Nonlinear Analysis: Modelling and Control, 2014, Vol. 19, No. 4, 611–625 611
http://dx.doi.org/10.15388/NA.2014.4.7

Empirical Bayes estimators of structural distribution of words in Lithuanian texts

Karolina Piaseckienė[a], **Marijus Radavičius**[b]

[a]Šiauliai University
P. Višinskio str. 19, LT-77156 Šiauliai, Lithuania
karol@delfi.lt

[b]Institute of Mathematics and Informatics, Vilnius University
Akademijos str. 4, LT-08663 Vilnius, Lithuania
marijus.radavicius@mii.vu.lt

Received: 19 August 2013 / **Revised:** 29 March 2014 / **Published online:** 25 August 2014

Abstract. Lithuanian language has great inflexion, free word order and other features which distinguish it from other languages. This raises a problem of testing for the Lithuanian language validity of findings established for other languages. In the paper, an empirical study of a collection of Lithuanian texts is performed. It is supposed that authors of texts are basic elements of the population under study and its heterogeneity stems out of the heterogeneity of preferences and choices of the authors. An attempt to estimate *structural distributions* of words in a collection of texts of different authors is made by making use of a simple statistical model and empirical Bayes approach.

Keywords: structural distribution, Zipf–Mandelbrot law, empirical Bayes, Poisson mixture, sparse data.

1 Introduction

Statistical methods are often used in quantitative linguistic analysis and natural language (NL) processing [1–3], so are probabilistic and statistical models. In *corpus linguistics*, a principal assumption is *randomness* of a corpus (see [4–6]; here randomness is understood as being *equiprobable*). A large corpus, however, contains very heterogeneous data, a mixture of texts of various genres and types (see [7] for empirical evidence), devoted to different purposes and various audiences saying nothing of vocabularies of authors' preferable words (their "mental lexicon" [8]), available and preferable NL structures. How can one in this case to distinguish features of the language itself from preferences of the authors? Thus, an explicit definition of the (finite) population under study is crucial, to a great extent it determines results of statistical analysis. It enables one to get a (representative) sample with appropriate statistical properties and to identify different (ideally independent) sources of (statistical) variation. However it is a problem in (corpus) linguistics. It seems that an implicit definition of the population commonly adopted in

linguistics studies treats *running words* in a corpus as basic elements which satisfy the *randomness condition*. Baroni and Evert [5] is an example of another approach which deals with statistical samples (populations) of *random* text documents. Nevertheless, linguistic features and preferences of authors of large text documents are over-represented in the samples (populations) as compared to authors of small text documents in this case as well.

In our work, authors of text documents are basic elements of population we are interested in. It is supposed that the heterogeneity of a collection of text documents (corpus) stems out of the heterogeneity of preferences and choices of the authors. An attempt to estimate *structural distributions* of words in a collection of texts of different authors is made by making use of a simple statistical model and empirical Bayes approach. Carlin and Luis [9] provide an overview with extensive reference list on empirical Bayes methods. Structural distribution is one of the main objects of study in statistical linguistics closely related to probabilistic approach, in particular to Zipf–Mandelbrot law [10, 11], Yule–Simon law [12, 13] etc., see [2, 14–16] and references therein. The main body of empirical studies of structural distribution and Zipf–Mandelbrot law considers English language corpora, as exceptions we refer to [17, 18].

Utka [18] presents a structural distribution of word frequency counts for a Lithuanian corpus of about 1 million word tokens. Lithuanian language has great inflexion, free word order and other features which distinguish it from other languages, especially English. We refer to [19] for thorough discussion of this topic. This raises a problem of testing for the Lithuanian language validity of findings established for other languages.

In our empirical study, the collection of texts are taken from the digital library collection [20][1] and consists of recommended school imaginative pieces of Lithuanian and foreign authors (novels, stories, poems, poetry, plays), overall 80 text documents of 63 authors with overall 206453 word types (different words in text) and 2567290 word tokens (running words in text).

The results of the empirical study show that the structural distribution estimator obtained as (weighted) average of structural distribution estimators calculated separately for each author in the collection has better performance than the direct structural distribution estimator obtained from the whole collection treated as a corpus.

The next section contains the background material. The definition of structural distribution of words is given and linked up with Herdan–Heaps and Zipf–Mandelbrot law. Section 3 is devoted to empirical Bayes method and its application for estimating a structural distribution. In Section 4, we briefly discuss the results of empirical study of the collection of Lithuanian texts.

2 Latent and structural distributions

Let S be a fixed population of subjects or sources of textual information which are assumed to be statistically independent. We are not interested in content and semantics

[1]The authors are grateful to Education Development Centre, the Ministry of Education and Science of the Republic of Lithuania and EU Structural Support of project "Development of the Key Competencies in Basic School (grades 5–8)".

of the texts but merely in their word types and tokens. Let W_s denote a set of word types in (vocabulary of) a source $s \in S$ and let $V_s := |W_s|$ be its total number of types (the vocabulary size). It is supposed that all vocabularies W_s are subsets of a general vocabulary $\mathcal{W}$. Thus the data we deal with is $\{(y_w(s), x_w(s)),\ w \in \mathcal{W},\ s \in S\}$, where $y_w(s)$ is frequency of a word type $w \in \mathcal{W}$ in a source $s \in S$, $x_w(s)$ is a vector of corresponding explanatory variables and represents some auxiliary information about both the word type $w \in \mathcal{W}$ and its source $s \in S$.

A fundamental assertion in quantitative linguistics states that in principle the vocabulary $\mathcal{W}$ is unbounded (see, e.g., [2, 21]). To put it formally let us introduce an asymptotic parameter $M \to \infty$ which represents overall size of text documents under consideration. For instance, if $\mathcal{W} = \bigcup_{s \in S} W_s$, the parameter M can be taken as $|S|$, the number of sources in S (another alternative is considered below). Then (in theoretical considerations) we require that

$$V = V^{(M)} := V(\mathcal{W}_M) \to \infty, \quad M \to \infty. \tag{1}$$

The sets W_s are treated as samples of size V_s from some (infinite) *superpopulation* of words generated by some stochastic mechanism (cf. [5, 6]). Various models of the word superpopulation Ω (without reference to this term) are proposed and discussed in linguistic literature. Actually, any probabilistic model (e.g., Zipf–Mandelbrot, Yule-Simon, etc.; see [2, 16, 22] and references therein) can be taken as a superpopulation model.

Numerous empirical observations starting from Estoup (1916) (the reference borrowed from [16]) show that up to a half of words in a corpus are hapaxes (*hapax legomena*), i.e. occur in a corpus only once. According to the classical rule of thumb the majority of frequency counts of categories in a frequency table of categorical data are required to be at least 5 [23, p. 396]. If this requirement is violated the accuracy of χ^2 approximation of distributions of test statistics may be insufficient for standard statistical inference and the categorical data (or the contingency table) is said to be *sparse*. Thus, word count data of a corpus is *sparse*. Assumption (1) is referred to as *sparse asymptotics* [24] or *large number of rare events* (LNRE) [15]. Below we describe two basic models of sparse categorical data.

2.1 Latent distribution model

Let us suppose for a while that there is only one source s of textual information so we can drop it from notation. Let word types in a general vocabulary $\mathcal{W}$ of the size $V = V^{(M)}$ be arranged in a certain order r to get $\underline{w} = \underline{w}_V(r) := (w_1, \ldots, w_V)$. The observed and expected word type frequencies, y and $\mu := \mathbf{E}y$, and the explanatory variables x (as well as other related objects) are accordingly arranged giving $\underline{y} := (y_1, \ldots, y_V)$, $\underline{\mu} := (\mu_1, \ldots, \mu_V)$ and $\underline{x} := (x_1, \ldots, x_V)$, respectively.

One of the simplest way to deal with the sparsity is to suppose that $\underline{\mu}$ is determined by a *latent distribution function* F on $[0, 1]$ via representation

$$\mu_i = \mu_+\big(F(t_i) - F(t_{i-1})\big), \quad \mu_+ := \sum_{i=1}^{V} \mu_i,$$

where $t_i := i/V$, $i = 0, 1, \ldots, V$ (cf. [15, 24]). This setting is often used in econometric studies of rankings (ordered data) [25]. It is usually assumed that there exists a bounded (and rather smooth) *latent distribution density* f, $f(u) = \mathrm{d}F(u)/\mathrm{d}u$. The latter assumption implies that the expected frequencies $\mu_i = O(\mu_+/V)$, $V \to \infty$. Let

$$\widehat{V}_m = \widehat{V}_m(s) = \sum_{i=1}^{V} \mathbf{I}\{y_i = m\}, \quad m = 0, 1, \ldots,$$
$$\widehat{V}_+ = \widehat{V}_+(s) := \sum_{i=1}^{V} \mathbf{I}\{y_i > 0\} = \sum_{m=1}^{\infty} \widehat{V}_m$$

be the number of word types $w \in \mathcal{W}$ observed exactly m times and a total number of actually observed word types (in the sampled source s), respectively. Here and below $\mathbf{I}(E)$ is an indictor of an event (set, relation) E. Hence the source s contains $\widehat{V}_+$ word types and y_+ word tokens, $y_+ := \sum_{i=1}^{V} y_i$, typically denoted by N in linguistic literature. Thus it is convenient to take the asymptotic parameter $M := \mu_+ = \mathbf{E}y_+$ since it applies also in case where the number of sources is fixed but the sources themselves are growing. Then $V = V^{(M)} \to \infty \;\Leftrightarrow\; M = M^{(V)} \to \infty$.

The sparsity of data can be characterized via various quantities, for instance,

$$\rho_A = \rho_A(M) := \frac{M}{V^{(M)}}, \qquad \rho_1 = \rho_1(M) := \frac{\mathbf{E}\widehat{V}_1}{\mathbf{E}\widehat{V}_+} \tag{2}$$

are the average expected frequency (the relative expected size) and the relative expected number of hapaxes in textual data, respectively. The quantity ρ_1 is introduced by Khmaladze [15] in order to define models (schemes) of large number of rare events (LNRE).

Definition 1. (See [15].) It is said that the observed frequencies $\underline{y}$ satisfy the LNRE model iff

$$\liminf_{M\to\infty} \rho_1(M) > 0, \quad \mathbf{E}\widehat{V}_+ \to \infty. \tag{3}$$

The LNRE model as well as the condition $\rho_A(N) = O(1)$ can viewed as a formal definition of sparse categorical data.

Remark 1. Latent distribution model assumes existence of an underlying *ordered* random variable r (variable measured at least in the order scale). Thus, it is not directly applicable to *nominal* data, in particular to word types in text documents. To overcome this difficulty one can introduce a certain dummy ordered variable somehow related to the nominal data under consideration. For a data of word counts, common dummy ordered variables are word occurring frequencies in a certain corpus or rankings of words arranged in increasing (decreasing) order of their frequencies in the corpus. Unfortunately, the uniqueness of word types is lost. The corresponding latent distribution of r (appropriately scaled) is called *structural distribution*. A formal definition of the structural distribution is given and discussed in the next subsection.

2.2 Structural distribution

Sometimes it is natural to suppose statistical inference to possess certain symmetry properties, in other words, to be invariant with respect to certain transformations. When dealing with word count data it is reasonable to assume that the order of word occurrence in a text is irrelevant, it is not interesting for a researcher. Thus, in this case, the statistical procedures to be applied should be invariant with respect to word order permutations in a text. It means that any quantitative characteristic of the "population" W can be completely represented by its empirical distribution. In particular, the expected frequencies of word types w in W are represented by their empirical distribution function (e.d.f.)

$$\widehat{F}(u) = \frac{1}{V}\sum_{i=1}^{V} \mathbf{I}\{\mu_i \leqslant u\}, \quad u \geqslant 0.$$

The e.d.f. $\widehat{F}$ is referred to as *empirical structural distribution*. One can expect that e.d.f. $\widehat{F}$ converges (as $V = V^{(M)} \to \infty$), possibly after some scaling, to a distribution function F, say.

Definition 2. (See [26].) Suppose e.d.f. $\widehat{F}(\rho t)$ with a scaling factor $\rho = \rho(M)$ converges weakly to a distribution function F as $M \to \infty$. Then F is called a *structural distribution* of the expected frequencies $\underline{\mu}$ (or simply of the vocabulary W) with the scaling factor ρ.

Under the Poisson sampling model, a value of y chosen at random with the equal probabilities from the observed frequencies $\{y_1, \dots, y_V\}$ satisfies

$$[y \mid \lambda] \stackrel{\mathcal{L}}{=} Poisson(\lambda), \quad \lambda \stackrel{\mathcal{L}}{=} \widehat{F} \quad (\stackrel{\mathcal{L}}{=} \text{ defines a distributions law}), \tag{4}$$

where $Poisson(\lambda)$ denotes the Poisson distribution law with a intensity (or mean) parameter $\lambda > 0$, i.e. $\mathbf{P}(y_i = k \mid \lambda = \mu_i) = \Pi_k(\mu_i) := \mu_i^k \mathrm{e}^{-\mu_i}/(k!)$, $k = 0, 1, \dots$. If the sequence $\rho^{-1}\underline{\mu}_V := (\rho^{-1}\mu_1, \dots, \rho^{-1}\mu_V)$ is a sequence of i.i.d. random variables with a common distribution F, then F obviously is the structural distribution of W with the scaling factor ρ. Thus the word counts distribution determined by (4) can be approximated by the Poisson mixture model

$$[y \mid \lambda] \stackrel{\mathcal{L}}{=} Poisson(\rho\lambda), \quad \lambda \stackrel{\mathcal{L}}{=} F.$$

Church and Gale [7] discuss applications of Poisson mixtures in linguistics.

Empirical studies (see, e.g., [2, 14]) show that better fit to real data is obtained with improper structural distributions. Evert [14] considers following the Zipf's law in the LNRE model with parameters $(\tau, b) \in (0, 1) \times (0, \infty)$. The improper structural density f of this distribution law is

$$f(z) = z^{-\tau-1}, \quad 0 < z \leqslant b < \infty,\ \tau \in (0, 1).$$

The parameter τ determines the *Zipf's exponent* $\alpha = 1/\tau$.

In order to avoid dealing with improper distributions, we can approximate empirical structural distribution by *finite Zipf–Mandelbrot* [14] or *truncated Pareto* distributions $F_\varrho(\cdot \mid \tau, \delta)$ with the density

$$f_\varrho(z \mid \tau, \delta) := \varrho^{-1} f_1\big(\varrho^{-1} z \bigm| \tau, \delta\big), \quad 0 < \delta < 1,\ \tau \in (0, \infty), \tag{5}$$

$$f_1(z) := c_1 z^{-\tau-1} \mathbf{I}\{\delta < z < 1\}, \tag{6}$$

where the scale parameter ϱ, the lower bound δ of the support of f_1 and hence the normalizing constant $c_1 = c_1(\tau, \delta) := \tau(\delta^{-\tau} - 1)^{-1}$ may depend on the asymptotic parameter M. Here we suppose that $\tau > 0$ is fixed,

$$\lim_{M\to\infty} \varrho(M) = b_0 \in (0, \infty], \qquad \lim_{M\to\infty} \delta(M)\varrho(M) = a_0 \in [0, b_0). \tag{7}$$

Assuming the positive lower bound for the support of an approximate structural distribution actually is not a restriction in applications since we never be able to estimate the likelihood of very rare word types whose occurrence probability is below some positive threshold (which depends on M).

In the rest of this subsection, we follow arguments by Evert [14] (see also [21]) to link the structural distribution approximation (5), (6) with Herdan's–Heap's and Zipf's laws. Let $\mathbf{E}_\rho$ ($\mathbf{P}_\rho$) denote the expectation (respectively, probability) with respect to the density f_ϱ. Then the expected number of word types having m counts

$$\mathbf{E}_\varrho \widehat{V}_m = V \mathbf{P}_\varrho\{y = m\} = V \int_0^\infty \Pi_m(u)\, \mathrm{d}F_\varrho(u) = V \frac{c_1 \varrho^\tau}{m!} \int_{\delta\varrho}^{\varrho} u^{m-\tau-1} \mathrm{e}^{-u}\, \mathrm{d}u. \tag{8}$$

Let us denote by $\Gamma(u; \beta)$ the incomplete gamma function and suppose $m \geqslant 1$ and $\tau \in (0, 1)$. Then (7) and (8) imply

$$\frac{\mathbf{E}_\varrho \widehat{V}_m}{V} \sim \frac{c_1 \rho^\tau (\Gamma(b_0; m-\tau) - \Gamma(a_0; m-\tau))}{\Gamma(m+1)}. \tag{9}$$

Analogously, for the expected vocabulary size, i.e. the expected number of word types observed at least once, we have

$$\frac{\mathbf{E}_\varrho \widehat{V}_+}{V} = \mathbf{P}_\varrho\{y > 0\} = \int_0^\infty \big(1 - \mathrm{e}^{-u}\big)\, \mathrm{d}F_\varrho(u) \sim c_1 c_0 \rho^\tau \frac{c_0 \tau (\delta\rho)^\tau}{1 - \delta^\tau} \tag{10}$$

with

$$c_0 = c_0(a_0, b_0, \tau) := \int_{a_0}^{b_0} u^{-(\tau+1)} \big(1 - \mathrm{e}^{-u}\big)\, \mathrm{d}u \sim \int_{\delta\varrho}^{\varrho} \frac{1 - \mathrm{e}^{-u}}{u^{\tau+1}}\, \mathrm{d}u.$$

Note that $a_0 = 0, b_0 = \infty$ yields $c_0 = \Gamma(1-\tau)/\tau$. The average expected frequency (2) of the vocabulary $\mathcal{W}$ generated by the approximate structural distribution $F_\varrho(\cdot \,|\, \tau, \delta)$ defined

in (5), (6) is

$$\rho_A = \frac{\mathbf{E}_\varrho y_+}{V} = \mathbf{E}_\varrho y = \varrho \int_0^\infty u \, \mathrm{d}F_1(u) = c_1 \varrho \int_\delta^1 \frac{\mathrm{d}u}{u^\tau} = \frac{c_1 \varrho \big(1 - \delta^{1-\tau}\big)}{1-\tau}. \tag{11}$$

From (2), (10) and (11) it follows that

$$\mathbf{E}_\varrho \widehat{V}_+ \sim c_0 \frac{1-\tau}{1-\delta^{1-\tau}} M \varrho^{\tau-1}. \tag{12}$$

Take $\varrho \sim c_\varrho M^\beta$, $c_\varrho > 0$, $\beta \in [0,1]$. Then (12) leads to a *power law*

$$\log(\mathbf{E}_\varrho \widehat{V}_+) = \log \frac{c_0(1-\tau)c_\varrho^{\tau-1}}{1-\delta^{1-\tau}} + (1 - \beta + \beta\tau) \log M + o(1). \tag{13}$$

As $M \to \infty$, the law of large numbers imply $\widehat{V}_+ \sim \mathbf{E}_\varrho \widehat{V}_+$ and $N = y_+ \sim \mathbf{E}_\varrho y_+ = M$ in probability. Thus expression (13) can be interpreted as approximate relation between the observed number of word types $\widehat{V}_+$ and the observed number of word tokens N in a large corpus

$$\log(\widehat{V}_+) \approx \mathrm{const} + (1 - \beta + \beta\tau) \log N \tag{14}$$

known as the Herdan's law [27] in quantitative linguistics and as the Heaps' law [28] in information retrieval.

In view of (9) and (10), a *relative frequency spectrum* of the vocabulary W generated by (5), (6)

$$\frac{\mathbf{E}_\varrho \widehat{V}_m}{\mathbf{E}_\varrho \widehat{V}_+} \sim \frac{\Gamma(b_0; m-\tau) - \Gamma(a_0; m-\tau)}{c_0(a_0, b_0, \tau)\Gamma(m+1)} \tag{15}$$

is (asymptotically as $M \to \infty$) independent of the vocabulary size M. Thus the sparsity condition (3) is satisfied. Suppose $a_0 = 0, b_0 = \infty$. Due to well-known approximation $\log(\Gamma(t+h)/\Gamma(t)) = h \log(t+h) + O(t^{-1})$, $t \to \infty$, (15) implies (for very large M)

$$\log(\mathbf{E}_\varrho \widehat{V}_m) = \log(\mathbf{E}_\varrho \widehat{V}_+) - \log \frac{\Gamma(1-\tau)}{\tau} - (1+\tau)\log(m) + O\big(m^{-1}\big), \quad m \to \infty.$$

This expression can be treated as a theoretical variant of *Zipf's second law*:

$$\log(\widehat{V}_m) \approx \log(\widehat{V}_+) - \log \frac{\Gamma(1-\tau)}{\tau} - (1+\tau)\log(m). \tag{16}$$

Here word type counts m are substituted for their ranks r used in *Zipf's first law*, see [21, p. 63].

Other parametric models of structural distributions are discussed in [2, 22], see also [5, 14, 16] and references therein. The (finite) Zipf–Maldebrot model is one of the best-fitting models in real applications.

Khmaladze [15] pointed out that the structural distribution can be treated as a latent mixing distribution in the empirical Bayes approach. In the next section, we present a simple and convenient for computations yet rather informative Bayes statistical model and adopted parametric empirical Bayes approach for estimating expected frequencies of word tokens and their structural distributions.

3 Empirical Bayes approach

Latent and structural distribution models introduced in the previous subsections ignore the available auxiliary information $\{x_w(s),\, w \in \mathcal{W},\, s \in S\}$. The Bayesian model presented here incorporates the auxiliary information by supposing that the conditional distribution of frequency of $y_w(s)$ of a word type w in a source s given a value x of $x_w(s)$ depends on x through scalar functions $p = p(x)$, $\mu = \mu(x)$ and $\kappa = \kappa(x)$. To be precise,

$$\big[y_w(s) \;\big|\; z_w(s) = 0\big] = 0, \tag{17}$$

$$\big[z_w(s) \;\big|\; x_w(s) = x\big] \stackrel{\mathcal{L}}{=} \mathit{Binomial}\big(1, 1 - p(x)\big), \tag{18}$$

$$\big[y_w(s) \;\big|\; z_w(s) = 1,\, \lambda_w(s) = \lambda\big] \stackrel{\mathcal{L}}{=} \mathit{Poisson}(\lambda), \tag{19}$$

$$\big[\lambda_w(s) \;\big|\; x_w(s) = x\big] \stackrel{\mathcal{L}}{=} \mathit{Gamma}\big(\kappa(x), \mu(x)\big). \tag{20}$$

Here $\{z_w(s),\, w \in \mathcal{W},\, s \in S\}$ are latent binary random variables (mutually) conditionally independent when values of the explanatory variables $\{x_w(s),\, w \in \mathcal{W},\, s \in S\}$ are given, $\{y_w(s),\, w \in \mathcal{W},\, s \in S\}$ are random variables (mutually) conditionally independent when values of latent positive random variables $\{\lambda_{sw},\, w \in \mathcal{W},\, s \in S\}$ and values of the explanatory variables are kept fixed, $\mathit{Gamma}(\kappa, \mu)$ denotes the Gamma distribution law with the mean $\mu > 0$, the variance $\kappa\mu^2$ and the distribution density

$$g(u \mid \kappa, \mu) := \frac{u^{1/\kappa - 1} \exp\{-u/(\mu\kappa)\}}{\Gamma(1/\kappa)(\mu\kappa)^{1/\kappa}}, \quad u > 0.$$

The value 0 of the binary latent variable $z_w(s)$ indicates that a word type w is irrelevant (not expected) for a source s, $p(x)$ is the irrelevance probability among cases with $x_w(s) = x$. The latent variable $\lambda_w(s)$ is the expected frequency of a relevant word type w in a source s.

The assumption of mutual conditional independence of y's is not realistic. However, from the viewpoint of asymptotic statistics, it is equivalent to the condition of weak (conditional) dependence of y's which seems to be a rather reasonable assumption when considering word count distributions in sufficiently large text documents.

The marginal (and conditional for given x's) distribution of y's is obtained by integrating out the unobservable random variables z's and λ's

$$\begin{aligned} Q_k(x) &:= \mathbf{P}\big(y_w(s) = k \;\big|\; x_w(s) = x\big) \\ &= p(x)\mathbf{I}\{k = 0\} + \big(1 - p(x)\big) \int_0^\infty \Pi_k(u) g\big(u \;\big|\; \kappa(x), \mu(x)\big)\, \mathrm{d}u, \end{aligned} \tag{21}$$

which actually is a mixture, respectively with the prior probabilities $p(x)$ and $1 - p(x)$, of the degenerate at 0 distribution and the negative binomial distribution q with the mean parameter $\mu = \mu(x)$ and the dispersion parameter $\kappa = \kappa(x)$:

$$q(k \mid \mu, \kappa) := \frac{\Gamma(1/\kappa + k)}{\Gamma(1/\kappa)k!} \left(\frac{\mu}{1+\mu}\right)^k \left(\frac{1}{1+\mu}\right)^{1/\kappa}, \quad k = 0, 1, \ldots.$$

Equations (17)–(20) define a conjugate Bayesian two-component Gamma-Poisson mixture model with mutually independent pairs of unknown parameters $(z_w(s), \lambda_w(s))$, $w \in \mathcal{W}$, $s \in S$, which have the prior distribution

$$\begin{aligned}
&\big[z_w(s) \;\big|\; p_w(s) = p\big] \stackrel{\mathcal{L}}{=} Binomial(1, 1-p),\\
&\big[\lambda_w(s) \;\big|\; z_w(s) = 1,\ \mu_{ws} = \mu,\ \kappa_{ws} = \kappa\big] \stackrel{\mathcal{L}}{=} Gamma(\kappa, \mu),\\
&\big[\lambda_w(s) \;\big|\; z_w(s) = 0,\ \mu_{ws} = \mu,\ \kappa_{ws} = \kappa\big] = 0
\end{aligned}$$

dependent on the *hyperparameters* $p_{ws} := p(x_w(s))$, $\mu_{sw} := \mu(x_w(s))$, $\kappa_{sw} := \kappa(x_w(s))$, $w \in \mathcal{W}$, $s \in S$. Hence the posterior distribution of the unknown parameters based on a sample $y\{D\} := \{y_w(s),\ w \in \mathcal{W},\ s \in D\}$, $D \subset S$, is again the two-component Gamma-Poisson mixture with the updated hyperparameters

$$\hat{p}_{sw} = \hat{p}_{sw}\big(y\{D\}\big) := \frac{p_{sw}\mathbf{I}\{y_w(s) = 0\}}{p_{sw}\mathbf{I}\{y_w(s) = 0\} + (1 - p_{sw})q(y_w(s) \mid \mu_{sw}, \kappa_{sw})}, \tag{22}$$

$$\hat{\mu}_{sw} = \hat{\mu}_{sw}(y\{D\}) := \frac{\mu_{sw}(1 + \kappa_{sw}y_w(s))}{1 + \kappa_{sw}\mu_{sw}}, \tag{23}$$

$$\hat{\kappa}_{sw} = \hat{\kappa}_{sw}(y\{D\}) := \frac{\kappa_{sw}}{1 + \kappa_{sw}y_w(s)}, \quad s \in D. \tag{24}$$

The main problem in Bayesian statistics is the prior distribution specification. In our setting, it means a specification of the hyperparameters p_{sw}, μ_{sw}, κ_{sw}, $w \in \mathcal{W}$, $s \in S$, (parametric approach) or the functions $p(\cdot)$, $\mu(\cdot)$ and $\kappa(\cdot)$ (nonparametric approach). According to *empirical Bayes approach*, the hyperparameters are estimated by fitting the marginal distributions (21) of y's to the available data $y\{D\}$. Assuming special parametric form of the functions $p(\cdot)$, $\kappa(\cdot)$ and $\mu(\cdot)$ allows one to solve this task efficiently. For instance, if $p(\cdot)$ and $\mu(\cdot)$ depend on linear predictors with the logit and logarithmic link functions, respectively, and $\kappa(\cdot)$ is a constant then equations (17)–(20) yield a regression model for K-mixture distributions (see [7, p. 3], also known in econometrics as *zero inflated negative binomial* (ZINB) regression model. Standard statistical software (R, SAS, STATA) can be applied to fit the model. Estimators of the unknown parameters are obtained by the maximum likelihood method and calculated by making use of iteratively re-weighted least squares or/and the EM algorithm.

Given the updated hyperparameters (22)–(24) the structural distribution of word types for a source $s \in S$ can be estimated directly as

$$\widehat{F}_s(u) = \frac{1}{V}\sum_{w \in \mathcal{W}} \big(\mathbf{I}\big\{\hat{\mu}_{ws} \leqslant u,\ y_w(s) > 0\big\} + (1 - \hat{p}_{sw})\mathbf{I}\big\{\hat{\mu}_{ws} \leqslant u,\ y_w(s) = 0\big\}\big). \tag{25}$$

The second summand in this expression estimates the contribution of *unseen* word types for a source $s \in S$. In order to obtain an estimator of the structural distribution of word types of the general vocabulary $\mathcal{W}$, one can take weighted average of structural distribution estimators (25)

$$\widehat{F}_*(u) := \frac{1}{\omega_+}\sum_{s \in S} \omega_s \widehat{F}_s\left(\frac{uN^*}{\hat{\mu}_{*s}}\right) \tag{26}$$

appropriately scaled with

$$\hat{\mu}_{*s} := \widehat{\mathbf{E}N_s} = \int_0^\infty u\widehat{F}_s\,(\mathrm{d}u)$$

to have the same estimated expected text sizes N^*. In empirical study (see the next section), the equal weights and the weights proportional to the source text (vocabulary) size are considered. The structural distribution of the general vocabulary $\mathcal{W}$ can be also estimated directly without intermediate estimation of the structural distributions of text sources:

$$\widehat{F}_W(u) = \frac{1}{V}\sum_{w\in\mathcal{W}} \mathbf{I}\big\{\hat{\mu}_{w+} \leqslant u\big\}, \tag{27}$$

$$\hat{\mu}_{w+} := \sum_{s\in S}\big(\hat{\mu}_{ws}\mathbf{I}\big\{y_w(s) > 0\big\} + (1-\hat{p}_{sw})\hat{\mu}_{ws}\mathbf{I}\big\{y_w(s) = 0\big\}\big). \tag{28}$$

4 Results of empirical study

In the study, the general vocabulary $\mathcal{W}$ is taken as $\bigcup_{s\in S} W_s$. The vector of explanatory variables $x_w(s)$ consists of two categorical variables, s and ℓ_w, and their interactions. The categorical variable $s \in S$ has $|S| = 63$ categories, the categorical variable $\ell_w \in \{2,\dots,10\}$ is the length group of a word type w. The group with $\ell_w = 2$ consists of word types of length 1 or 2, word types in group with $\ell_w = 10$ have 10 or more letters, in the rest groups word type length and group number coincides. We also use a derivative feature *native* indicating whether an author is native Lithuanian or he/she is foreign.

Figure 1 gives an illustration of the Herdan–Heaps law (14), also shows the distribution of text and vocabulary sizes among the authors. The data in log-log scale fits the

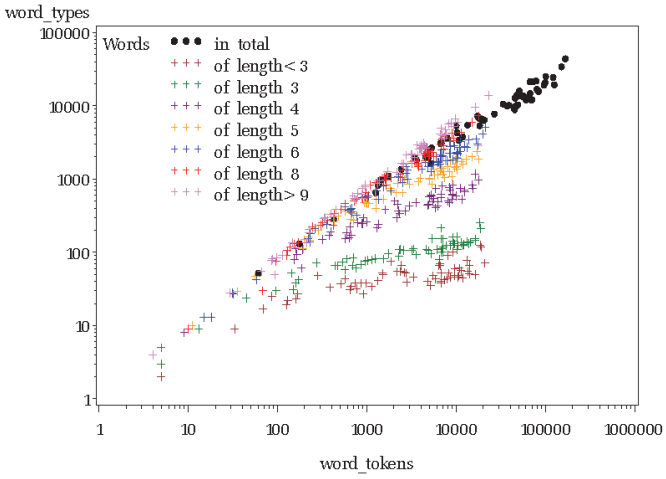

Fig. 1. The Herdan-Heaps law for the words of different length and in total.

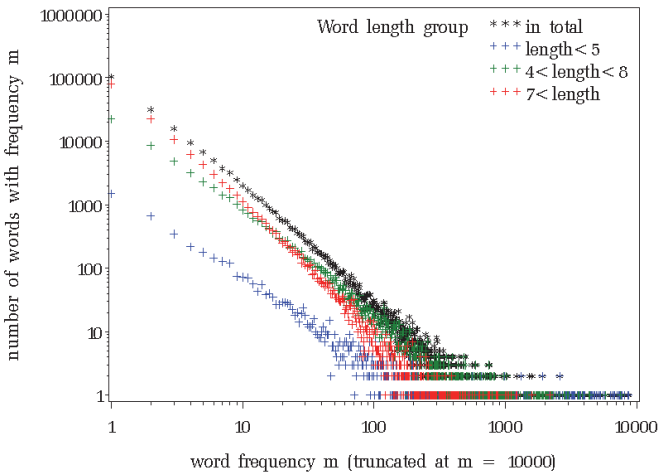

Fig. 2. The Zipf's second law for the words of different length and in total.

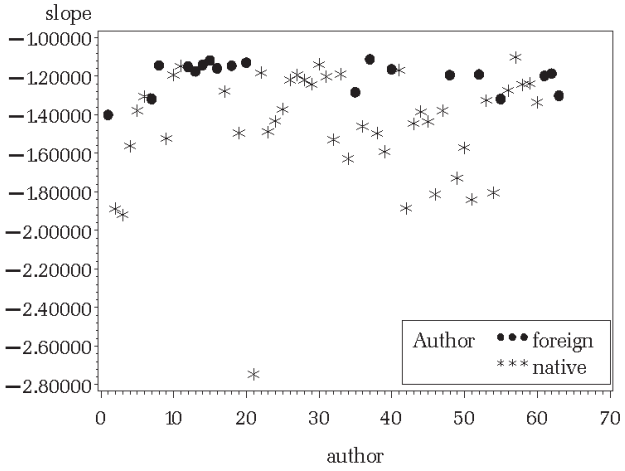

Fig. 3. Scatter plot of slope estimates in the Zipf's second law (16).

straight line very well even for short text documents. However this does not hold for shorter word types (when $\ell_w < 6$). Graphical illustrations of the Zipf's second law (16) are presented in Fig. 2. Again one can notice significant dependence of the slopes in the Zipf's law on the length of word types (more vivid for longer ones with $\ell_w > 7$). Fig. 3 contains scatter plot of estimated slopes in approximate linear equations (16) by formal fitting a linear regression model to each source data (we do not check the model fit and validity). A tendency of slopes for foreign authors as compared with native ones to have smaller absolute values of the slope is apparent. These observations show that choosing $x_w(s) = (s, \ell_w)$ (as an initial step in this direction) is quite reasonable.

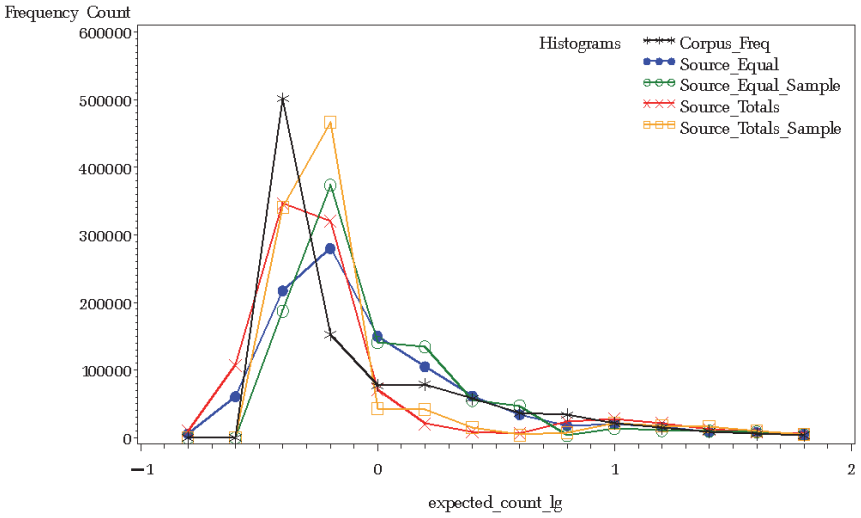

Fig. 4. Histograms of estimates related to the estimator $\widehat{F}_*$, see (26).

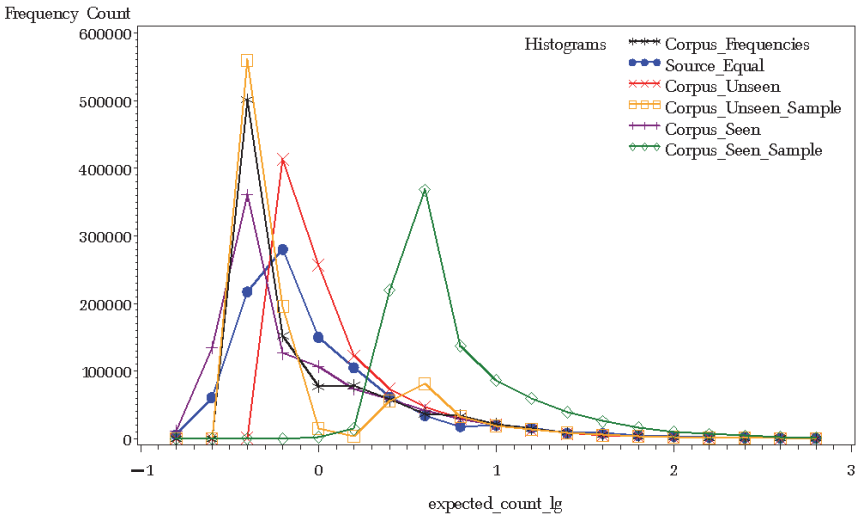

Fig. 5. Histograms of estimates related to the estimator $\widehat{F}_W$, see (27), (28).

The empirical Bayes approach applied to the available data enables us to estimate the number of *unseen word types* in each source and hence respectively adjust estimators of the structural distributions. The effect of the adjustment is apparent in Fig. 4 and Fig. 5, where histograms of structural distribution estimators obtained by making use of different methods are drawn. The structural distribution estimators are respectively

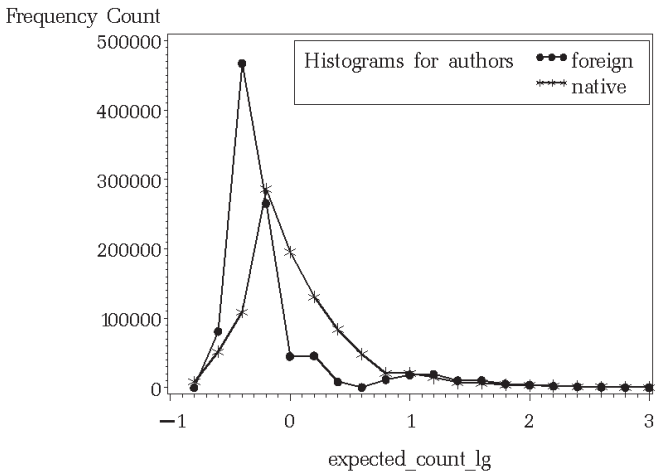

Fig. 6. The structural distribution estimates for native and foreign authors

scaled to match the total $N^* = 10^6$ of word tokens and their histograms are standardized for a text document with vocabulary size of 10^6 word types. The histogram graphs are truncated at a certain value of the expected frequency, either 10^3 or 10^2. *Source_Equal* and *Source_Totals* label the estimates obtained by (26) with the equal weights $\{\omega_s\}$ and the weights proportional to the text size N_s of sources $s \in S$, respectively. The estimates with the weights proportional to the source vocabulary size are very close to that with equal weights and therefore are not presented here. The estimate calculated by (27), (28) is labelled *Corpus_Unseen*, while label *Corpus_Seen* stands for the estimate derived from (27) by deleting the estimated contribution of unseen word types. In addition, histograms of analogous structural distribution estimates but now based on a random subsample $S_9 \subset S$ of size 9 are drawn. This is indicated by adding the word *Sample* to the labels. For comparison, we include the histogram of estimate *Source_Equal* and also the histogram of respectively scaled observed frequencies (labelled *Corpus_Frequences*) in the both figures.

The weighted structural distribution estimates *Source_Equal_Sample* and *Source_Totals_Sample* yield reasonable predictions of the respective estimates *Source_Equal* and *Source_Totals* based on the whole sample S and probably of the true structural distribution F. The estimates *Corpus_Unseen* and *Corpus_Unseen_Sample* seems to be biased in this case. To illustrate the textual data heterogeneity, histograms of *Source_Equal* type estimates of structural distributions for two groups of authors (native vs. foreign) are presented in Fig. 6. Note a more subtle word frequency pattern of foreign authors as compared to scatter plot of slopes in the Zipf's second law in Fig. 3. On the one hand translated texts tend to use more standard vocabulary (reduction of expected frequencies in the interval $(0, 0.8)$ in the $\log_{10}$ scale), on the other hand they contain words related to culture and specific being of other nations and hence rare in original Lithuanian texts (the pike at -0.6).

References

1. S. Abney, Statistical methods and linguistics, in: J.L. Klavans, P. Resnik (Eds.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, MIT Press, Cambridge, 1996, pp. 1–26.
2. R.H. Baayen, *Word Frequency Distributions*, Kluwer Academic Publishers, Dordrecht, 2001.
3. N.A. Smith, *Linguistic Structure Prediction*, Synthesis Lectures on Human Language Technologies, Vol. 13, Morgan & Claypool Publishers, 2011.
4. R.H. Baayen, The randomness assumption in word frequency statistics, in: *Research in Humanities Computing, Vol. 5*, Oxford University Press, Oxford, 1996, pp. 17–31.
5. M. Baroni, S. Evert, Words and echoes: Assessing and mitigating the non-randomness problem in word frequency distribution modeling, in: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (Prague, Czech Republic, June 23–30, 2007)*, pp. 904–911.
6. S. Evert, How random is a corpus? The library metaphor, *Zeitschrift für Anglistik und Amerikanistik*, **54**:177–190, 2006.
7. K.W. Church, W.A. Gale, Poisson mixtures, *Journal of Natural Language Engineering*, **1**:163–190, 1995.
8. A.E. Allahverdyan, W. Deng, Q.A. Wang, Explaining Zipf's law via mental lexicon, 2013, `http://arxiv.org/abs/1302.4383`.
9. B.P. Carlin, T.A. Louis, Empirical Bayes: Past, present and future, *J. Am. Stat. Assoc*, **95**:1286–1289, 2000.
10. B. Mandelbrot, An informational theory of the structure of language, in: W. Jackson (Ed.), *Communication Theory: Papers Read at a Symposium on Applications of Communication Theory, Held at the Institution of Electrical Engineers, London, September 22nd-26th 1952*, Butterworths, London, 1953, pp. 503–513.
11. G.K. Zipf, *The Psycho-Biology of Language: An Introduction to Dynamic Philology*, Houghton Mifflin, Boston, MA, 1935.
12. H.A. Simon, On a class of skew distribution functions, *Biometrika*, **42**:425–440, 1955.
13. G.U. Yule, A mathematical theory of evolution, *Philos. Trans. R. Soc. Lond., Ser. B*, **213**:21–87, 1925.
14. S. Evert, A simple LNRE model for random character sequences, in: *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles (Louvain-la-Neuve, Belgium, 2004)*, pp. 411–422.
15. E.V. Khmaladze, The statistical analysis of large number of rare events, Technical report MS-R8804, Department of Mathematical Statistics, CWI, Amsterdam, 1987.
16. E.V. Khmaladze, Zipf"s law, in: *Encyclopaedia of Mathematics, Supplement III*, Kluwer Academic Publishers, Dordrecht, 2002.
17. L.Q. Ha, D.W. Stewart, P. Hanna, F.J. Smith, Zipf and Type-Token rules for the English, Spanish, Irish and Latin languages, *Web Journal of Formal, Computational and Cognitive Linguistics*, **1**:1–12, 2006.

18. A. Utka, Characteristic features of the most frequent words and wordforms in Lithuanian, *Lituanica*, **61**:48–55, 2005.

19. D. Šveikauskienė, Formal description of the syntax of the Lithuanian language, *Information Technologies and Control*, **34**:245–256, 2005.

20. `http://ebiblioteka.mkp.emokykla.lt`.

21. A. Kornai, How many words are there?, *Glottometrics*, **4**:61–86, 2002.

22. R.H. Baayen, Statistical models for word frequency distributions: A linguistic evaluation, *Comput. Humanities*, **26**:347–363, 1993.

23. A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, Hoboken, NJ, 2002.

24. Y.M.M. Bishop, S.E. Fienberg, P.W. Holland, *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, 1975.

25. M. Aerts, I. Augustynas, P. Jansen, Sparse consistency and smoothing for multinomial data, *Stat. Probab. Lett*, **33**:41–48, 1997.

26. B. van Es, C.A.J. Klaassen, R.M. Mnatsakanov, Estimating the structural distribution function of cell probabilities, *Austrian Journal of Statistics*, **32**:85–98, 2003.

27. G. Herdan, *Quantitative Linguistics*, Butterworths, London, 1964.

28. H.S. Heaps, *Information Retrieval: Computational and Theoretical Aspects*, Academic Press, Orlando, FL, 1978.