combined English texts.

Arabic Text	The percentage of dissimilarities	English Text	The percentage of dissimilarities
A07A04	43.1734	E03E06	46.9697
A01A03	45.0704	E13E18	49.1228
A03A08	41.3249	E07E09	41.6667

In the Table 5 we show the results of the percentage dissimilarity for 8 analyzed texts.

Table 5: Results of symbol *n*-gram method for 4 Arabic texts and 4 English texts (the statistics of the texts are in Table 1 and Table 2).

Text	The percentage of dissimilarities	Text	The percentage of dissimilarities
E1	53.3383	A1	31.0743
E2	8.0508	A2	26.2561
E3	43.3132	A3	35.2332
E4	27.9174	A4	34.0771

The values higher than 40% call for some attention to texts, in the presented case texts E1 and E3. In the case of the combined text A7A4, the result 43.1734% express quite good a situation (the text is written by two authors), in the case of English texts E1 and E3 we recommend to do some next analysis.

We analyzed 40 Arabic and 40 English texts from [8, 10]. We found 5 English texts (E1, E3, E19, E24, E30) and 4 Arabic text (A14, A19, A22, A35) with the higher percentage dissimilarity than 40%.

5.2 Histograms of words

In the Fig. 7, histograms of words in text parts of the text A7A4 are drawn for parameters of K function σ = 0.1 and five values of μ = 0.25, 0.4, 0.55, 0.7, 0.85.

If histograms are normalized into same interval then they can be compared according to a dissimilarity function. A comparison of the histogram distances of two text parts shows dissimilarities in the combined texts. Histograms in the Fig. 7 were done on the same intervals and they illustrate that they were done on different text parts. A distance of two histograms in the same positions in two different text parts can be evaluated using developed distances (10, 11, 12).

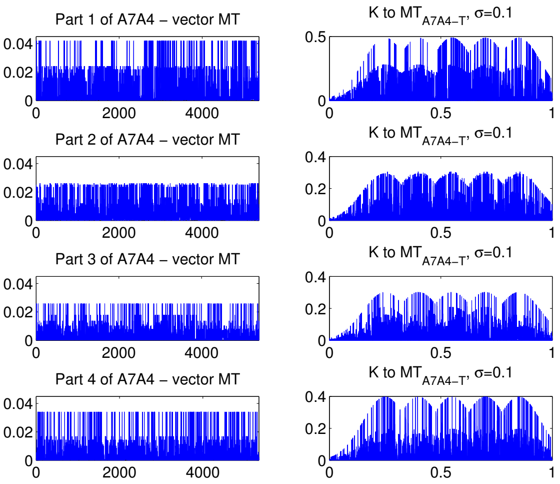

Figure 7: The histograms of four text parts in Arabic text A7A4. The histograms are drawn for $\sigma = 0.1$ and five values of $\mu = 0.25, 0.4, 0.55, 0.7, 0.85$.

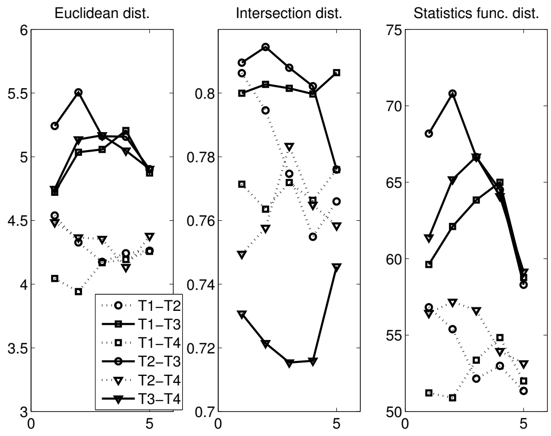

Figure 8: The dissimilarities of the text part histograms for Arabic text A7A4. The text part T1 have higher distance values to the other text parts except intersection distance.

Table 6: Results of histogram distances of four text parts in the Combined Arabic text A7A4.

		Euclidean distance			
A7A4	H1	H2	H3	H4	H5
T1-T2	4.0879	4.3036	5.1475	4.6842	4.4178
T1-T3	3.8826	3.8836	4.9584	4.4012	4.2190
T2-T3	3.2443	3.4931	3.4383	3.5383	3.2439
T1-T4	4.1218	4.0015	5.2247	4.8278	4.6146
T2-T4	3.4153	3.7377	3.8938	4.0560	3.7768
T3-T4	3.0872	2.9601	3.1528	3.3129	3.1432
Euclidean distance - mean value: 3.9423					
Intersection distance					
A7A4	H1	H2	H3	H4	H5
T1-T2	0.6743	0.6997	0.6502	0.6489	0.6704
T1-T3	0.6537	0.6598	0.6000	0.5798	0.6304
T2-T3	0.6562	0.6032	0.5691	0.5872	0.6500
T1-T4	0.6618	0.6641	0.6376	0.6652	0.7036
T2-T4	0.6581	0.6432	0.6608	0.6837	0.7256
T3-T4	0.7252	0.7091	0.7214	0.7581	0.7861
Intersection distance - mean value: 0.6645					
Statistics function distance					
A7A4	H1	H2	H3	H4	H5
T1-T2	46.4776	51.0914	58.3920	55.1670	46.0019
T1-T3	42.8903	44.2577	52.7783	49.0519	41.5391
T2-T3	37.6820	39.4750	39.5410	40.1174	34.3253
T1-T4	43.4616	44.5446	55.3728	54.4515	46.9147
T2-T4	37.7778	42.2563	45.5686	46.3336	40.3130
T3-T4	32.1282	31.5597	33.0406	32.9503	29.6607
Statistics function distance - mean value: 43.1707					

The results for the text A7A4 are shown in Table 6 and Fig. 8. The results show that the distances of the text part T1 to the other text parts have higher values than distances between the residual text parts. In the first step we recommend to use mean values of distances as border parameter values, $BP_{ED} = 3.9423, BP_{\cap DE} = 0.6645$ and $BP_{SD} = 43.1707$. The distances of the text part T1 to the other parts are higher than border parameter values for Euclidean and statistics distances mainly. The values of the intersection distance are very closed to the border parameter value.

The results for the text A4 are in Table 7 and Fig. 9. The text part T3 has higher distance values than the border parameter values in the case of Euclidean and statistics function distance, it means the text A4 should be analyzed again (according to the first method the text A4 has the percentage of dissimilarities 34.0771%).

6 Conclusion

In this paper, two methods for computation of texts characteristics were developed. The algorithms evaluate a dissimilarity of texts and they should be used for some clas-

Table 7: Results of histogram distances of four text parts in the Arabic text A4.

Euclidean distance					
A4	H1	H2	H3	H4	H5
T1-T2	4.5386	4.3280	4.1756	4.2428	4.2627
T1-T3	4.7217	5.0358	5.0575	5.2058	4.8726
T2-T3	5.2419	5.5057	5.1615	5.1572	4.9024
T1-T4	4.0451	3.9419	4.1687	4.1936	4.2578
T2-T4	4.4856	4.3660	4.3544	4.1344	4.3785
T3-T4	4.7468	5.1360	5.1681	5.0481	4.9043
Euclidean distance - mean value: 4.6580					
Intersection distance					
A4	H1	H2	H3	H4	H5
T1-T2	0.8063	0.7946	0.7747	0.7549	0.7660
T1-T3	0.7999	0.8027	0.8015	0.7997	0.8064
T2-T3	0.8095	0.8145	0.8079	0.8022	0.7760
T1-T4	0.7714	0.7635	0.7719	0.7663	0.7759
T2-T4	0.7495	0.7577	0.7834	0.7649	0.7584
T3-T4	0.7307	0.7215	0.7154	0.7160	0.7456
Intersection distance - mean value: 0.7736					
Statistics function distance					
A4	H1	H2	H3	H4	H5
T1-T2	56.8219	55.3926	52.1574	52.9932	51.3536
T1-T3	59.6212	62.1068	63.8350	65.0018	58.7666
T2-T3	68.1809	70.8000	66.6500	64.4992	58.2916
T1-T4	51.2219	50.9001	53.3641	54.8419	52.0024
T2-T4	56.4430	57.1799	56.6264	53.9514	53.1499
T3-T4	61.3877	65.1853	66.6711	64.0820	59.1484
Statistics function distance - mean value: 58.7542					

sification of texts and an analysis of authorship detection. Our problem of text part dissimilarity needs to analyze if all parts of the text were written by the same author (or as it was declared by authors). A combination of both methods should be more successful in our research. The first method was applied for 40 Arabic and 40 English texts from the corpus [10] and [8] and in an evaluation of texts to use some classification methods. The result is 25% of English and 20% Arabic texts belong to texts with higher values of dissimilarity percentage. The second method was applied to 10 Arabic and 10 English texts. The results support the evaluation of texts according to the first method. Our research will continue in a searching of a new distance functions and their weighted combinations to get a better evaluation of texts.

Acknowledgement: The research is supported by the Slovak Scientific Grant Agency VEGA, Grant No. 1/0142/15. We thank to Bc. Peter Sedmák from P. J. Šafárik University in Košice for his help in a programming in Java.

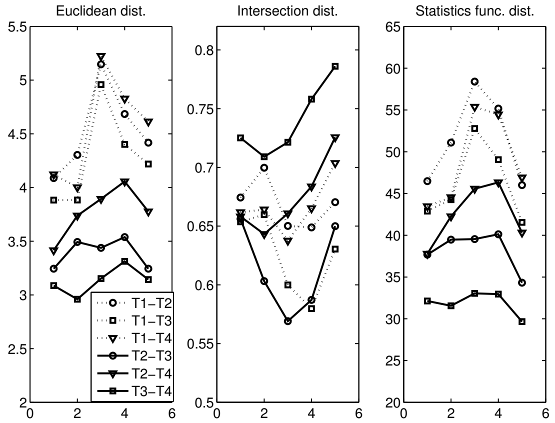

Figure 9: The dissimilarities of the text part histograms for Arabic text A4. The histograms of words in text part T3 have bigger distances to the other text parts.

References

[1] Meyer zu Eissen S., Stein B., Intrinsic Plagiarism detection, In: Lalmas et al. (Eds.), Proceedings of the 28th ECIR (10-12 April 2006, London, UK), Springer, 2006, 565-569

[2] Haj Hassan F. I., Chaurasia M. A., N-Gram Based Text Author Verification,
Proceedings of the ICIIM (7-8 January 2012, Chengdu, China), IACSIT press, 2012 , 36, 67–71

[3] Kuta M., Kitowki, J., Optimisation of Character n-gram Profiles Method for Intrinsic Plagiarism Detection, In: L. Rutkowski et al. (Eds.), Proceedings of 13th ICAISC (1-5 June 2014, Zakopane, Poland), Springer, 2014, 500-511

[4] Stamatatos E., Intrinsic Plagiarism Detection Using Character *n*-gram Profiles, In: G. Sidorov, A. H. Aguirre, C. A. R. Garcia (Eds.), Proceedings of the 3rd PAN Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (September 2009, Donostia-San Sebastian, Spain), Springer, 2009, 38-46

[5] Stamatatos, E., A survey of modern authorship attribution methods, J. Am. Soc. Inf. Sci. Technol, 2010, 538-556

[6] Escalante H. J., Solorio T., Montes-y-Gomez M., Local Histograms of Character n-grams for Authorship Attribution, In: F. Castro, A. Gelbukh, M. Gonzales (Eds.), Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (19-24 June 2011, Portland, USA), Springer, 2011, 288-298

[7] Lebanon G., Mao Y., Dillon J., The locally weighted bag of words framework for text representation, JMLR, 2007, 8, 2405-2441

[8] Corpus of English texts: (PAN-PC-11), http://www.uniweimar.de/en/media/chairs/webis/corpora/pan-pc-11/

[9] Meyer zu Eissen S., Stein B., Kulig M., Plagiarism detection without reference collections, In. Advances in Data Analysis, Proceedings of the 30th Annual Conference of the German Classification Society (2006, Berlin, Germany), Springer, 2006, 359 - 366

[10] King Saud University Corpus of Classical Arabic Texts.
http://ksucorpus.ksu.edu.sa

[11] Bensalem I., Rosso P., Chikhi S., A New Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection, In: P.Forner et al. (Eds.), Proceedings of the 4th International Conference of the CLEF Initiative, (23-26 September 2013, Valencia, Spain), Springer, 2013, 53-58

[12] Almarimi A., Andrejková G., Text Anomalies Detection Using Histograms of Words, ACSIJ, 2016, 19, 63-68