S©MR ISSN 1813-3304

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ

Siberian Electronic Mathematical Reports
http://semr.math.nsc.ru

Том 17, стр. 1959–1974 (2020) УДК 519.233
DOI 10.33048/semi.2020.17.132 MSC 62F03

A STATISTICAL TEST FOR CORRESPONDENCE OF TEXTS TO THE ZIPF—MANDELBROT LAW

A. CHAKRABARTY, M.G. CHEBUNIN, A.P. KOVALEVSKII,
I.M. PUPYSHEV, N.S. ZAKREVSKAYA, Q. ZHOU

Abstract. We analyse correspondence of texts to a simple probabilistic model. The model assumes that the words are selected independently from an infinite dictionary, and the probability distribution of words corresponds to the Zipf—Mandelbrot law. We count the numbers of different words in the text sequentially and get the process of the numbers of different words. Then we estimate the Zipf—Mandelbrot law's parameters using the same sequence and construct an estimate of the expectation of the number of different words in the text. After that we subtract the corresponding values of the estimate from the sequence and normalize along the coordinate axes, obtaining a random process on a segment from 0 to 1. We prove that this process (the empirical text bridge) converges weakly in the uniform metric on $C(0,1)$ to a centered Gaussian process with continuous a.s. paths. We develop and implement an algorithm for calculating the probability distribution of the integral of the square of this process. We present several examples of application of the algorithm for analysis of the homogeneity of texts in English, French, Russian, and Chinese.

Keywords: Zipf's law, weak convergence, Gaussian process.

Chakrabarty, A., Chebunin, M.G., Kovalevskii, A.P., Pupyshev, I.M., Zakrevskaya, N.S., Zhou, Q., A statistical test for correspondence of texts to the Zipf—Mandelbrot law.

The reported study was funded by RFBR and NSFC according to the research project No. 19-51-53010.

Received September, 28, 2020, published November, 27, 2020.

1. Introduction

Our analysis is based on the fact that a text in any natural language can be divided into words. The source material for our analysis is a text with separated words and excluded punctuation. In addition, all capital letters (if any) are replaced by lowercase.

We test the hypothesis H that a text matches a simple probabilistic model. The model satisfies the following three assumptions:

1) the dictionary contains countably many words that are enumerated $i = 1, 2, \ldots$;

2) words are sampled from the dictionary in the i.i.d. fashion according to discrete distribution $F = \{p_i\}_{i \geq 1}$ where p_i is the probability of word i;

3) the distribution F has a power tail:

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

where q and θ are the distribution parameters and c is the normalising constant.

These assumptions have a long history. Power decay of probabilities together with unboundedness of the dictionary were proposed by Zipf (1936). Mandelbrot (1965) noted that shift q is required for the model to better match real texts. Modern large-scale studies (see Petersen et al., 2012) show that the process of the emergence of new words never stops, but very long sequences of texts show a slightly lower frequency of new words than it is predicted by the formula (1). So we apply our test for examples of not very long texts only.

The second assumption of the independence of the choice of consecutive words is obviously false for any meaningful text. We can easily reject it statistically. If we calculate the relative frequency of the word '*and* ' in an English text, and then the relative frequency of the sequence '*and and* ' (that is, two and in succession), then, according to H, the second should be approximately equal to the square of the first. In practice, the second is much smaller; often, it is simply zero.

But we are not ready to abandon the assumption of independence. This assumption is the base of our analysis. Therefore, we choose a characteristic that changes little when rearranging neighboring words. This characteristic is the number of different words of the text.

Let R_n be the number of different words in the text of n words. Under the assumption of independence, Bahadur proved the law of large numbers $R_n/\mathbf{E}R_n \to 1$ in probability, and Karlin proved strong law of large numbers $R_n/\mathbf{E}R_n \to 1$ a.s. and the central limit theorem: $(R_n - \mathbf{E}R_n)/\sqrt{\mathbf{Var}R_n}$ converges weakly to the standard normal law.

Thus, we consider a text as a random sequence, we construct the sequence $R_1, \ldots, R_n$ and study it using the methods of the theory of random processes: we invent parameter estimates and construct the empirical text bridge, that is, a random process built on the parameter estimates and the sequence of numbers of different words. We find the limit Gaussian process in the sense of weak convergence in $C(0,1)$. Then we calculate the distribution function G of the integral of the squared limit process using the Smirnov (1937) formula.

Eigenvalues of the covariance function that are necessary for applying the Smirnov formula are calculated approximately as the eigenvalues of the matrix $Q = (q_{ij})_{i,j=1}^{L}$

composed of coefficients

$$q_{ij} = \int_0^1 \int_0^1 \widehat{K}(s,t) \sin \pi i s \sin \pi j t \, ds dt. \tag{2}$$

We calculate q_{ij} by a fast algorithm that reduces double integrals to definite integrals.

Asymptotics of R_n and similar statistics (in particular, the number of unique words, that is, words with exactly one occurrence in the text) have been studied by a number of authors. The Gaussian approximation under assumptions (1) was studied by Karlin, and beyond these assumptions by Dutko (1989), Gnedin, Hansen and Pitman (2007), Hwang and Janson (2008), Barbour and Gnedin (2009). Barbour (2009) proposed translated Poisson approximation for the number of unique words. New papers by Ben-Hamou, Boucheron and Ohannessian (2017) and Decrouez, Grabchak and Paris (2018) proved new general facts about these statistics.

The main result on which our study is based is the functional central limit theorem for the sequence $R_1, \dots, R_n$. It was proven by Chebunin and Kovalevskii (2016) in preparation for this study. Note that the Gaussian process which is the limit for this sequence can be found also in Durieu and Wang (2016) as a limit for another prelimit process. Its generalization is in Durieu, Samorodnitsky and Wang (2019).

Zipf parameter estimates were proposed by Nicholls (1987), Chebunin and Kovalevskii (2019b), but we need a special estimate for which we can calculate joint limit distribution of it and of the sequence $R_1, \dots, R_n$. This estimate is proposed in Chebunin and Kovalevskii (2019a). Zakrevskaya and Kovalevskii (2019) used the estimate in analysis of Shakespeare's sonnets.

Bahadur proved that under H the mathematical expectation of the number of different words grows according to an asymptotically power law. This fact is known to experts in natural language processing as Herdan's law (Herdan, 1960) or Heaps' law (Heaps, 1978). Van Leijenhorst and van der Weide (2005), Eliazar (2011) analyzed the relationship between the Zipf's law and Heaps' law based on probabilistic models that were different from H.

Gerlach and Altmann (2013) noted specifically that there is no mathematically correct statistical test for correspondence of a text to the Zipf's law. We proposed such a test in Chebunin and Kovalevskii (2019a). We are developing the algorithm and applying it to texts in different languages in the present paper.

We propose an estimate for the Mandelbrot parameter q, prove its consistence, then we construct an approximation of the process of the number of different words using this estimate, and prove the weak convergence of the normalized difference between the process and the approximation to a centered Gaussian process. We calculate the covariance function of the limiting process and the distribution of the integral of the square of this process. We use these results for an algorithm for calculating the p-value of the hypothesis of text homogeneity and apply the algorithm to the analysis of the homogeneity of texts in different languages.

The rest of the paper is organised as follows. The necessary theoretical results are in Section 2, the algorithm is in Section 3, examples of text analysis are in Section 4, and a discussion is in Section 5.

2. Theoretical results

The Hurvitz zeta function is

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

Note that constant c in (1) is

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

Let n be a number of words in a text.

Let R_k be the number of different words among first k words of the text.

Let $R_0=0$. We have $R_1=1$, $R_0<R_1\le R_2\le\ldots\le R_n$.

Bahadur (1960) proved that

$$\mathbf{E}R_j\sim c^{\theta}\Gamma(1-\theta)j^{\theta}, \tag{3}$$

where $\Gamma(x)=\int_0^{\infty}y^{x-1}e^{-y}\,dy$ is the Euler gamma function. Bahadur also proved convergence in probability $R_j/\mathbf{E}R_j\xrightarrow{p}1$.

Karlin (1967) proved that $R_j/\mathbf{E}R_j\xrightarrow{a.s.}1$, which is equivalent to

$$R_j\sim C_1j^{\theta}\text{ a.s.}, \tag{4}$$

thanks to (3). Here $C_1=C_1(\theta,q)=c^{\theta}\Gamma(1-\theta)$.

Chebunin and Kovalevskii (2016) proved the Functional Central Limit Theorem, that is, the weak convergence of the process $\{(R_{[nt]}-\mathbf{E}R_{[nt]})/\sqrt{\mathbf{E}R_n},\ 0\le t\le 1\}$ to a centered Gaussian process Z_θ with continuous a.s. sample paths and covariance function

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

Denote $\log^+x=\max(\log x,\,0)$.

We propose the following estimator for parameter θ (Chebunin and Kovalevskii, 2019a)

$$\widehat{\theta}=\int_0^1\log^+R_{[nt]}\,dA(t)$$

with a function $A(\cdot)$ that is the sum of a step function and a piecewise continuosly differentiable function on [0, 1] and

$$\int_0^1\log t\,dA(t)=1,\quad A(t)=0,\ t\in[0,\,\delta],\ \text{for some }\delta\in(0,\,1),\ A(1)=0. \tag{5}$$

The next theorem follows from Theorems 2.1, 2.2 by Chebunin and Kovalevskii (2019a).

Theorem 2.1 If (5) holds then the estimator $\widehat{\theta}$ is strongly consistent, and

$$\sqrt{\mathbf{E}R_n}(\widehat{\theta}-\theta)-\int_0^1t^{-\theta}Z_n(t)\,dA(t)\to_p 0.$$

From Theorem 2.1, it follows that $\widehat{\theta}$ converges to θ with rate $(\mathbf{E}R_n)^{-1/2}$, and normal random variable $\int_0^1t^{-\theta}Z_\theta(t)\,dA(t)$ has variance $\int_0^1\int_0^1(st)^{-\theta}K(s,t)\,dA(s)\,dA(t)$.

Example 2.1 Take

$$A(t)=\begin{cases}0, & 0\le t\le 1/2;\\ -(\log 2)^{-1}, & 1/2<t<1;\\ 0, & t=1.\end{cases}$$

Then

$$\widehat{\theta} = \log_2(R_n/R_{[n/2]}), \quad n \geq 2. \tag{6}$$

Note that, in this example, for any function g on $[0,1]$,

$$\int_0^1 g(t)\, dA(t) = \frac{g(1) - g(1/2)}{\log 2}.$$

Let us introduce the process $\widehat{Z}_n$:

$$\widehat{Z}_n(k/n) = \left(R_k - (k/n)^{\widehat{\theta}} R_n\right) / \sqrt{R_n},$$

$0 \leq k \leq n$. Let for $0 \leq t \leq 1/n$ and $0 \leq k \leq n-1$

$$\widehat{Z}_n\left(\frac{k}{n} + t\right) = \widehat{Z}_n(k/n) + nt\left(\widehat{Z}_n((k+1)/n) - \widehat{Z}_n(k/n)\right).$$

Theorem 2.2 (Theorem 4.1 by Chebunin and Kovalevskii, 2019a) If (5) holds then $\widehat{Z}_n$ converges weakly to $\widehat{Z}_\theta$ as $n \to \infty$, where

$$\widehat{Z}_\theta(t) = Z^0_\theta(t) - t^\theta \log t \int_0^1 u^{-\theta} Z_\theta(u)\, dA(u),$$

$Z^0_\theta(t) = Z_\theta(t) - t^\theta Z_\theta(1)$, $0 \leq t \leq 1$.

The correlation function of Z^0_θ is given by

$$K^0(s,t) = \mathbf{E} Z^0_\theta(s) Z^0_\theta(t) = K(s,t) - s^\theta K(1,t) - t^\theta K(s,1) + s^\theta t^\theta K(1,1).$$

Corollary 2.1 Let $\widehat{W}^2_n = \int\limits_0^1 \left(\widehat{Z}_n(t)\right)^2 dt$ and (5) holds. Then $\widehat{W}^2_n$ converges weakly to $\widehat{W}^2_\theta = \int\limits_0^1 \left(\widehat{Z}_\theta(t)\right)^2 dt$.

$\widehat{W}^2_n$ has the following representation

$$\widehat{W}^2_n = \frac{1}{3n} \sum_{k=1}^{n-1} \widehat{Z}_n\left(\frac{k}{n}\right)\left(2\widehat{Z}_n\left(\frac{k}{n}\right) + \widehat{Z}_n\left(\frac{k+1}{n}\right)\right).$$

The p-value of the goodness-of fit test is $1 - \widehat{F}_\theta(\widehat{W}^2_{n,obs})$. Here $\widehat{F}_\theta$ is the cumulative distribution function of $\widehat{W}^2_\theta$, and $\widehat{W}^2_{n,obs}$ is the observed value of $\widehat{W}^2_n$.

One can estimate F_θ by simulations or find it explicitely using the Smirnov's formula (Smirnov, 1937): if $\widehat{W}^2_\theta = \sum_{k=1}^\infty \frac{\eta_k^2}{\lambda_k}$, $\eta_1, \eta_2, \ldots$ are independent and have standard normal distribution, $0 < \lambda_1 < \lambda_2 < \ldots$, then

$$F_\theta(x) = 1 + \frac{1}{\pi} \sum_{k=1}^\infty (-1)^k \int_{\lambda_{2k-1}}^{\lambda_{2k}} \frac{e^{-\lambda x/2}}{\sqrt{-D(\lambda)}} \cdot \frac{d\lambda}{\lambda}, \; x > 0, \tag{7}$$

$$D(\lambda) = \prod_{k=1}^\infty \left(1 - \frac{\lambda}{\lambda_k}\right).$$

The integrals in the RHS of (7) must tend to 0 monotonically as $k \to \infty$, and λ_k^{-1} are the eigenvalues of kernel $\widehat{K}(s,t) = \mathbf{E}\widehat{Z}_\theta(s)\widehat{Z}_\theta(t)$, see Smirnov (1937).

We are exploring the two-parameter model in this paper. Therefore, we now develop a new approximation of the process of the number of different words using the estimate of the Mandelbrot parameter q.

Let us introduce the empirical bridge of a text $\widetilde{Z}_n$ by substituting

$$r_k = r_k(\widehat{\theta}, \widehat{q}) = \sum_{i=1}^{\infty} \left(1 - (1 - \widehat{p}_i)^k\right) \tag{8}$$

instead of $(k/n)^{\widehat{\theta}} R_n$ in $\widehat{Z}_n$, that is,

$$\widetilde{Z}_n(k/n) = (R_k - r_k)/\sqrt{R_n}, \tag{9}$$

$0 \le k \le n$.

Here

$$\widehat{p}_i = c(\widehat{\theta}, \widehat{q})(i + \widehat{q})^{-1/\widehat{\theta}}, \quad i \ge 1, \tag{10}$$

$$\widehat{q} = \min\{\widetilde{q} > -1 : \ r_n(\widehat{\theta}, \widetilde{q}) = R_n\}. \tag{11}$$

Theorem 2.3 *If H is true then $\widehat{q} \to q$ in probability as $n \to \infty$, and $\widehat{Z}_n - \widetilde{Z}_n \to 0$ in probability uniformly on q in any segment in $(-1, \infty)$.*

Proof

From Lemma 1 in Gnedin et al. (2007) we have

$$|\mathbf{E}R_n - \mathbf{E}R_{\Pi(n)}| < \frac{2}{n}\mathbf{E}R_{n,2} \le \frac{2}{n} \times \frac{n}{2} = 1.$$

Karlin (1967) proposed representation

$$\mathbf{E}R_{\Pi(t)} = \int_0^{\infty} \alpha(ty) \frac{1}{y^2} e^{-1/y} dy,$$

where

$$\alpha(x) = \max\{j | p_j \ge \frac{1}{x}\}.$$

In our case

$$\alpha(x) = [(cx)^{\theta} - q].$$

We represent

$$\Gamma(k - \gamma) = \int_0^{\infty} y^{\gamma - k - 1} e^{-1/y} dy.$$

So

$$(cn)^{\theta}\Gamma(1-\theta) - q - 1 \le \mathbf{E}R_{\Pi(n)} \le (cn)^{\theta}\Gamma(1-\theta) - q,$$
$$\mathbf{E}R_n = (cn)^{\theta}\Gamma(1-\theta) - q + T_n, \quad |T_n| < 2,$$

then

$$|r_n - (\widehat{cn})^{\widehat{\theta}}\Gamma(1-\widehat{\theta}) + \widehat{q}| < 2 \text{ a.s.}$$

Notice, that for any $x > 0$

$$\zeta(\theta^{-1}, x) = \sum_{i=0}^{\infty} (i+x)^{-\theta^{-1}} > \sum_{i=0}^{\infty} \int_i^{i+1} (y+x)^{-\theta^{-1}} dy = \int_0^{\infty} (y+x)^{-\theta^{-1}} dy = c_1 x^{1-\theta^{-1}},$$

where $c_1 = c_1(\theta) = (\theta^{-1} - 1)^{-1}$, and

$$\zeta(\theta^{-1}, x) = \sum_{i=0}^{\infty} (i+x)^{-\theta^{-1}} < x^{-\theta^{-1}} + \sum_{i=1}^{\infty} \int_{i-1}^{i} (y+x)^{-\theta^{-1}} dy = x^{-\theta^{-1}} + c_1 x^{1-\theta^{-1}}.$$

It's not hard to see that

$$c^{\theta} = (\zeta(\theta^{-1}, q+1))^{-\theta} \sim c_1 q^{1-\theta} \quad \text{as } q \to \infty.$$

From the strong law of large numbers for R_n and $\widehat{\theta} \to \theta$ a.s., we can choose a constant $K = K(\theta, q) = 2c^{\theta}\Gamma(1-\theta)$ such that

$$\mathbf{P}(R_n < Kn^{\widehat{\theta}}) \to 1 \text{ as } n \to \infty.$$

Therefore, we can choose q^* such, that

$$(c(\widehat{\theta}, q^*))^{\widehat{\theta}}\Gamma(1-\widehat{\theta}) \underset{a.s.}{\sim} (c(\theta, q^*))^{\theta}\Gamma(1-\theta) > K.$$

Hence

$$\mathbf{P}(\widehat{q}_n < q^*) \to 1 \text{ as } n \to \infty.$$

On the other hand, from (3) we have

$$R_n \underset{a.s.}{\sim} (c(\theta, q)n)^{\theta}\Gamma(1-\theta),$$

$$r_n \underset{a.s.}{\sim} (c(\widehat{\theta}, \widehat{q})n)^{\widehat{\theta}}\Gamma(1-\widehat{\theta}),$$

and $r_n = R_n$, $\widehat{\theta} \to \theta$ a.s., so

$$\frac{(c(\widehat{\theta}, \widehat{q})n)^{\widehat{\theta}}}{(c(\theta, q)n)^{\theta}} \xrightarrow{a.s.} 1.$$

From Theorem 2.1, $n^{\widehat{\theta}-\theta} \to 1$ in probability, so $c(\widehat{\theta}, \widehat{q}) \to c(\theta, q)$ in probability, and $c(\theta, q)$ is a continious and strictly monotone function of q for any $\theta \in (0, 1)$. Thus $\widehat{q} \to q$ in probability.

Now we estimate

$$\max_{1 \le k \le n} \frac{|(k/n)^{\widehat{\theta}}R_n - r_k|}{\sqrt{R_n}}.$$

We have

$$\begin{aligned} |(k/n)^{\widehat{\theta}}R_n - r_k| &= |(k/n)^{\widehat{\theta}}r_n - r_k| \\ &< |(k/n)^{\widehat{\theta}}((\widehat{c}n)^{\widehat{\theta}}\Gamma(1-\widehat{\theta}) - \widehat{q}) - (\widehat{c}k)^{\widehat{\theta}}\Gamma(1-\widehat{\theta}) + \widehat{q}| + 4 \\ &= |\widehat{q}(1-(k/n)^{\widehat{\theta}})| + 4 \le \widehat{q} + 4. \end{aligned}$$

As $R_n \to \infty$ a.s. and $\widehat{q} \to q$ in probability, we have

$$\max_{1 \le k \le n} \frac{|(k/n)^{\widehat{\theta}}R_n - r_k|}{\sqrt{R_n}} \to 0$$

in probability.

The proof is complete.

3. Algorithm

Note that

$$\begin{aligned} r_k &= \sum_{i=1}^{\infty} \left(1 - \sum_{j=0}^{k} C_k^j (-1)^j (\widehat{p}_i)^j\right) \\ &= \sum_{j=1}^{k} C_k^j (-1)^{j+1} \sum_{i=1}^{\infty} \widehat{c}^j (i+\widehat{q})^{-j/\widehat{\theta}} \\ &= \sum_{j=1}^{k} C_k^j (-1)^{j+1} \zeta(j/\widehat{\theta}, 1+\widehat{q}) \left(\zeta(1/\widehat{\theta}, 1+\widehat{q})\right)^{-j}. \end{aligned}$$

So we have the finite formula to calculate $r_1, \dots, r_n$. This formula is not good for large k due to high complexity of calculations of binomial coefficients. So we use an appoximation by substituting an intergal instead of series residual

$$r_k \approx \sum_{i=1}^{M}(1-(1-\widehat{p}_i)^k) + \int\limits_{M+0.5+\widehat{q}}^{\infty}(1-\exp(-k\widehat{c}y^{-\widehat{\alpha}}))dy$$

$$= \sum_{i=1}^{M}(1-(1-\widehat{p}_i)^k) + (k\widehat{c})^{\widehat{\theta}}\int\limits_{0}^{k\widehat{c}N^{-\widehat{\alpha}}} z^{-\widehat{\theta}}e^{-z}dz - N(1-\exp(-k\widehat{c}N^{-\widehat{\alpha}})), \tag{12}$$

$\widehat{\alpha} = 1/\widehat{\theta}$, $N = M + 0.5 + \widehat{q}$.

We calculate the integral using incomplete Gamma function. We find $\widehat{q}$ by dichotomy method for $r_n = R_n$.

Elementary calculations give

$$\widehat{K}(s,t) = \mathbf{E}\widehat{Z}_\theta(s)\widehat{Z}_\theta(t)$$

$$= K^0(s,t) - t^\theta \log t \frac{K(s,1)-2^\theta K(s,1/2)}{\log 2} - s^\theta \log s \frac{K(t,1)-2^\theta K(t,1/2)}{\log 2}$$

$$+ s^\theta t^\theta (\log s + \log t)\frac{K(1,1)-2^\theta K(1,1/2)}{\log 2}$$

$$+ s^\theta t^\theta \log s \log t \frac{K(1,1)-2^{\theta+1}K(1,1/2)+2^{2\theta}K(1/2,1/2)}{\log^2 2}.$$

Now we represent the kernel $\widehat{K}$ by the matrix $Q = (q_{ij})_{i,j=1}^{L}$ by calculation of

$$q_{ij} = \int_0^1\int_0^1 \widehat{K}(s,t)\sin\pi is \sin\pi jt\, dsdt. \tag{13}$$

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

$$\widehat{K}(s,t) = K(s,t) - s^\theta K(1,t) - t^\theta K(s,1) + s^\theta t^\theta K(1,1)$$

$$-t^\theta \log t \frac{K(s,1)-2^\theta K(s,1/2)}{\log 2} - s^\theta \log s \frac{K(t,1)-2^\theta K(t,1/2)}{\log 2}$$

$$+ s^\theta t^\theta (\log s + \log t)\frac{K(1,1)-2^\theta K(1,1/2)}{\log 2}$$

$$+ s^\theta t^\theta \log s \log t \frac{K(1,1)-2^{\theta+1}K(1,1/2)+2^{2\theta}K(1/2,1/2)}{\log^2 2}.$$

Let us denote

$$J_{ij} = \int_0^1\int_0^1 (s+t)^\theta \sin\pi is \sin\pi jt\, dsdt,$$

$$a(\theta,k,x) = \int_0^x t^\theta \sin\pi kt\, dt, \quad b(\theta,k,x) = \int_0^x t^\theta \cos\pi kt\, dt,$$

$$A_i = \int_0^1 t^\theta \sin\pi it\, dt = a(\theta,i,1), \quad B_i = \int_1^2 t^\theta \sin\pi it\, dt = a(\theta,i,2) - a(\theta,i,1),$$

$$C_i = \int_0^1 t^{\theta+1}\cos\pi it\, dt = b(\theta+1,i,1),$$

$$D_i = \int_1^2 t^\theta (2-t)\cos \pi i t \, dt = 2b(\theta, i, 2) - 2b(\theta, i, 1) - b(\theta+1, i, 2) + b(\theta+1, i, 1),$$

$$E_{ij} = \int_0^1 t^\theta \sin \pi i t \cos \pi j t \, dt, \quad F_i = \int_0^1 K(t,1) \sin \pi i t \, dt,$$

$$G_i = \int_0^1 t^\theta \log t \sin \pi i t \, dt, \quad H_i = \int_0^1 K(t, 1/2) \sin \pi i t \, dt.$$

Then we have

$$\begin{aligned} q_{ij} = J_{ij} &- \frac{1}{\pi j}(A_i - E_{ij}) - \frac{1}{\pi i}(A_j - E_{ji}) \\ &- A_i F_j - A_j F_i + K(1,1) A_i A_j \\ &- G_j \frac{F_i - 2^\theta H_i}{\log 2} - G_i \frac{F_j - 2^\theta H_j}{\log 2} \\ &+ (A_i G_j + A_j G_i) \frac{K(1,1) - 2^\theta K(1,1/2)}{\log 2} \\ &+ G_i G_j \frac{K(1,1) - 2^{\theta+1} K(1,1/2) + 2^{2\theta} K(1/2,1/2)}{\log^2 2}. \end{aligned}$$

We calculate J_{ij} substituting $t = u - s$, $s \le u \le s+1$. So

$$J_{ij} = \int_0^1 u^\theta \, du \int_0^u \sin \pi i s \sin \pi j (u-s) \, ds + \int_1^2 u^\theta \, du \int_{u-1}^1 \sin \pi i s \sin \pi j (u-s) \, ds.$$

If $i \ne j$ then we have

$$\begin{aligned} J_{ij} &= \frac{1}{2} \int_0^1 u^\theta \, du \int_0^u (\cos \pi(is - ju + js) - \cos \pi(is + ju - js)) \, ds \\ &+ \frac{1}{2} \int_1^2 u^\theta \, du \int_{u-1}^1 (\cos \pi(is - ju + js) - \cos \pi(is + ju - js)) \, ds \\ &= \frac{1}{2} \int_0^1 u^\theta \left(\frac{\sin \pi i u + \sin \pi j u}{\pi(i+j)} - \frac{\sin \pi i u - \sin \pi j u}{\pi(i-j)} \right) du \\ &+ \frac{1}{2} \int_1^2 u^\theta \left(\frac{\sin(\pi(i+j) - \pi j u) + \sin(\pi(i+j) - \pi i u)}{\pi(i+j)} \right. \\ &\left. - \frac{\sin(\pi(i-j) + \pi j u) + \sin(\pi(i-j) - \pi i u)}{\pi(i-j)} \right) du \\ &= \frac{i A_j - j A_i - (-1)^{i+j}(i B_j - j B_i)}{\pi(i^2 - j^2)}. \end{aligned}$$

If $i = j$ then we have

$$\begin{aligned} J_{ii} &= \frac{1}{2} \int_0^1 u^\theta \, du \int_0^u (\cos \pi(2is - iu) - \cos \pi i u) \, ds \\ &+ \frac{1}{2} \int_1^2 u^\theta \, du \int_{u-1}^1 (\cos \pi(2is - iu) - \cos \pi i u) \, ds \\ &= \frac{1}{2} \int_0^1 \left(u^\theta \frac{\sin \pi i u}{\pi i} - u^{\theta+1} \cos \pi i u \right) du \\ &+ \frac{1}{2} \int_1^2 u^\theta \frac{\sin(2\pi i - \pi i u) - \sin(2\pi i u - 2\pi i - \pi i u)}{2\pi i} \, du - \frac{1}{2} \int_1^2 u^\theta (2-u) \cos \pi i u \, du \end{aligned}$$

$$= \frac{A_i - B_i}{2\pi i} - \frac{C_i + D_i}{2}.$$

We know

$$a(\theta, k, x) = \pi k x^{\theta+2}\, {}_1F_2\left(\theta/2+1, 3/2, \theta/2+2, -k^2\pi^2x^2/4\right)/(\theta+2),$$

$$b(\theta, k, x) = x^{\theta+1}\, {}_1F_2\left(\theta/2+1/2, 1/2, \theta/2+3/2, -k^2\pi^2x^2/4\right)/(\theta+1),$$

$$E_{ij} = \frac{1}{2}A_{i+j} + \frac{1}{2}A_{i-j},$$

$$F_i = (-1)^i B_i + \frac{(-1)^i - 1}{\pi i},$$

$$G_i = \int_0^1 t^\theta \log t \sum_{k=0}^{\infty} \frac{(-1)^k(\pi i t)^{2k+1}}{(2k+1)!}\, dt = -\sum_{k=0}^{\infty} \frac{(-1)^k(\pi i)^{2k+1}}{(2k+1)!(2k+\theta+2)^2}$$

$$= -\pi i\, {}_2F_3\left(\theta/2+1, \theta/2+1, 3/2, \theta/2+2, \theta/2+2, -i^2\pi^2/4\right)/(\theta+2)^2,$$

$$H_i = \int_{1/2}^{3/2} t^\theta \sin(\pi i t - \pi i/2)\, dt + (\cos(\pi i/2) - 1)2^{-\theta}/(\pi i) - \int_{1/2}^{1} t^\theta \sin \pi i t\, dt$$

$$= \pi i \cos(\pi i/2)\Big((3/2)^{\theta+2}\, {}_1F_2\left(\theta/2+1, 3/2, \theta/2+2, -9i^2\pi^2/16\right)$$

$$-2^{-\theta-2}\, {}_1F_2\left(\theta/2+1, 3/2, \theta/2+2, -i^2\pi^2/16\right)\Big)/(\theta+2)$$

$$-\sin(\pi i/2)\Big((3/2)^{\theta+1}\, {}_1F_2\left(\theta/2+1/2, 1/2, \theta/2+3/2, -9i^2\pi^2/16\right)$$

$$-2^{-\theta-1}\, {}_1F_2\left(\theta/2+1/2, 1/2, \theta/2+3/2, -i^2\pi^2/16\right)\Big)/(\theta+1)$$

$$+(\cos(\pi i/2)-1)2^{-\theta}/(\pi i) - A_i + \pi i 2^{-\theta-2}\, {}_1F_2\left(\theta/2+1, 3/2, \theta/2+2, -i^2\pi^2/16\right)/(\theta+2).$$

${}_1F_2$ and ${}_2F_3$ are generalized hypergeometric functions, hyp1f2 and hyp2f3 in Python.

We calculate eigenvalues λ_i, $1 \le i \le L$, of the matrix Q. Then we use the Smirnov formula to calculate the p-value.

Let $W_n^2 = \int\limits_0^1 \left(Z_n^0(t)\right)^2 dt$. It is equal to

$$W_n^2 = \frac{1}{3n} \sum_{k=1}^{n-1} Z_n^0\left(\frac{k}{n}\right)\left(2Z_n^0\left(\frac{k}{n}\right) + Z_n^0\left(\frac{k+1}{n}\right)\right). \tag{14}$$

Then W_n^2 converges weakly to $W_\theta^2 = \int\limits_0^1 \left(Z_\theta^0(t)\right)^2 dt$.

So the test rejects the basic hypothesis if $W_n^2 \ge C$. The p-value of the test is $1 - F_\theta(W_{n,obs}^2)$. Here F_θ is the cumulative distribution function of W_θ^2 and $W_{n,obs}^2$ is a concrete value of W_n^2 for observations under consideration.

So, we developed the following text homogeneity analysis algorithm.

Algorithm

1. Remove all punctuation marks.
2. Calculate the process of numbers of different words.
3. Calculate the estimate $\widehat{\theta}$ of the parameter θ by (6).
4. Calculate the estimate $\widehat{q}$ of the parameter q by (11) with (12) for $k = n$ and (10).

5. Calculate the normalised difference of the process of numbers of different words and its estimate (the empirical bridge $\widetilde{Z}_n$) by (9) with (12).
6. Calculate the integral $W^2_{n,obs} = W^2_n$ of the square of the empirical bridge by (14).
7. Represent the kernel $\widehat{K}$ by the matrix $Q = (q_{ij})_{i,j=1}^{L}$ by (13). The integrals are calculated by methods of Section 3.
8. Find eigenvalues λ_i of the matrix.
9. Calculate the p-value $1 - F_\theta(W^2_{n,obs})$ by (7).

L and M are the parameters of the algorithm, L is the matrix dimension, M is the constant in (12).

4. Examples of text analysis

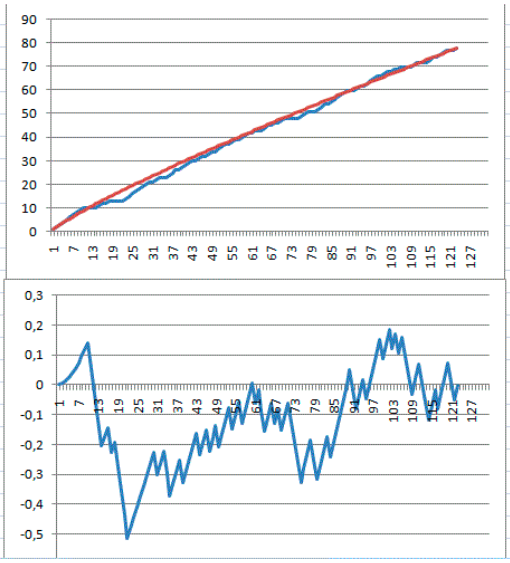
We use the algorithm with $L = 100$, $M = 1000$.

We find $\widehat{q}$ by dichotomy method for $r_n = R_n$ with 20 iterations on segment $[-0.9, 40]$.

Here are given an example of French poetry, *Les Regrets* by Joachim du Bellay (1558), sonnet 1 (Pic. 1) and Shakespeare's sonnet 1 (Pic. 2).

These are examples of homogeneous texts.

1	je	1	1	0
2	ne	2	1.961141	0.0044
3	veux	3	2.88887	0.012583
4	point	4	3.787758	0.024032
5	fouiller	5	4.661642	0.038312
6	au	6	5.513751	0.055057
7	sein	7	6.346808	0.073959
8	de	8	7.163111	0.094759
9	la	9	7.964605	0.117235
10	nature	10	8.752938	0.141202
11	je	10	9.529515	0.053272
12	ne	10	10.29553	-0.03346
13	veux	10	11.05201	-0.11912
14	point	10	11.79983	-0.20379
15	cercher	11	12.53974	-0.17434
16	l'esprit	12	13.27241	-0.14407
17	de	12	13.99839	-0.22627
18	l'univers	13	14.71818	-0.19455
19	je	13	15.43221	-0.27539
20	ne	13	16.14086	-0.35563
21	veux	13	16.84447	-0.4353
22	point	13	17.54333	-0.51443
23	sonder	14	18.23772	-0.47983
24	les	15	18.92785	-0.44474
25	abysmes	16	19.61396	-0.4092



n	123
R_n	78
theta_hat	0.859137
q_hat	-0.13832
omega^2	0.036427
p-value	0.567342

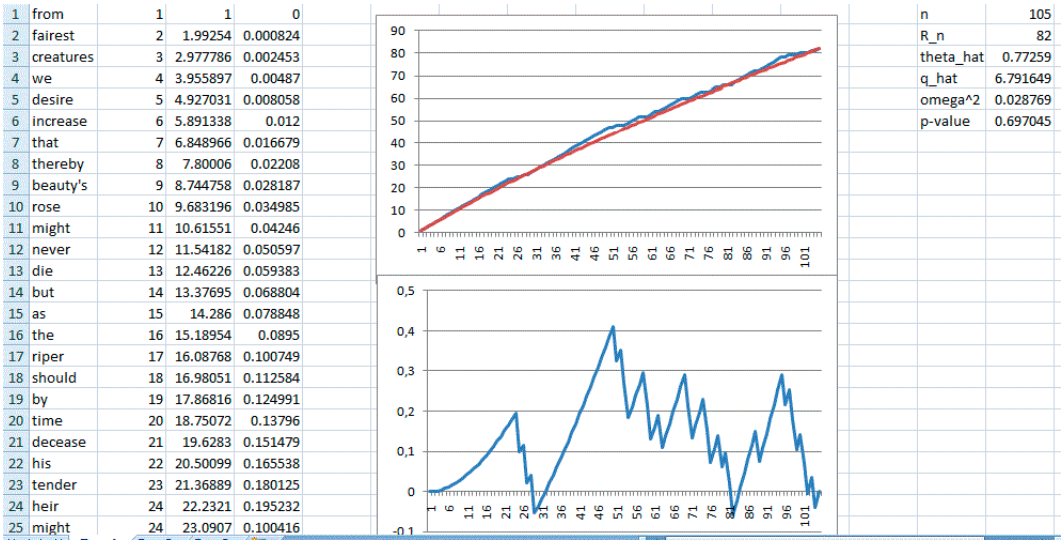
Pic. 1. *Les Regrets* by Joachim du Bellay (1558), sonnet 1

This approach can work with texts on any language.

See the first stanza of the first chapter of *Eugene Onegin* by Pushkin (Pic. 3) for example of Russian homogeneous text.

We use hieroglyphs instead of words when analyzing Chinese texts. We substitute hieroglyphs with their HTML codes. We analyse *Danqing Painting* by Du Fu (Pic. 4). This text is homogeneous, too.

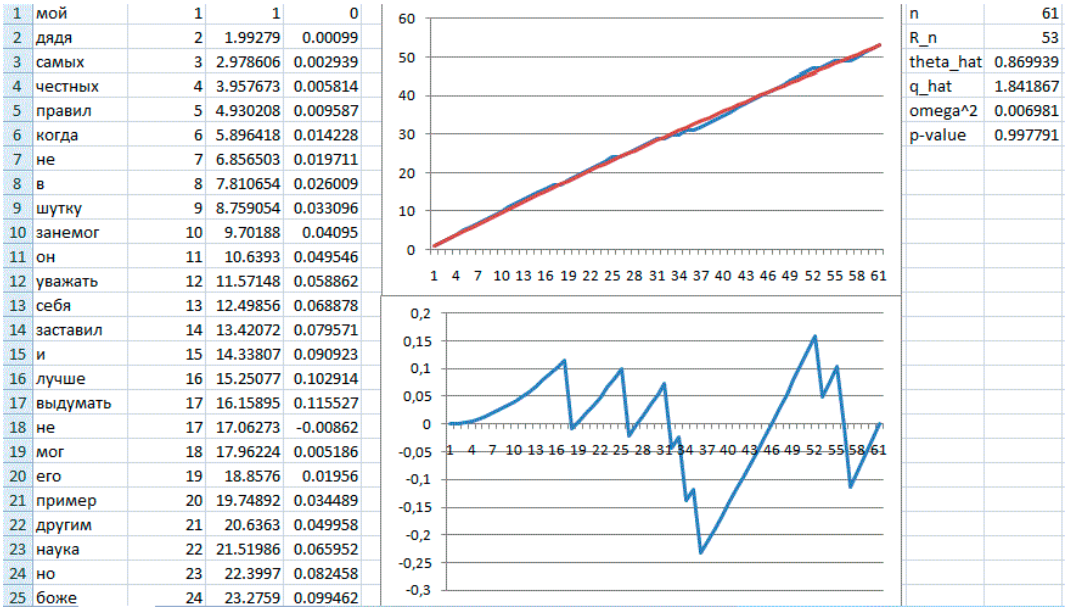
An example of a nonhomogeneous text is Shakespeare's sonnet 1 with 3 repeated lines (Pic. 5).



1	from	1	1	0
2	fairest	2	1.99254	0.000824
3	creatures	3	2.977786	0.002453
4	we	4	3.955897	0.00487
5	desire	5	4.927031	0.008058
6	increase	6	5.891338	0.012
7	that	7	6.848966	0.016679
8	thereby	8	7.80006	0.02208
9	beauty's	9	8.744758	0.028187
10	rose	10	9.683196	0.034985
11	might	11	10.61551	0.04246
12	never	12	11.54182	0.050597
13	die	13	12.46226	0.059383
14	but	14	13.37695	0.068804
15	as	15	14.286	0.078848
16	the	16	15.18954	0.0895
17	riper	17	16.08768	0.100749
18	should	18	16.98051	0.112584
19	by	19	17.86816	0.124991
20	time	20	18.75072	0.13796
21	decease	21	19.6283	0.151479
22	his	22	20.50099	0.165538
23	tender	23	21.36889	0.180125
24	heir	24	22.2321	0.195232
25	might	24	23.0907	0.100416

n	105
R_n	82
theta_hat	0.77259
q_hat	6.791649
omega^2	0.028769
p-value	0.697045

Рис. 2. Shakespeare's sonnet 1



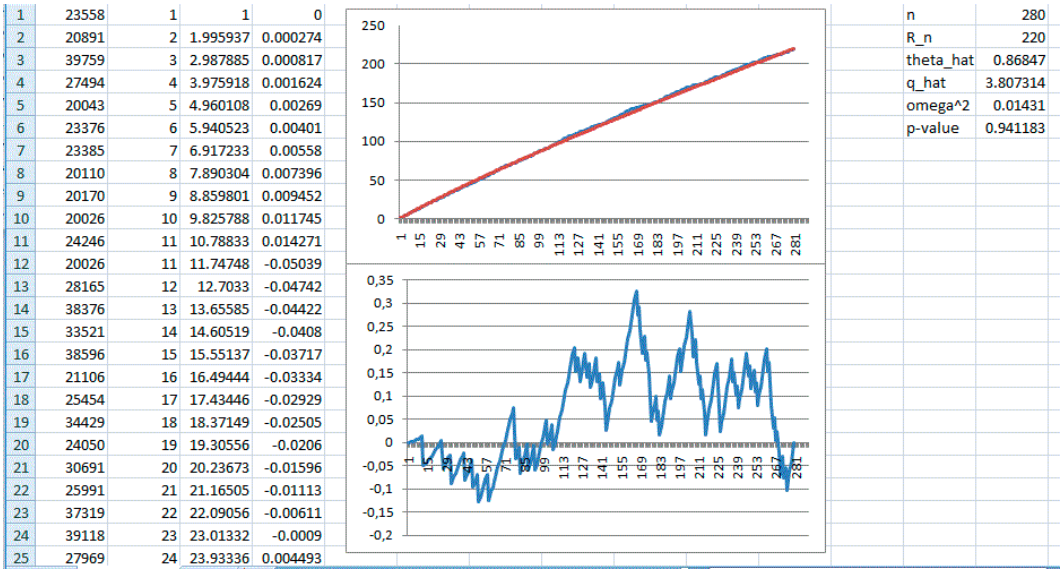
1	мой	1	1	0
2	дядя	2	1.99279	0.00099
3	самых	3	2.978606	0.002939
4	честных	4	3.957673	0.005814
5	правил	5	4.930208	0.009587
6	когда	6	5.896418	0.014228
7	не	7	6.856503	0.019711
8	в	8	7.810654	0.026009
9	шутку	9	8.759054	0.033096
10	занемог	10	9.70188	0.04095
11	он	11	10.6393	0.049546
12	уважать	12	11.57148	0.058862
13	себя	13	12.49856	0.068878
14	заставил	14	13.42072	0.079571
15	и	15	14.33807	0.090923
16	лучше	16	15.25077	0.102914
17	выдумать	17	16.15895	0.115527
18	не	17	17.06273	-0.00862
19	мог	18	17.96224	0.005186
20	его	19	18.8576	0.01956
21	пример	20	19.74892	0.034489
22	другим	21	20.6363	0.049958
23	наука	22	21.51986	0.065952
24	но	23	22.3997	0.082458
25	боже	24	23.2759	0.099462

n	61
R_n	53
theta_hat	0.869939
q_hat	1.841867
omega^2	0.006981
p-value	0.997791

Рис. 3. The first stanza of the first chapter of *Eugene Onegin* by Pushkin

Another example of nonhomogeneity is a text from different languages. The example is Du Bellay's sonnet 1 + Shakespeare's sonnet 1 (Pic. 6).

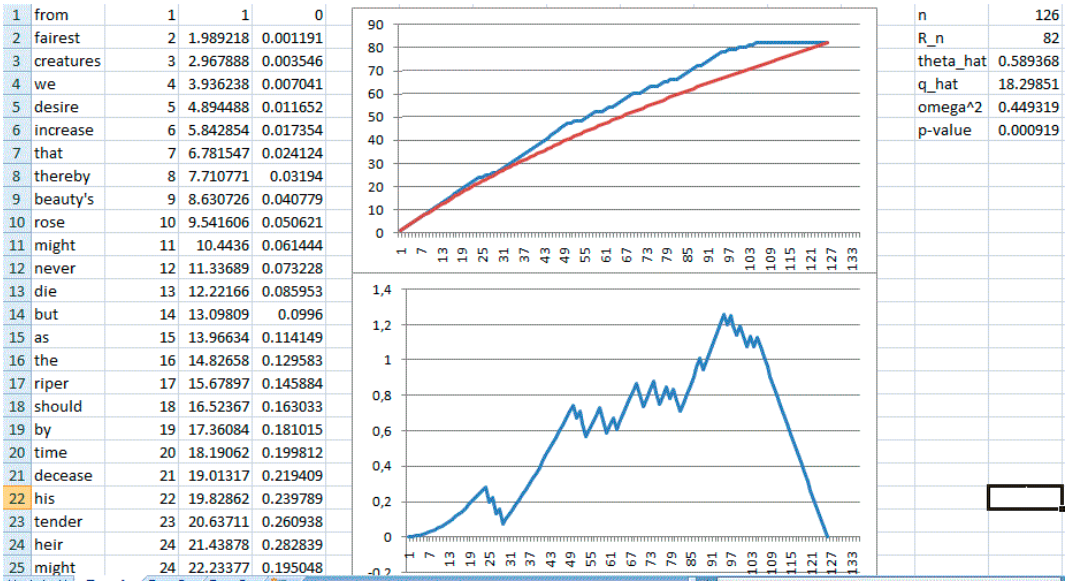
To find nonhomogeneity for a text in one language we need more longer texts. Here the first three stanzas of *Childe Harold's Pilgrimage* by Byron and the first three Shakespeare's sonnets (Pic. 7).



1	23558	1	1	0
2	20891	2	1.995937	0.000274
3	39759	3	2.987885	0.000817
4	27494	4	3.975918	0.001624
5	20043	5	4.960108	0.00269
6	23376	6	5.940523	0.00401
7	23385	7	6.917233	0.00558
8	20110	8	7.890304	0.007396
9	20170	9	8.859801	0.009452
10	20026	10	9.825788	0.011745
11	24246	11	10.78833	0.014271
12	20026	11	11.74748	-0.05039
13	28165	12	12.7033	-0.04742
14	38376	13	13.65585	-0.04422
15	33521	14	14.60519	-0.0408
16	38596	15	15.55137	-0.03717
17	21106	16	16.49444	-0.03334
18	25454	17	17.43446	-0.02929
19	34429	18	18.37149	-0.02505
20	24050	19	19.30556	-0.0206
21	30691	20	20.23673	-0.01596
22	25991	21	21.16505	-0.01113
23	37319	22	22.09056	-0.00611
24	39118	23	23.01332	-0.0009
25	27969	24	23.93336	0.004493

n	280
R_n	220
theta_hat	0.86847
q_hat	3.807314
omega^2	0.01431
p-value	0.941183

Рис. 4. *Danqing Painting* by Du Fu



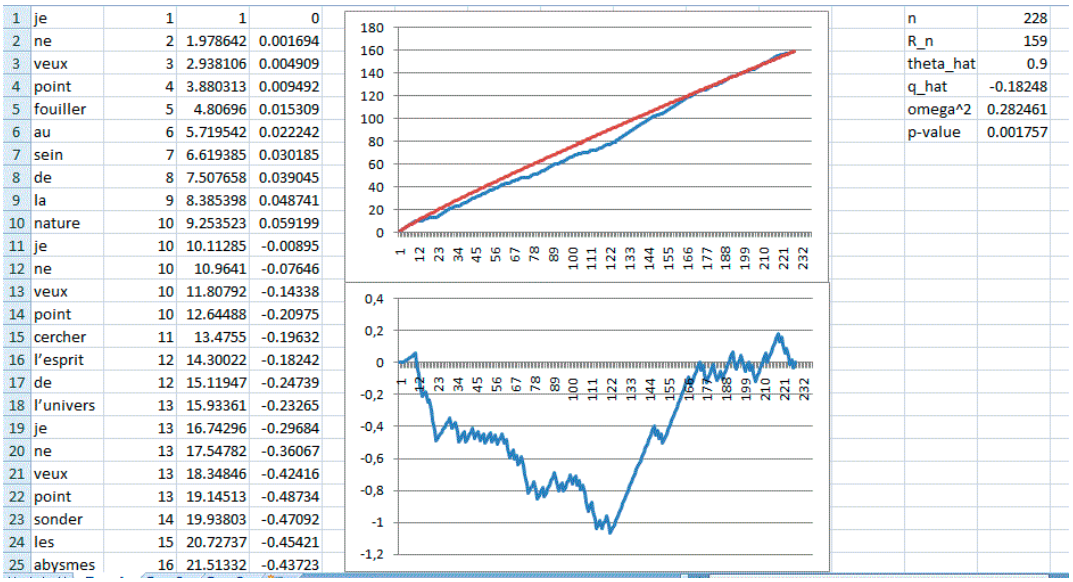
1	from	1	1	0
2	fairest	2	1.989218	0.001191
3	creatures	3	2.967888	0.003546
4	we	4	3.936238	0.007041
5	desire	5	4.894488	0.011652
6	increase	6	5.842854	0.017354
7	that	7	6.781547	0.024124
8	thereby	8	7.710771	0.03194
9	beauty's	9	8.630726	0.040779
10	rose	10	9.541606	0.050621
11	might	11	10.4436	0.061444
12	never	12	11.33689	0.073228
13	die	13	12.22166	0.085953
14	but	14	13.09809	0.0996
15	as	15	13.96634	0.114149
16	the	16	14.82658	0.129583
17	riper	17	15.67897	0.145884
18	should	18	16.52367	0.163033
19	by	19	17.36084	0.181015
20	time	20	18.19062	0.199812
21	decease	21	19.01317	0.219409
22	his	22	19.82862	0.239789
23	tender	23	20.63711	0.260938
24	heir	24	21.43878	0.282839
25	might	24	22.23377	0.195048

n	126
R_n	82
theta_hat	0.589368
q_hat	18.29851
omega^2	0.449319
p-value	0.000919

Рис. 5. Shakespeare's sonnet 1 with 3 repeated lines

5. Discussion

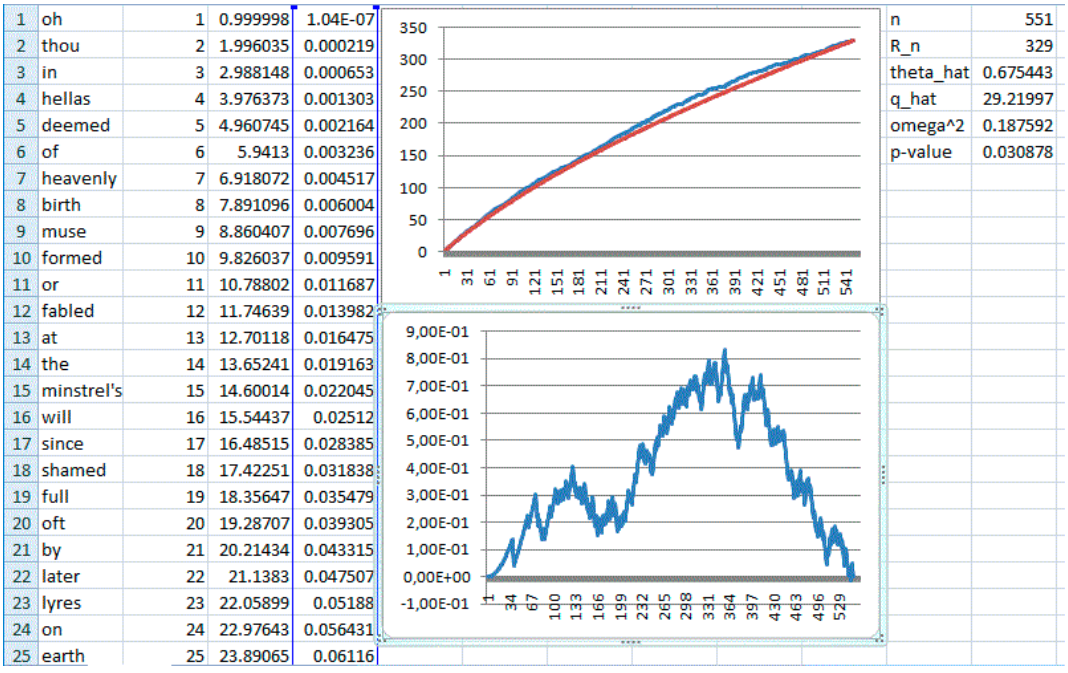
There are some open questions in the application of this approach. The first open question is the Poisson approximation. If the number of identical words is small, then the Gaussian approximation is inaccurate. Barbour (2009) proposed an approximation by a translated Poisson distribution. We need a functional version of his theorem.



1	je	1	1	0
2	ne	2	1.978642	0.001694
3	veux	3	2.938106	0.004909
4	point	4	3.880313	0.009492
5	fouiller	5	4.80696	0.015309
6	au	6	5.719542	0.022242
7	sein	7	6.619385	0.030185
8	de	8	7.507658	0.039045
9	la	9	8.385398	0.048741
10	nature	10	9.253523	0.059199
11	je	10	10.11285	-0.00895
12	ne	10	10.9641	-0.07646
13	veux	10	11.80792	-0.14338
14	point	10	12.64488	-0.20975
15	cercher	11	13.4755	-0.19632
16	l'esprit	12	14.30022	-0.18242
17	de	12	15.11947	-0.24739
18	l'univers	13	15.93361	-0.23265
19	je	13	16.74296	-0.29684
20	ne	13	17.54782	-0.36067
21	veux	13	18.34846	-0.42416
22	point	13	19.14513	-0.48734
23	sonder	14	19.93803	-0.47092
24	les	15	20.72737	-0.45421
25	abysmes	16	21.51332	-0.43723

n	228
R_n	159
theta_hat	0.9
q_hat	-0.18248
omega^2	0.282461
p-value	0.001757

Рис. 6. Du Bellay's sonnet 1 + Shakespeare's sonnet 1



1	oh	1	0.999998	1.04E-07
2	thou	2	1.996035	0.000219
3	in	3	2.988148	0.000653
4	hellas	4	3.976373	0.001303
5	deemed	5	4.960745	0.002164
6	of	6	5.9413	0.003236
7	heavenly	7	6.918072	0.004517
8	birth	8	7.891096	0.006004
9	muse	9	8.860407	0.007696
10	formed	10	9.826037	0.009591
11	or	11	10.78802	0.011687
12	fabled	12	11.74639	0.013982
13	at	13	12.70118	0.016475
14	the	14	13.65241	0.019163
15	minstrel's	15	14.60014	0.022045
16	will	16	15.54437	0.02512
17	since	17	16.48515	0.028385
18	shamed	18	17.42251	0.031838
19	full	19	18.35647	0.035479
20	oft	20	19.28707	0.039305
21	by	21	20.21434	0.043315
22	later	22	21.1383	0.047507
23	lyres	23	22.05899	0.05188
24	on	24	22.97643	0.056431
25	earth	25	23.89065	0.06116

n	551
R_n	329
theta_hat	0.675443
q_hat	29.21997
omega^2	0.187592
p-value	0.030878

Рис. 7. The first three stanzas of *Childe Harold's Pilgrimage* by Byron and the first three Shakespeare's sonnets

Another open question is the implementation of the Simon model. Let $R_{n,1}$ be the number of words that occur exactly once. Under the assumptions made, there should be convergence $R_{n,1}/R_n \to \theta$ a.s. (Karlin, 1967). But real texts behave differently. Typically, the number of words that occur once is significantly less than $\widehat{\theta} R_n$.

Simon (1955) proposed the next stochastic model: the $(n+1)$-th word in the text is new with probability p; it coincides with each of the previous words with probability $(1-p)/n$. The drawback of Simon's model is that the number of different words grows linearly. We need some kind of hybrid of an infinite urn model and Simon's model.

Acknowledgements

The research was supported by RFBR grant 19-51-53010. The authors would like to thank Sergey Foss and an anonimous referee for helpful and constructive comments and suggestions.

References

[1] R.R. Bahadur, *On the number of distinct values in a large sample from an infinite discrete distribution*, Proc. Natl Inst. Sci. India, **26A**, Supp. II (1960), 67–75. Zbl 0151.23803

[2] A.D. Barbour, *Univariate approximations in the infinite occupancy scheme*, Alea, **6** (2009), 415–433. MR2576025

[3] A.D. Barbour, A.V. Gnedin, 2009. *Small counts in the infinite occupancy scheme*, Electron. J. Probab., **14** (2009), 365–384. Zbl 1189.60048

[4] A. Ben-Hamou, S. Boucheron, M.I. Ohannessian, *Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications*, Bernoulli, **23**:1 (2017), 249–287. Zbl 1366.60016

[5] M. Chebunin, A. Kovalevskii, 2016. *Functional central limit theorems for certain statistics in an infinite urn scheme*, Stat. Probab. Lett., **119** (2016), 344–348. Zbl 1398.60051

[6] M. Chebunin, A. Kovalevskii, *A statistical test for the Zipf's law by deviations from the Heaps' law*, Sib. Electron. Mat. Izv., **16** (2019), 1822–1832. Zbl 1433.62060

[7] M. Chebunin, A. Kovalevskii, *Asymptotically normal estimators for Zipf's law*, Sankhya, Ser. A, **81**:2 (2019), 482–492. Zbl 1437.62097

[8] G. Decrouez, M. Grabchak, Q. Paris, Finite sample properties of the mean occupancy counts and probabilities, Bernoulli, **24**:3 (2018), 1910–1941. Zbl 1429.60016

[9] O. Durieu, Y. Wang, *From infinite urn schemes to decompositions of self-similar Gaussian processes*, Electron. J. Probab., **21** (2016), Paper No. 43. Zbl 1346.60039

[10] O. Durieu, G. Samorodnitsky, Y. Wang, *From infinite urn schemes to self-similar stable processes*, Stochastic Processes Appl., **130**:4 (2020), 2471–2487. Zbl 1434.60105

[11] M. Dutko, *Central limit theorems for infinite urn models*, Ann. Probab., **17**:3 (1989), 1255–1263. Zbl 0685.60023

[12] I. Eliazar, *The Growth Statistics of Zipfian Ensembles: Beyond Heaps' Law*, Physica (Amsterdam), **390** (2011), 3189.

[13] M. Gerlach, E.G. Altmann, *Stochastic Model for the Vocabulary Growth in Natural Languages*, Physical Review X 3 (2013) 021006.

[14] A. Gnedin, B. Hansen, J. Pitman, *Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws*, Probab. Surv., **4** (2007), 146–171. Zbl 1189.60050

[15] H.S. Heaps, *Information Retrieval: Computational and Theoretical Aspects*, Academic Press, New York etc., 1978. Zbl 0471.68075

[16] G. Herdan, *Type-token mathematics. A textbook of mathematical linguistics*, Mouton and Co., 's-Gravenhage, 1960. Zbl 0163.40904

[17] H.-K. Hwang, S. Janson, *Local limit theorems for finite and infinite urn models*, Ann. Probab., **36**:3 (2008), 992–1022. Zbl 1138.60027

[18] D.C. van Leijenhorst, Th.P. van der Weide, *A Formal Derivation of Heaps' Law*, Inf. Sci., **170**:2-4 (2005), 263–272. Zbl 1070.60009

[19] B. Mandelbrot, *Information Theory and Psycholinguistics*, In: B.B. Wolman and E. Nagel, *Scientific psychology*, Basic Books. 1965
[20] P.T. Nicholls, *Estimation of Zipf parameters*, J. Am. Soc. Inf. Sci., **38**:8 (1987), 443–445.
[21] A.M. Petersen, J.N. Tenenbaum, S. Havlin, H.E. Stanley, M. Perc, *Languages cool as they expand: Allometric scaling and the decreasing need for new words*, Scientific Reports **2** (2012), Article No. 943. https://doi.org/10.1038/srep00943
[22] N.V. Smirnov, *On the ω^2 distribution*, Mat. Sb. n. Ser., **2** (1937), 973–993. Zbl 0018.41202
[23] N. Zakrevskaya, A. Kovalevskii, *An omega-square statistics for analysis of correspondence of small texts to the Zipf–Mandelbrot law*, In: *Applied methods of statistical analysis. Statistical computation and simulation*, Proceedings of the International Workshop, NSTU, Novosibirsk, 2019, 488–494.
[24] G.K. Zipf, *The Psycho-Biology of Language*, Houghton Mifflin, Boston, 1935.

Anik Chakrabarty
Novosibirsk State University,
1, Pirogova str.,
Novosibirsk, 630090, Russia
Email address: a.chakrabarty@g.nsu.ru

Mikhail Georgievich Chebunin
Sobolev Institute of Mathematics,
4, Koptyuga ave.,
Novosibirsk State University,
1, Pirogova str.,
Novosibirsk, 630090, Russia
Email address: chebuninmikhail@gmail.com

Artyom Pavlovich Kovalevskii
Novosibirsk State Technical University,
20, K. Marksa ave.,
Novosibirsk, 630073, Russia
Novosibirsk State University,
1, Pirogova str.,
Novosibirsk, 630090, Russia
Email address: artyom.kovalevskii@gmail.com

Ilya Mikhailovich Pupyshev
Novosibirsk State Technical University,
20, K. Marksa ave.,
Novosibirsk, 630073, Russia
Novosibirsk State University,
1, Pirogova str.,
Novosibirsk, 630090, Russia
Email address: iluxa1@ngs.ru

Natalia Stanislavovna Zakrevskaya
Novosibirsk State Technical University,
20, K. Marksa ave.,
Novosibirsk, 630073, Russia
Email address: natali.erlagol@gmail.com

Qianqian Zhou
School of Mathematical Sciences,
Nankai University,
Tianjin, 300071, China
Email address: qianqzhou@yeah.net