Statistical tests for text homogeneity: using forward and backward processes of numbers of different words

Berhane Abebe[1,2], Mikhail Chebunin[3,2], Artyom Kovalevskii[4,2*], Natalia Zakrevskaya[4]

[1] Mainefhi College of Science, Mainefhi, Eritrea
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Karlsruhe Institute of Technology, Institute of Stochastics, Karlsruhe, Germany
[4] Novosibirsk State Technical University, Novosibirsk, Russia
[*] Corresponding author's email: artyom.kovalevskii@gmail.com

DOI: https://doi.org/10.53482/2022_53_401

ABSTRACT

The processes of growth in the number of diverse words in a text, when reading in the forward and backward directions, are studied in this article. Based upon the statistics achieved from the difference between these two processes, we construct a statistical test. This statistical test is used for text homogeneity checks. The elementary model states that words in a text are selected from some dictionary independent of each other according to the Zipf–Mandelbrot law. P-values of the statistical test are calculated based on the elementary probabilistic model using the asymptotic normality of corresponding statistics. At last but not least, this statistical test is applied for the analysis of homogeneity of sequences of sonnets.

Keywords: Zipf's law, weak convergence, Gaussian process, statistical test, text homogeneity, urn model.

1 Introduction

The motivation behind this work was the procedure for writing essays by students in the Internet era: a student's essay sometimes is simply a combination of two or more texts found using a search engine. As a result, we cannot determine the student's intellectual contribution. Therefore, we need an algorithm that allows us to identify the presence of heterogeneous fragments in a text. Our models and methods are completely probabilistic.

We calculate R_k the number of different words among the first k words, and R'_k the number of different words among the last k ones. We let $R_0 = R'_0 = 0$.

Let us illustrate forward and backward processes of numbers of different words by the following simple example. All words have been converted to its lowercase letters and punctuation marks are excluded from the text. Then we calculate sequential numbers of different words R_k in the forward direction. Then we read the words in the backward direction and calculate corresponding sequential numbers of different words R'_k. The text under analysis is: 'Hamlet: To be, or not to be:'. The results are in Table 1.

Table 1: Forward and backward readings of the text: 'Hamlet: To be, or not to be:'.

Forward reading								
k	0	1	2	3	4	5	6	7
words	-	hamlet	to	be	or	not	to	be
R_k	0	1	2	3	4	5	5	5
Backward reading								
k	0	1	2	3	4	5	6	7
words	-	be	to	not	or	be	to	hamlet
R'_k	0	1	2	3	4	4	4	5

We construct a theoretical background for the analysis of forward and backward processes in Section 2. We describe our material (sequences of sonnets) in Section 3 and analyse it in Section 4. Conclusion is in Section 5. Appendix contains proofs.

2 Theoretical background

An infinite urn scheme is the theoretical model for these text statistics. Words are chosen from a countably infinite dictionary one by one independent of each other and are numbered $1, 2, \ldots$. Let X_i be the number of the word at ith position, where $1 \le i \le n$,

$$\mathbf{P}(X_i = j) = p_j > 0, \quad j \ge 1, \quad p_1 + p_2 + \ldots = 1, \quad p_1 \ge p_2 \ge \ldots .$$

Bahadur (1960) proved the next results for R_n under these assumptions:

$$\mathbf{E}R_n = \sum_{i=1}^{\infty}(1-(1-p_i)^n), \quad \mathbf{Var}R_n \le \mathbf{E}R_n,$$

$$\mathbf{E}R_n \to \infty, \quad \mathbf{E}R_n/n \to 0$$

Karlin (1975) proved the Strong Law of Large Numbers (SLLN):

$$R_n/\mathbf{E}R_n \overset{a.s.}{\to} 1.$$

The next Regularity Condition plays an important role in the following:

$$\kappa(x) := \max\{k > 0 : \ p_k \ge 1/x\} = x^{\theta}L(x), \quad 0 < \theta < 1,$$

$L(\cdot)$ is the slowly varying function of the real argument: $L(tx)/L(x) \to 1$ as $x \to +\infty$ for any real $t > 0$.

Equivalent condition in terms of word probabilities is:

$$p_i = i^{-1/\theta} l(i),$$

$l(\cdot)$ is the another slowly varying function.

The model is an elementary probabilistic model that generalises the Zipf's Law (Zipf, 1936) of power decreasing of word probabilities.

Slowly varying functions include all the functions that have a finite positive limit at infinity, so the Regularity Condition includes the case of the very general form of Zipf's Law that correspond to formula (2) in Ferrer i Cancho and Solé (2001). But the Regularity Condition permits a more wide class of probability distributions that can appear in different generalizations of the Zipf's Law.

Karlin (1975) proved that if the regularity condition holds then $(R_n - \mathbf{E}R_n)/\sqrt{\mathbf{Var}R_n}$ converges weakly to the standard normal distribution,

$$\mathbf{E}R_n \sim \Gamma(1-\theta)\kappa(n), \quad \mathbf{Var}R_n/\mathbf{E}R_n \to 2^\theta - 1,$$

$\Gamma(\cdot)$ is the Euler gamma.

So $(R_n - \mathbf{E}R_n)/\sqrt{\mathbf{E}R_n}$ converges weakly to the centered normal distribution with variance $2^\theta - 1$.

Chebunin and Kovalevskii (2016) proved that there is convergence of the centered and normalized process of numbers of different words,

$$Z_n = \{Z_n(t),\ 0 \le t \le 1\} = \{(R_{[nt]} - \mathbf{E}R_{[nt]})/\sqrt{\mathbf{E}R_n},\ 0 \le t \le 1\}$$

converges weakly in $D(0,1)$ with uniform metrics to a centered Gaussian process Z_θ with continuous a.s. sample paths and covariance function

$$K(s,t) = (s+t)^\theta - \max(s^\theta, t^\theta).$$

We introduce a process in backward time

$$Z'_n = \{Z'_n(t),\ 0 \le t \le 1\} = \{(R'_{[nt]} - \mathbf{E}R'_{[nt]})/\sqrt{\mathbf{E}R'_n},\ 0 \le t \le 1\}.$$

The next theorem take place.

Theorem 1 (for joint distribution)

If the Regularity Condition holds then $(Z_n, Z'_n) = \left\{(Z_n(t), Z'_n(t)),\ 0 \le t \le 1\right\}$ *converges weakly in the uniform metrics in* $D(0,1)^2$ *to a 2-dimensional Gaussian process* (Z, Z') *with zero expectation and covariance function*

$$\mathbf{E}Z(s)Z(t) = \mathbf{E}Z'(s)Z'(t) = K(s,t), \quad \mathbf{E}Z(s)Z'(t) = K'(s,t),$$

$$K'(s,t) = ((s+t)^{\theta} - 1)\mathbf{1}(s+t > 1).$$

Corollary 1 (for the difference of processes) *If the Regularity Condition holds then*

$$J_n = \frac{\sum_{k=1}^{n}(R_k - R'_k)}{n\sqrt{R_n}}$$

converges weakly to a centered normal random variable with variance $\frac{\theta}{\theta+2}$.

The proofs of Theorem 1 and Corollary 1 are given in the Appendix.

Corollary 1 gives the opportunity to test the homogeneity of a text using any consistent estimate θ^* of parameter θ. The p-value is calculated using the tail of the standard normal distribution and the observed value J_{obs} of J_n:

$$\text{p-value} = 2\overline{\Phi}\left(|J_{obs}|\sqrt{1+2/\theta^*}\right).$$

In order to use Corollary 1 in applications, we need some estimate of the unknown parameter θ. Various classes of such estimates have been obtained and analysed by (Chakrabarty et al., 2020; Guillou and Hall, 2001; Hill, 1975; Nicholls, 1987; Ohannessian and Dahleh, 2012).

But we need an estimate that is symmetric to the forward and backward processes. Hence, we introduce a new estimate and study its properties.

From the SLLN, we have $\log R_n \sim \theta \log n$ a.s. Therefore, we may propose the following estimates for the parameter θ:

$$\theta_n = \log_2 \frac{R_n}{R_{\lfloor n/2 \rfloor}}, \quad \theta'_n = \log_2 \frac{R'_n}{R'_{\lfloor n/2 \rfloor}}.$$

Note that $R_n = R'_n$. Let

$$\widehat{\theta}_n = (\theta_n + \theta'_n)/2.$$

Then

$$\widehat{\theta}_n = \log_2\left(R_n/\sqrt{R_{\lfloor n/2 \rfloor}R'_{\lfloor n/2 \rfloor}}\right), \quad n \geq 2.$$

All these estimates are consistent due to SLLN.

3 Material

We used 3 sequences of sonnets from open sources: Thomas Wyatt (32 sonnets), 1542; William Shakespeare (154 sonnets), 1609; Charlotte Smith, Elegiac sonnets (sonnets I - LIX), 1784.

Sonnets by sir Thomas Wyatt (31 sonnets, the last sonnet in two parts) and sonnets by William Shakespeare (154 sonnets) can be found at https://shakespeares-sonnets.com/

Elegiac sonnets by Charlotte Smith can be found at

https://quod.lib.umich.edu/e/evans/N22357.0001.001?rgn=main;view=fulltext

4 Analysis of sonnets

We use the algorithm from Section 2 to analyse sequences of sonnets.

Shakespeare's sonnets contain $n = 17516$ words, $R_n = 3258$ different words. The graph of rank verses frequency of words in Shakespeare's sonnets is given in Figure 1, and its Zipfian diagram is in Figure 2. Table 2 contains the most frequent words.

Table 2: First 24 tokens in Shakespeare's sonnets.

Token	Rank	Frequency
and	1	489
the	2	444
to	3	409
of	4	371
my	5	364
i	6	341
in	7	322
that	8	320
thy	9	266
thou	10	234
with	11	181
for	12	171
is	13	169
not	14	167
but	15	164
me	16	164
a	17	163
thee	18	162
love	19	160
so	20	145
be	21	141
as	22	121
all	23	117
you	24	110

The process of numbers of different words in Shakespeare's sonnets (Heaps' diagram) in forward direction is drawn in Figure 3.

We know that

$$\mathbf{E}R_k = \sum_{i=1}^{\infty}(1-(1-p_i)^k).$$

We can estimate the unknown expectation by

$$\widetilde{R}_k = \sum_{i=1}^{R_n}(1-(1-p_i^*)^k)$$

with

$$p_i^* = n_i/n,$$

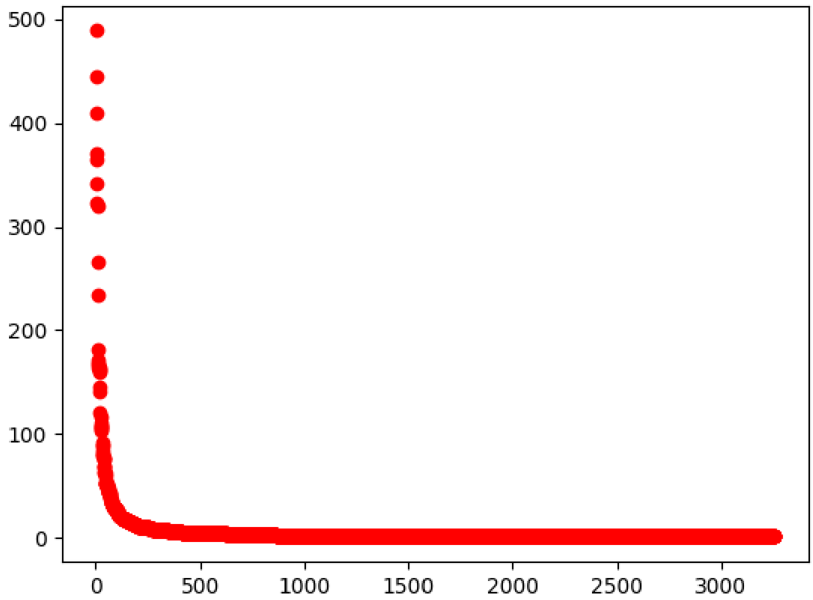

Figure 1: Frequencies of words in Shakespeare's sonnets.

n_i be the number of occurrences of a word with rank i.

Figure 4 illustrates the poor fitness of this approximation. The cause of this poor fitness is that the sum does not approach to infinity, but rather to R_n only. We have not estimates of probabilities for ranks greater than R_n. So we need the regularity conditions to estimate all the probabilities to infinity.

The Mandelbrot (1965) modification of the Zipf's Law is:

$$p_i = c(i+q)^{-1/\theta}, \;\; i \geq 1, \;\; 0 < \theta < 1, \;\; q > -1.$$

Here

$$c = (\zeta(1/\theta, q+1))^{-1},$$

$$\zeta(\alpha, x) = \sum_{i=0}^{\infty} (i+x)^{-\alpha}$$

is the Hurvitz zeta function.

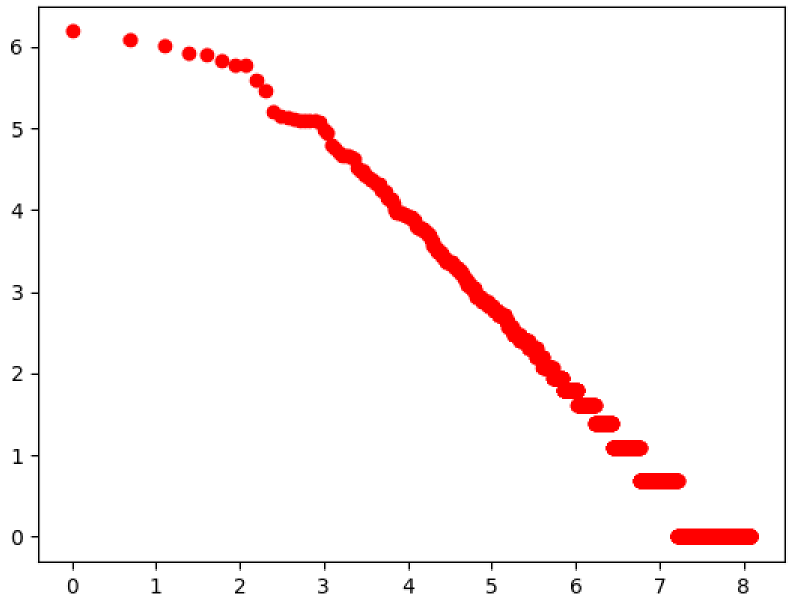

Figure 2: Logarithms of frequencies of words to logarithms of ranks in Shakespeare's sonnets (Zipfian diagram).

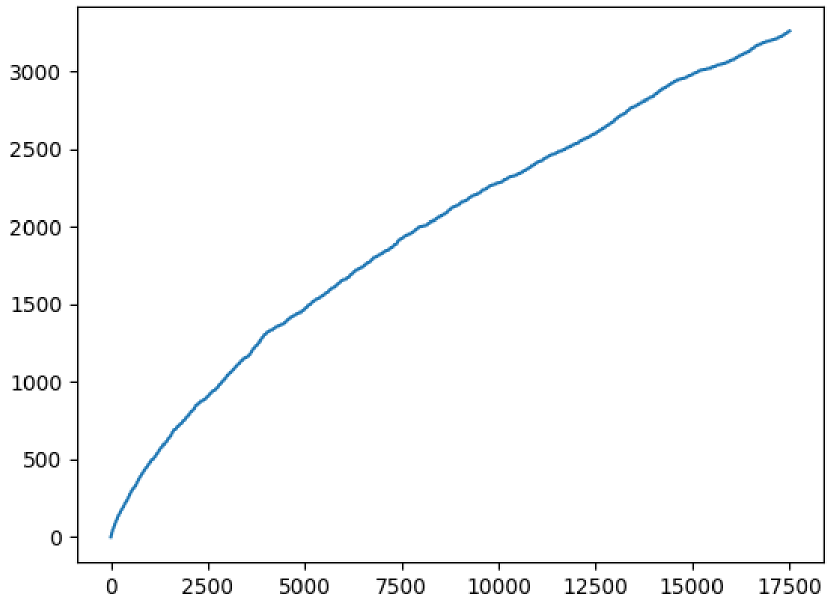

Figure 3: The process of numbers of different words in Shakespeare's sonnets (Heaps' diagram).

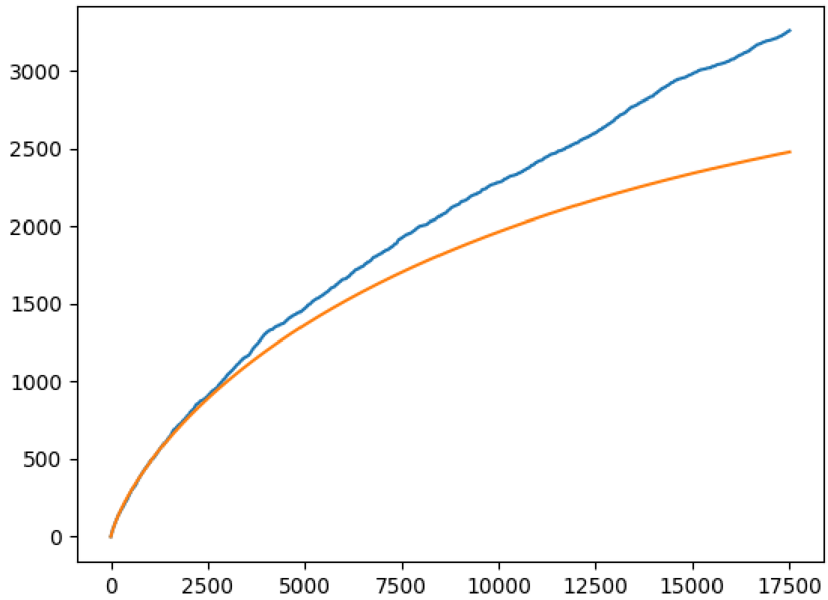

Figure 4: The process of R_k with its empirical approximation $\widetilde{R}_k$.

We use approximation

$$r(k) = \sum_{i=1}^{\infty} \left(1 - (1 - \widehat{p}_i)^k\right), \quad 0 \le k \le n.$$

Here

$$\widehat{p}_i = c(i + q_n)^{-1/\theta_n}, \quad i \ge 1,$$

q_n is such that $r(n) = R_n$.

Figure 5 illustrates the goodness of this approximation.

Figures 6–8 demonstrate forward and backward processes of numbers of different words for Wyatt's, Shakespeare's and Smith's sonnets. Statistical tests confirm good homogeneity, see Table 1. We use statistics J_n with known limiting distribution and the statistics

$$\omega_n^2 = \int_0^1 (Z_n(t) - Z_n'(t))^2 dt.$$

This statistics converges weakly to some limiting distribution but its cdf is complicated. So we do not calculate p-values for it.

Figures 9–14 demonstrate forward and backward processes of numbers of different words for all variants of concatenations of sonnets of these 3 authors. There is obvious non-homogeneity, and p-values are

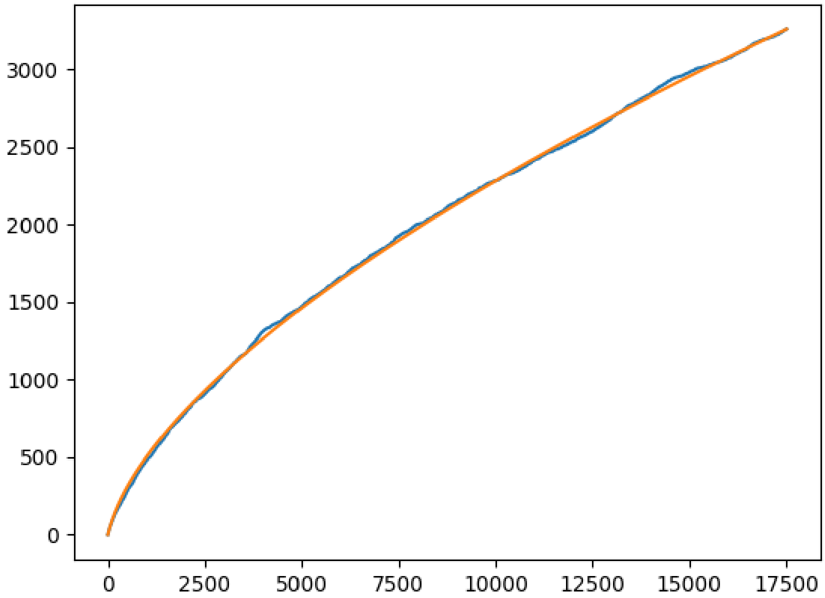

Figure 5: The forward process of numbers of different words for Shakespeare's sonnets and its approximation $r(k)$ with $\theta_n = 0.6267$, $q_n = 46.39$.

lesser. J_n give relatively big p-values for concatenations of Smith and Wyatt but ω_n^2 precisely differ for texts from one author ($\omega_n^2 \leq 0.883$) and for concatenations ($\omega_n^2 \geq 3.7753$). All the results of calculations are in Table 3.

Table 3: Sequences of sonnets and its concatenations.

Author(s)	J_n	θ_n	θ'_n	$\widehat{\theta}_n$	p-value	ω_n^2
Wyatt	-0.1139	0.7556	0.7459	0.7507	0.8275	0.0681
Shakespeare	0.2939	0.6267	0.6274	0.6271	0.5475	0.3868
Smith	-0.8748	0.6788	0.62	0.6494	0.0772	0.883
Shakespeare+Wyatt	-3.7886	0.8082	0.5634	0.6858	0.0000	20.3048
Wyatt+Shakespeare	4.2126	0.5837	0.7948	0.6893	0.0000	22.4295
Smith+Shakespeare	4.6056	0.552	0.7925	0.6723	0.0000	27.3113
Shakespeare+Smith	-4.8183	0.8146	0.5444	0.6795	0.0000	28.7613
Smith+Wyatt	-0.5909	0.8108	0.7256	0.7682	0.2620	4.5616
Wyatt+Smith	-0.4583	0.7627	0.7924	0.7775	0.3863	3.7753

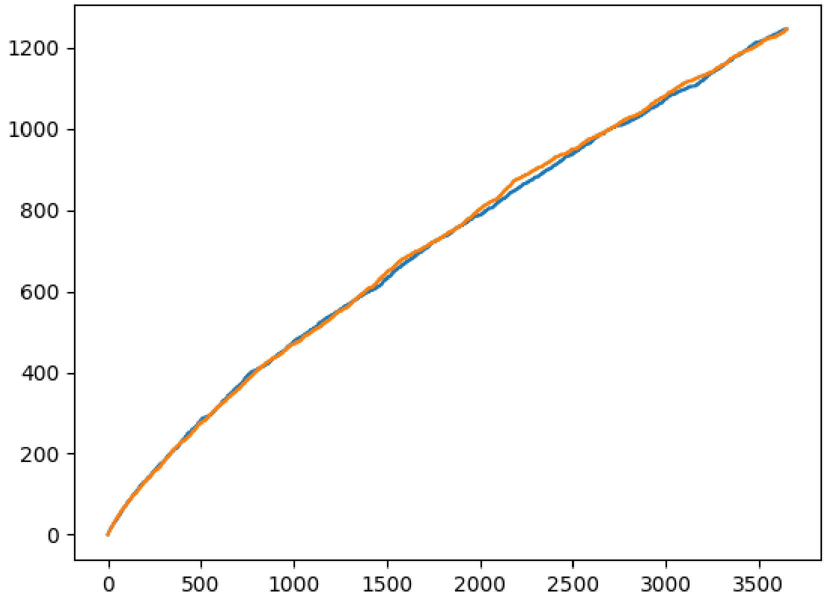

Figure 6: Forward and backward processes of numbers of different words for Wyatt's sonnets.

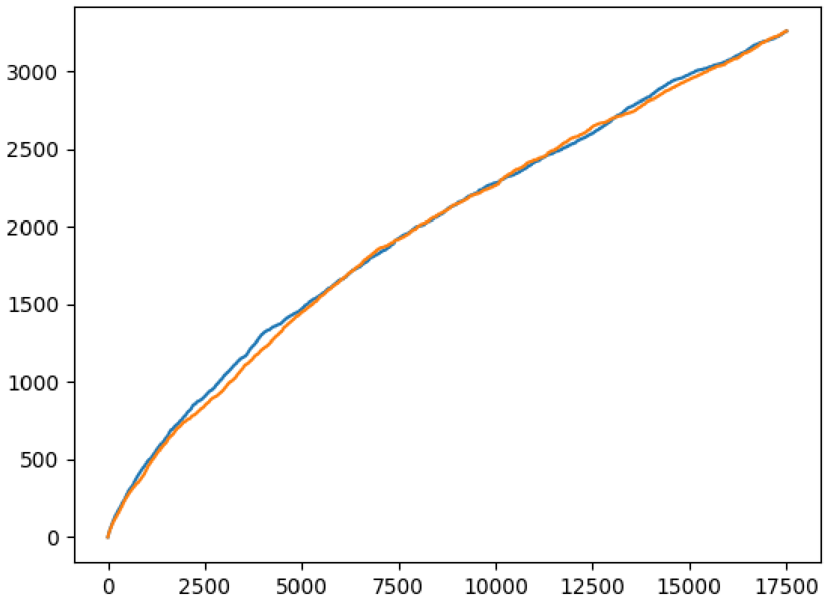

Figure 7: Forward and backward processes of numbers of different words for Shakespeare's sonnets.

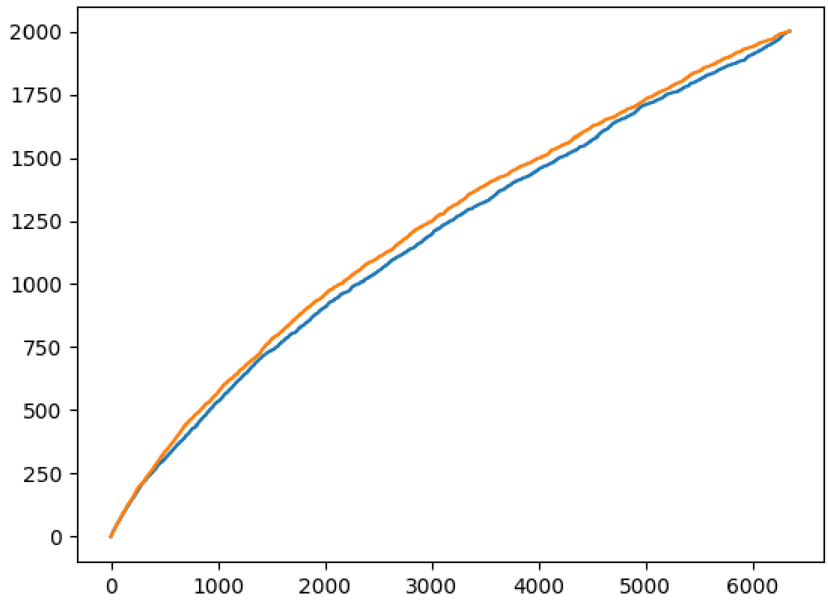

Figure 8: Forward and backward processes of numbers of different words for Smith's sonnets.

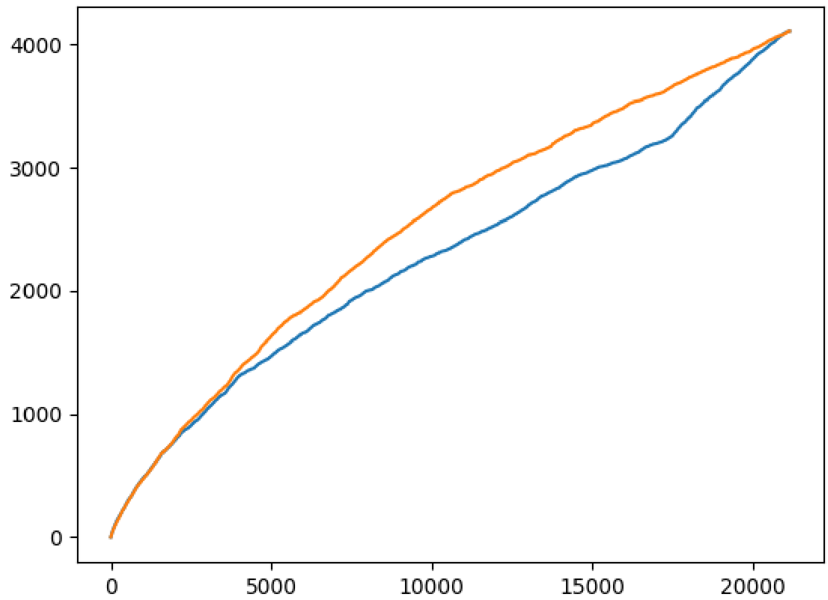

Figure 9: Forward and backward processes of numbers of different words for Shakespeare+Wyatt.

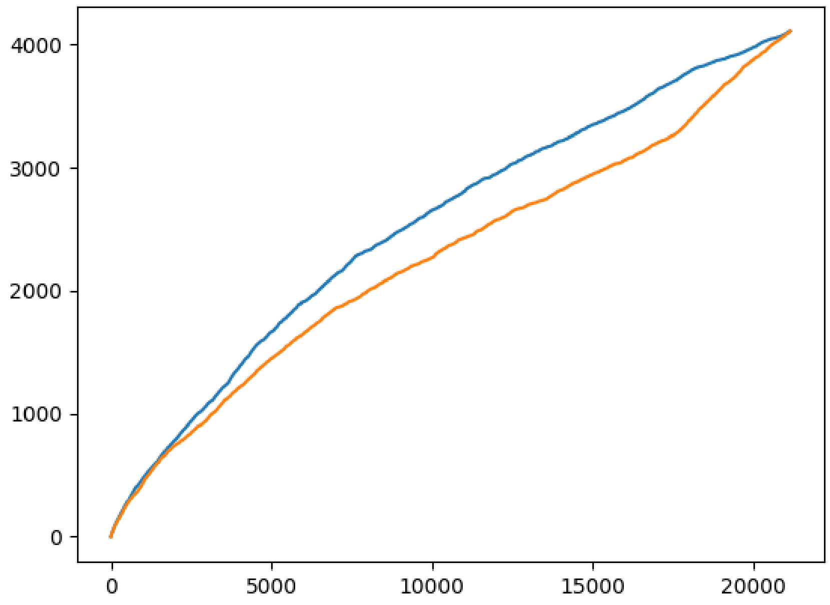

Figure 10: Forward and backward processes of numbers of different words for Wyatt+Shakespeare.

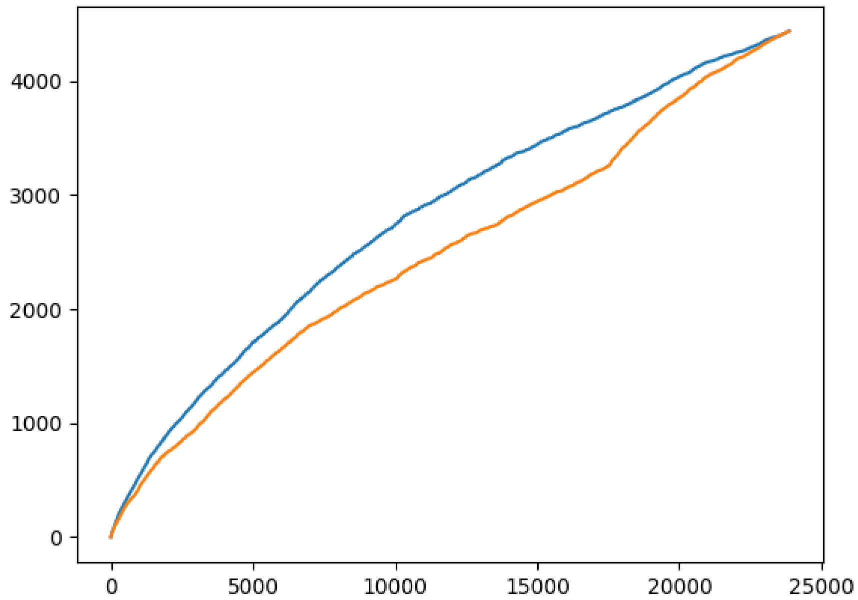

Figure 11: Forward and backward processes of numbers of different words for Smith+Shakespeare.

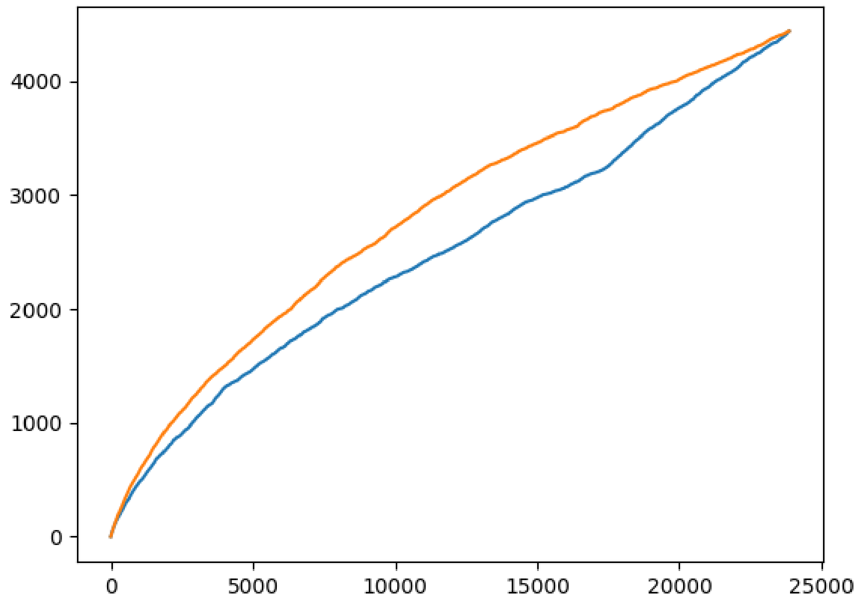

Figure 12: Forward and backward processes of numbers of different words for Shakespeare+Smith.

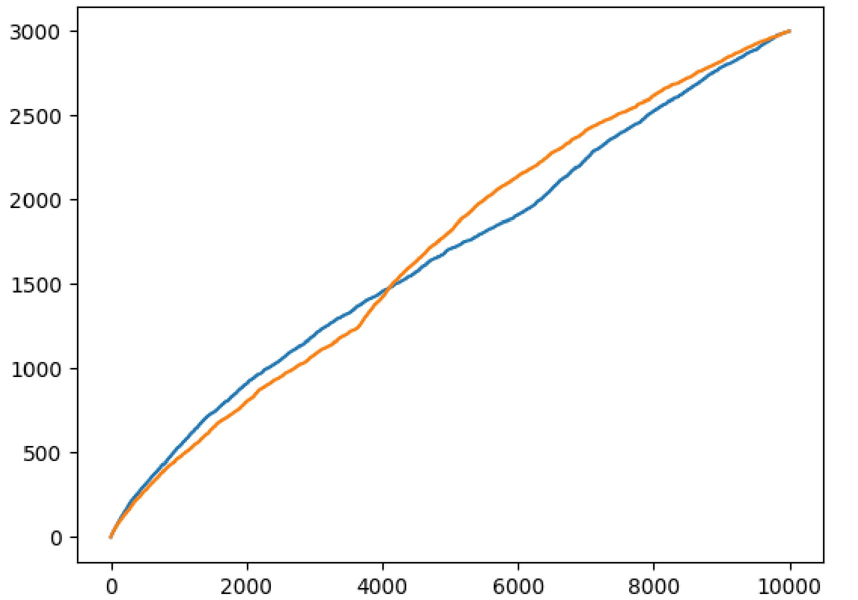

Figure 13: Forward and backward processes of numbers of different words for Smith+Wyatt.

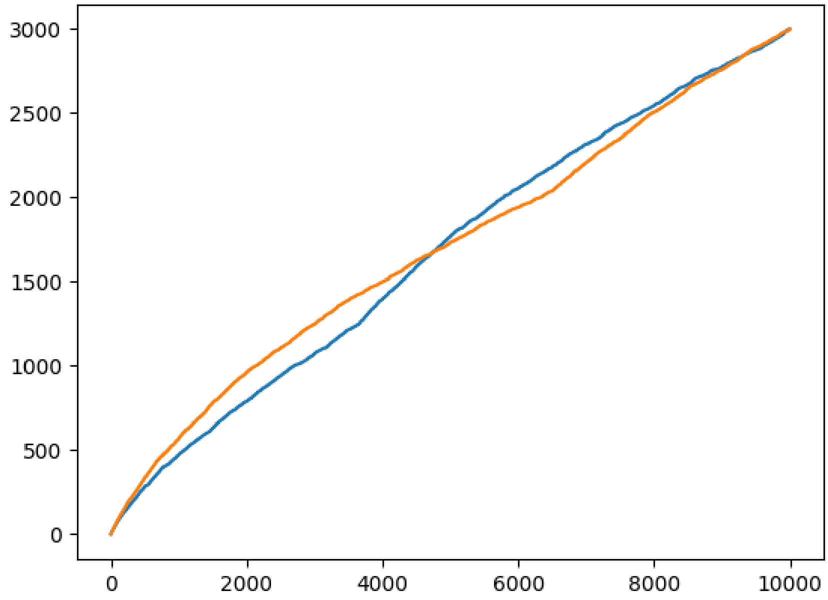

Figure 14: Forward and backward processes of numbers of different words for Wyatt+Smith.

5 Conclusion

We have invented a new class of statistical tests for analysis the homogeneity of texts. These tests can be applied to texts in many languages. They use the appearance of new words when reading the text in the forward and reverse order. The above examples show that the sequences of texts of one author are recognized as homogeneous, and of two authors as heterogeneous.

We use two statistics, that is, two functionals of the process under study for both directions. There are many other functionals, and their comparison in terms of the effectiveness of testing the hypothesis of homogeneity is the subject of further study.

The tests work correctly, only if a text under analysis is relatively long. If someone want to deal with short texts then he (she) may need to employ some correction of p-values that can be calculated by simulation of texts using the elementary probability model.

Another interesting problem is division of a text into homogeneous fragments using the process under study. We plan to develop the algorithm by proving the corresponding theorems. Then the algorithm will be implemented and tested on real mixed texts.

From the point of view of plagiarism analysis, concatenation of texts from several sources using machine translation and minimal editing is just the simplest way to create this kind of essay. More complex ways of creating plagiarism, involving careful mixing of texts from many different sources, require the development of more accurate methods of statistical analysis.

Acknowledgments

The work is supported by Mathematical Center in Akademgorodok under agreement No. 075-15-2019-1675 with the Ministry of Science and Higher Education of the Russian Federation.

The authors like to thank an anonymous referee for helpful and constructive comments and suggestions.

References

Bahadur, R. (1960). On the number of distinct values in a large sample from an infinite discrete distribution. *Proceedings of the National Institute of Sciences of India*, *26*(A), pp. 67–75.

Chakrabarty, A., Chebunin, M., Kovalevskii, A., Pupyshev, I., Zakrevskaya, N., Zhou, Q. (2020). A statistical test for correspondence of texts to the Zipf-Mandelbrot law. *Siberian Electronic Mathematical Reports*, *17*, pp. 1959–1974. https://doi.org/10.33048/semi.2020.17.13

Chebunin, M., Kovalevskii, A. (2016). Functional central limit theorems for certain statistics in an infinite urn scheme. *Statistics and Probability Letters*, *119*, pp. 344–348. https://doi.org/10.1016/j.spl.2016.08.1099

Ferrer i Cancho, R., Solé, R. V. (2001). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, *8*(3), pp. 165–173. https://doi.org/10.1076/jqul.8.3.165.4101

Guillou, A., Hall, P. (2001). A diagnostic for selecting the threshold in extreme value analysis. *Journal of the Royal Statistical Society: Series B*, *63*(2), pp. 293–305. https://doi.org/10.1111/1467-9868.00286

Hill, B. M. (1975). A Simple General Approach to Inference About the Tail of a Distribution. *The Annals of Statistics*, *3*(5), pp. 1163–1174. https://doi.org/10.1214/aos/1176343247

Karlin, S. (1975). Central Limit Theorems for Certain Infinite Urn Schemes. *Journal of Mathematics and Mechanics*, *17*(4), pp. 373–401. https://doi.org/10.1512/iumj.1968.17.17020

Mandelbrot, B. (1965). Information Theory and Psycholinguistics. In: Wolman, B. B. (Ed.). *Scientific psychology*. Basic Books.

Nicholls, P. (1987). Estimation of Zipf parameters. *Journal of the American Society for Information Science*, *38*(4), pp. 443–445. https://doi.org/10.1002/ (SICI)1097-4571(198711)38:6<443::AID-ASI4>3.0.CO;2-E

Ohannessian, M., Dahleh, M. (2012). Rare probability estimation under regularly varying heavy tails. In: *Proceedings of the 25th Annual Conference on Learning Theory*. PMLR.

Zipf, G. K. (1936). *The Psycho-Biology of Language*. Routledge.

Appendix. Proofs

Proof of Theorem 1

Let $\mathbb{X}_i(n)$ be a number of balls in urn i. Let $\Pi = \{\Pi(t), t \geq 0\}$ be a Poisson process with parameter 1. The Poissonized version of Karlin model assumes the total number of $\Pi(n)$ balls. According to well-known thinning property of Poisson flows, stochastic processes $\left\{\mathbb{X}_i(\Pi(t)) \stackrel{\text{def}}{=} \mathbf{\Pi}_i(t), t \geq 0\right\}$ are compound Poisson with intensities p_i and are mutually independent for different i 's. The definition implies that for any fixed $n \geq 1$, $\tau, t \in [0, 1]$

$$R_{\Pi(tn)} = \sum_{k=1}^{\infty} \mathbf{I}\left(\mathbf{\Pi}_k(tn) > 0\right) = \sum_{k=1}^{\infty} \mathbf{I}_k(tn),$$

$$R'_{\Pi(\tau n)} = \sum_{k=1}^{\infty} \mathbf{I}\left(\mathbf{\Pi}_k(n) - \mathbf{\Pi}_k((1-\tau)n) > 0\right) = \sum_{k=1}^{\infty} \mathbf{I}'_k(\tau n).$$

Let $\tau, t \in [0, 1]$

$$\mathbf{cov}\left(R_{\Pi(tn)}, R'_{\Pi(\tau n)}\right) = \sum_{k=1}^{\infty} \mathbf{cov}(\mathbf{I}_k(tn), \mathbf{I}'_k(\tau n))$$

$$= \sum_{k=1}^{\infty} \left(\mathbf{P}\left(\mathbf{\Pi}_k(tn) > 0, \mathbf{\Pi}_k(n) - \mathbf{\Pi}_k((1-\tau)n) > 0\right) - (1 - e^{-p_k tn})(1 - e^{-p_k \tau n})\right)$$

Note that if $t + \tau > 1$, then

$$\mathbf{P}(\mathbf{\Pi}_k(tn) > 0, \mathbf{\Pi}_k(n) - \mathbf{\Pi}_k((1-\tau)n) > 0) = \mathbf{P}(\mathbf{\Pi}_k(tn) - \mathbf{\Pi}_k((1-\tau)n) > 0)$$

$$+\mathbf{P}(\mathbf{\Pi}_k(tn) - \mathbf{\Pi}_k((1-\tau)n) = 0, \mathbf{\Pi}_k((1-\tau)n) > 0, \mathbf{\Pi}_k(n) - \mathbf{\Pi}_k(tn) > 0)$$

$$= 1 - e^{-p_k(t+\tau-1)n} + e^{-p_k(t+\tau-1)n}(1 - e^{-p_k(1-\tau)n})(1 - e^{-p_k(1-t)n})$$

$$= 1 - e^{-p_k tn} - e^{-p_k \tau n} + e^{-p_k n}.$$

Hence

$$\mathbf{cov}\left(R_{\Pi(tn)}, R'_{\Pi(\tau n)}\right) = \mathbf{I}(t + \tau > 1) \sum_{k=1}^{\infty} \left(e^{-p_k n} - e^{-p_k(t+\tau)n}\right)$$

$$= \mathbf{I}(t + \tau > 1)(\mathbf{E}R_{\Pi((t+\tau)n)} - \mathbf{E}R_{\Pi(n)}).$$

Since

$$\mathbf{E}R_{\Pi(tn)}/\mathbf{E}R_n \sim t^{\theta},$$

then

$$\mathbf{cov}\left(R_{\Pi(tn)}, R'_{\Pi(\tau n)}\right)/\mathbf{E}R_n \sim K'(t,\tau).$$

For any fixed $m \geq 1$, $0 < t_1 < t_2 < \cdots < t_m \leq 1$, triangle array of $2m$-dimensional random vectors

$$\left\{((\mathbf{I}_k(nt_i) - \mathbf{E}\mathbf{I}_k(nt_i))/\sqrt{\mathbf{E}R_n}, (\mathbf{I}'_k(nt_i) - \mathbf{E}\mathbf{I}'_k(nt_i))/\sqrt{\mathbf{E}R_n}, i \leq m), k \leq n\right\}_{n\geq 1}$$

satisfies Lindeberg condition. So we have the convergence of finite-dimensional distributions for the poissonised process.

Since $R_{\Pi(nt)} \stackrel{d}{=} R'_{\Pi(nt)}$, the relative compactness and the approximation of the original process follow from steps 3 and 4 in the proof of Theorem 1 in Chebunin and Kovalevskii (2016).

Theorem 1 is proved.

Proof of Corollary 1

Due to SLLN, $\sqrt{R_n}/\sqrt{\mathbf{E}R_n} \to 1$ a.s.

So it is enough to prove the convergence for

$$J'_n = \frac{\sum_{k=1}^n (R_k - R'_k)}{n\sqrt{\mathbf{E}R_n}} = \int_0^1 (Z_n(t) - Z'_n(t))\, dt.$$

J'_n is the bounded continuous functional of (Z_n, Z'_n), so, due to Theorem 1, it converges to its limit

$$J' = \int_0^1 (Z(t) - Z'(t))\, dt.$$

This limit is the centered normal random variable, and

$$\mathbf{Var}\, J' = \mathbf{E}\left(\int_0^1 (Z(t) - Z'(t))dt\right)^2 = \int_0^1 \int_0^1 \mathbf{E}(Z(s) - Z'(s))(Z(t) - Z'(t))\, dsdt$$

$$= 2\int_0^1 \int_0^1 (K(s,t) - K'(s,t))\, dsdt = \frac{\theta}{\theta + 2}.$$

The proof is complete.