The Generalized Poisson Distributions as Models of Word Length Frequencies

Gordana Đuraš[1], Ernst Stadlober[2] and Emmerich Kelih[3]

[1]Institute of Applied Statistics, JOANNEUM RESEARCH, Graz, Austria
[2]Institute of Statistics, Graz University of Technology, Austria
[3]Institute for Slavic Studies, University of Vienna, Austria

Email: gordana.djuras@joanneum.at

Abstract. This study focuses on the analysis of the distribution of word length in Slovenian and Russian texts of different types. Here, word length is measured by the number of syllables per word. Zero-syllable words have not been taken into account in the analysis. The statistical investigation is based on 120 Slovenian and 120 Russian texts. The index of dispersion $d = \frac{s^2}{\bar{x}-1}$ is used as a classifier of texts as follows. In case of $d < 1$ there is under-dispersion, for $d = 1$ equi-dispersion (standard Poisson) and for $d > 1$ overdispersion of word length distributions. As an appropriate alternative for modelling such kind of count data we suggest two-parametric generalizations of the Poisson distribution which allow to model both over- and under-dispersion. Different estimation procedures suggested in literature, such as the method of moments, the maximum likelihood method and an approach based on the sample mean and on the first frequency class, are compared by applying them to our Slovenian and Russian texts under study. In addition, the performance of the parameter estimators is investigated by a simulation study.

Keywords: Count data, Poisson under-/over-dispersion, Parameter estimation tecniques

1 Introduction

An essential question when modeling word length frequencies is how to choose an appropriate probability model to describe the observed values. The final choice depends, however on the interaction of diverse *extra-textua*l factors, such as the concrete language under study, individual authorship or text type, influencing both word length and word length frequencies. Also *theory-driven* factors, in particular the definition of the word and the choice of a measuring unit for its length have to be defined in advance [1]. Taking into account all these boundary conditions allows to reduce the possible set of distributions. Here, word length is measured by the number of syllables where zero-syllable words, typical for Slavic languages, are considered to be a part of the subsequent word and are excluded from the analysis as such.

The simplest and the most widely used distribution for analyzing count data is the Poisson distribution. The equality of mean and variance, known as *equi-dispersion*, is an important property of the Poisson distribution. But this requirement is sometimes too restrictive for count data. In many practical situations empirical data sets exhibit departures from equi-dispersion which can be either *over-dispersion* (the

variance of the count variable exceeds its mean) or *under-dispersion* (the variance is smaller than the mean) with respect to the Poisson model and thus cannot be considered to be fitted by the Poisson distribution. To find out if the count data come from the Poisson distribution we can apply the index of dispersion of a count variable X being defined as the variance to mean ratio [10]. Due to the fact that the texts under study contain no zero-syllable words, 1-displaced versions of the proposed models became relevant for our further research. For this reason, we propose to use the following index of dispersion $\delta = \mathrm{var}(X)/(\mathrm{E}(X) - 1)$ and estimate it by its empirical value $d = s^2/(\bar{x} - 1)$. The 1-displaced Poisson distribution has δ = 1, so it provides an adequate fit only for empirical samples with $d \approx 1$, i.e. for count data where the sample mean diminished by one is near to the sample variance. For this reason, δ can be used as a measure for detecting departures from the 1-displaced Poisson model.

In order to find a general model of word length frequency distributions we construct two-parametric generalizations of the Poisson distributions that cover the whole δ range and are applicable to all dispersion situations.

2 Data Base of the Study

The 120 Slovenian and the 120 Russian texts, serving as data basis for this study include four different text types, namely *journalism*, *poems*, *private letters* and *prose*, thirty texts of each text type being analyzed. These texts were chosen systematically in order to generate a balanced experimental design with an equal number of observations for each text group. The particular selection of the four text types aims to cover the broad textual spectrum and is based on findings from recent word length studies [1, 6]. Table 1 displays the characteristic statistical measures of our text sample: mean word length ($\bar{x}$), sample variance (s^2), text length (TL) and index of dispersion (d).

Table 1. Statistical measures of Slovenian and Russian texts under study

			$\bar{x}$		s^2		TL		d	
	Text Type	N	min	max	min	max	min	max	min	max
SLO	Journalism	30	2.05	2.46	1.22	1.96	328	1166	1.09	1.35
	Poems	30	1.48	1.90	0.37	0.84	58	626	0.60	1.14
	Private Letters	30	1.72	1.98	0.78	0.98	401	1979	0.97	1.15
	Prose	30	1.73	1.98	0.70	1.04	288	4401	0.95	1.16
RUS	Journalism	30	2.40	2.83	1.46	2.17	320	901	1.03	1.34
	Poems	30	1.76	2.40	0.73	1.60	77	1014	0.76	1.41
	Private Letters	30	1.83	2.52	0.90	2.11	48	488	0.79	1.53
	Prose	30	2.02	2.52	1.15	1.83	236	3154	0.93	1.24

A closer look at the distributions of word length in Slovenian and Russian texts shows that texts within one single language, but also between languages, significantly differ with regard to the above statistical measures. Figure 1 visualizes the differences in the mean word length between the two languages and the four text

types given. Obviously, the average word lengths in Russian texts are significantly longer than those in Slovenian texts (Wilcoxon rank-sum test with p<0.01). The linguistic explanation for the apparent differences is that Russian compared to Slovenian has a more complex syllable structure, and tends to have longer word forms on the morphological level, in particular compound words. For further cross linguistic studies of word length in Slavic languages see [7, 8].

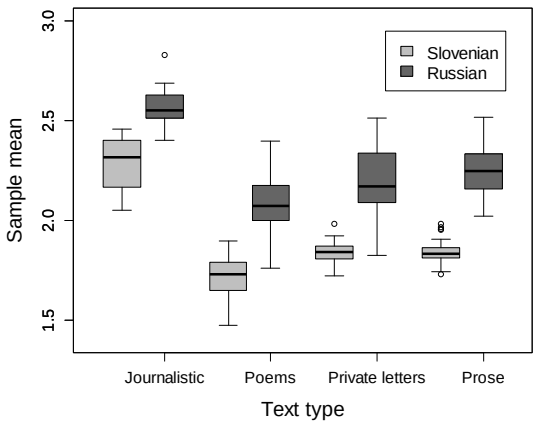

Fig. 1. Differences in word length between languages for different text types

3 Construction of Distributions

In this section we discuss two different approaches for constructing two-parametric generalizations of the Poisson distribution applicable to all three dispersion situations: over-, equi- and under-dispersion.

3.1 The Generalized Poisson Model

Consider the random summation of N mutually independent Borel random variables Y_i with the common probability mass function (pmf) defined by [2, p. 158]

$$P(Y = y) = \frac{(\lambda y)^{y-1} e^{-\lambda y}}{y!}, y = 1,2, \dots \quad (1)$$

and zero otherwise, where $0 < \lambda < 1$. Assuming the number of components N to be a Poisson distributed random variable independent of each Y_i, we obtain the generalized Poisson (GP) distribution as a result. Additionally to Poisson parameter θ we also have parameter λ, which when negative causes some kind of truncation. The probability mass function (pmf) corresponding to the 1-displaced distribution is defined only over positive integers and given by the formula

$$\pi_{x|\lambda,\theta} = P(X = x) = \begin{cases} \dfrac{\theta(\theta + x\lambda - \lambda)^{x-2} e^{-(\theta + x\lambda - \lambda)}}{(x-1)!}, & x = 1, \dots, m+1 \\ 0, & x > (m+1), \text{ for } \lambda < 0 \end{cases} \quad (2)$$

where $\theta > 0, \max(-1, -\theta/m) \leq \lambda < 1$ and m is the largest positive integer such that $\theta + m\lambda > 0$ when λ is negative. Additionally, the condition $m \geq 4$ is proposed in order to ensure that there are at least five non-zero probability classes in the truncated model when $\lambda < 0$ [2, p. 4]. Notice that in the case when $0 \leq \lambda < 1$ the support of the above model does not have to be truncated, hence we have $m = \infty$.

The mean and the variance of the distribution (2) are $\mathrm{E}(X) = \theta/(1-\lambda) + 1$ and $\mathrm{var}(X) = \theta/(1-\lambda)^3$, hence the index of dispersion is $\delta = 1/(1-\lambda)^2$. The parameter λ provides information about the type of the distribution, whereas the parameter θ indicates the intensity of the Poisson process. Obviously, for $\lambda = 0$ the GP model (2) simplifies to the common Poisson model. For $0 < \lambda < 1$ we have $\delta > 1$ (over-dispersion), i.e. the model allows for modeling counts where $s^2 > \bar{x} - 1$ is fulfilled. When $\lambda < 0$ we have $\delta < 1$ (under-dispersion), hence enables to describe $s^2 < \bar{x} - 1$ data cases.

Furthermore, the GP distribution, suitable for over- and under-dispersion, approximates both the negative binomial (NB) and the binomial (B) model. Since the comparison of any two distributions is reasonable only if some common characteristics are fixed, and all investigated distributions are two-parametric, we fix the first two moments or explicitly mean and index of dispersion and express the 1-displaced model parameters in terms of μ and δ as follows:

- $\mathrm{GP}(\lambda, \theta)$: $\mu = \frac{\theta}{1-\lambda} + 1$, $\sigma^2 = \frac{\theta}{(1-\lambda)^3}$, thus $\lambda = 1 - \sqrt{1/\delta}$ and $\theta = (\mu - 1)\sqrt{1/\delta}$
- $\mathrm{NB}(r, p)$: $\mu = \frac{(1-p)r}{p} + 1$, $\sigma^2 = \frac{(1-p)r}{p^2}$, thus $p = \frac{1}{\delta}$ and $r = \frac{\mu-1}{\delta-1}$
- $\mathrm{B}(n, p^*)$: $\mu = np^* + 1, \sigma^2 = np^*(1 - p^*)$, thus $p^* = 1 - \delta$ and $n = \frac{\mu-1}{1-\delta}$

Figures 2 and 3 illustrate differences in the probability distributions among GP, NB and B models for different values of μ and δ. The rows indicate if there is any difference that results from increase in dispersion, whereas the columns describe modifications according to the change in mean. The pmfs of the above distributions are computed by using the ratio of two successive probabilities [5]. Figure 2 shows that for low over-dispersion (when δ is close to 1) there is a negligible difference in the pmf's. When μ increases, the pmf's differ as δ increases. But the difference becomes significant only for very large over-dispersion, not relevant for our study (cf. Table 1).

The difference between the binomial and the corresponding GP distribution (cf. Figure 3) is so small that the two lines overlap in most cases. The slight disagreement is only obvious for δ = 0.25, but vanishes as μ increases. However, such low under-dispersion cases are not of interest here (cf. Table 1).

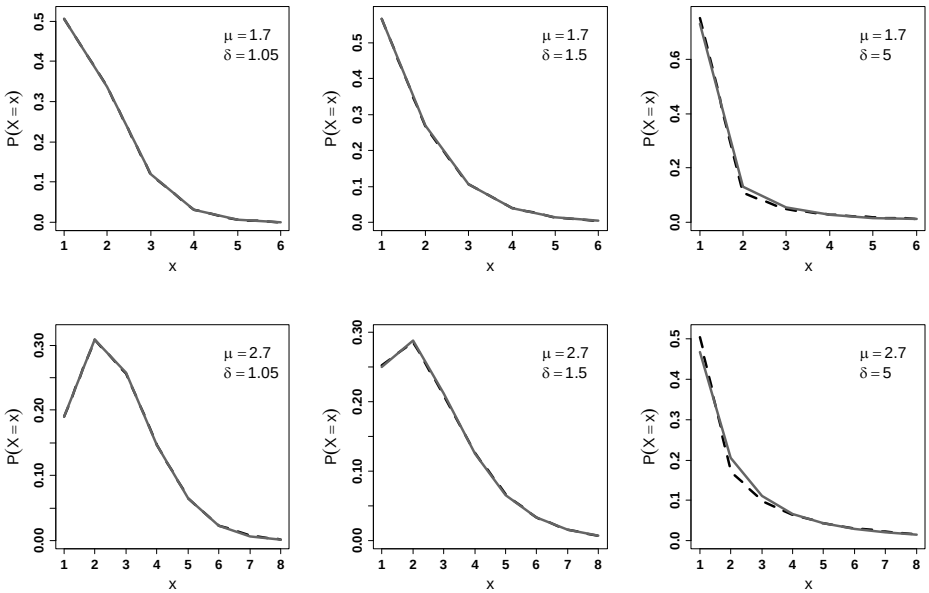

Fig. 2. Comparison of pmf's: GP(λ, θ) (solid line) and NB(r, p) (dashed line) for over-dispersed cases with δ=1.05, 1.5, 5 and mean μ=1.7, 2.7.

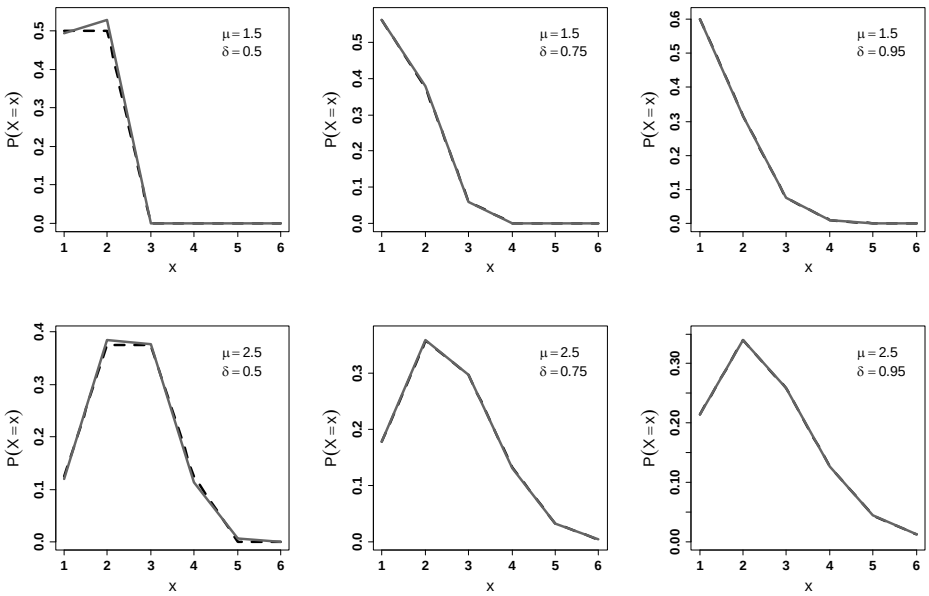

Fig. 3. Comparison of pmf's: GP(λ, θ) (solid line) and B(n, p^*) (dashed line) for under-dispersed cases with δ= 0.5, 0.75, 0.95 and mean μ = 1.5, 2.5.

3.2 The Singh-Poisson Model

Another possibility to construct a two-parametric generalized Poisson distribution is to combine the Poisson with the degenerate (one-point) distribution where the probability mass is cumulated at zero-point [4]. In its 1-displaced form, the pmf of a discrete random variable X having SP distribution is given by

$$\pi_{x|\alpha,\theta} = P(X = x) = \begin{cases} 1 - \alpha + \alpha e^{-\theta}, \ x = 1 \\ \alpha\theta^{x-1}e^{-\theta}/(x-1)!, \ x = 2, 3, \ldots \end{cases} \qquad (3)$$

where $\theta > 0$ and $0 < \alpha \leq \alpha_{\max} = 1/\left(1 - e^{-\theta}\right)$. Here, $\alpha_{\max}$ denotes the maximal possible value of α for given θ and results from the constraint $1 - \alpha + \alpha e^{-\theta} \geq 0$.
The first two moments are given by $\mathrm{E}(X) = 1 + \alpha\theta$ and $\mathrm{var}(X) = \alpha\theta(1 + \theta - \alpha\theta)$, hence the index of dispersion is $\delta = 1 + \theta(1 - \alpha)$. Clearly, under- or over-dispersion is governed only by parameter α, as θ is positive. For $\alpha = 1$ we have equi-dispersion, when $0 < \alpha < 1$ over-dispersion, whereas in case of $1 < \alpha < \alpha_{\max}$ underdispersion with respect to Poisson variation. For further details see [5].

4 Parameter Estimation

To estimate unknown parameters of the introduced models we consider the three most common estimation procedures: method of moments (MM), maximum likelihood method (ML) and estimation based on sample mean and first frequency class (FF).
The moment estimators are obtained by equating the sample moments to the corresponding theoretical counterparts, and are given by:

- for GP: $\hat{\theta}_{\mathrm{MM}} = \sqrt{\frac{(\bar{x}-1)^3}{m_2}}$ and $\hat{\lambda}_{\mathrm{MM}} = 1 - \sqrt{\frac{\bar{x}-1}{m_2}}$
- for SP: $\hat{\theta}_{\mathrm{MM}} = \frac{m_{(2)}}{\bar{x}-1} - 2$ and $\hat{\alpha}_{\mathrm{MM}} = \frac{\bar{x}-1}{\hat{\theta}_{\mathrm{MM}}}$.

The maximum likelihood estimator is the value that maximizes the log-likelihood function which by solving score equations results in

- for GP: $\hat{\theta}_{\mathrm{ML}} = (\bar{x} - 1)(1 - \lambda)$, where $\hat{\lambda}_{\mathrm{ML}}$ is solution of
$$\sum_{i=1}^{k} \frac{f_i(i-2)(i-1)}{(\bar{x}-1) + (i-\bar{x})\lambda} - n(\bar{x} - 1) = 0$$
- for SP: $\hat{\alpha}_{\mathrm{ML}} = \frac{n-f_1}{n\left(1-e^{-\hat{\theta}_{\mathrm{ML}}}\right)}$, where $\hat{\theta}_{\mathrm{ML}}$ is solution of $\frac{\theta(n-f_1)}{n(\bar{x}-1)} + e^{-\theta} - 1 = 0$.

To obtain estimators based on mean and first frequency class we equate the sample mean $\bar{x}$ and the relative frequency of the first class f1/n to the theoretical mean μ and the probability of the first class π_1 and obtain $\hat{\theta}_{\mathrm{FF}} = \log\left(\frac{n}{f_1}\right)$ and $\hat{\lambda}_{\mathrm{FF}} = 1 - \frac{1}{\bar{x}-1}\log\left(\frac{n}{f_1}\right)$ for the GP model. For the SP model it can be shown that $\hat{\alpha}_{\mathrm{FF}} = \hat{\alpha}_{\mathrm{ML}}$ and $\hat{\theta}_{\mathrm{FF}} = \hat{\theta}_{\mathrm{ML}}$.

5 Monte Carlo Simulation Study

We investigate whether the GP model or the SP model performs better for different estimation techniques mentioned above by carrying out a simulation study where all three dispersion situations are considered. Again, we fix the degree of dispersion and the first theoretical moment to obtain the parametrization of the SP model parameters as: $\theta = \delta + \mu - 2$ and $\alpha = (\mu - 1)/\theta$. For the parametrization of the GP model parameters see Section 3.1.

This approach enables us to find out the consequence of applying the wrong model whenever the true parameter values, specified in terms of μ and δ, are known. Additionally, we can receive an impression of how effective the proposed estimation techniques are.

With μ and δ known, the model settings result from above expressions as: (i) for $(\delta, \mu) = (1.34, 2.45)$ we have $(\lambda, \theta_{\mathrm{GP}}) = (0.14, 1.25)$ and $(\alpha, \theta_{\mathrm{SP}}) = (0.81, 1.79)$, (ii) for $(\delta, \mu) = (1.02, 1.85)$ we have $(\lambda, \theta_{\mathrm{GP}}) = (0.01, 0.84)$ and $(\alpha, \theta_{\mathrm{SP}}) = (0.98, 0.87)$, whereas (iii) for $(\delta, \mu) = (0.78, 1.63)$ we have $(\lambda, \theta_{\mathrm{GP}}) = (-0.13, 0.71)$ and $(\alpha, \theta_{\mathrm{SP}}) = (1.54, 0.41)$. The sampling experiments are carried out to produce M = 500 Monte Carlo samples from both models, and diverse dispersion situations, each of size n = 1000. The whole procedure is implemented in the statistical software R. The ML estimation is solved by the Newton-Raphson algorithm for GP model, and for SP by using function **optim()**, available in the software R. For each of the data situations considered here we calculated the mean values of M parameter estimates, as well as the estimated standard errors. To generate GP and SP random variables we use the inversion method [11]. Table 3 summarizes the simulation results obtained. We encountered no difficulty when performing these simulation experiments.

Table 2. Estimation results for data simulated from the GP and SP models

GP(λ, θ)	**λ**		**θ**	
	$\bar{\lambda}_{\mathrm{MM}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{MM}}})$	$\bar{\lambda}_{\mathrm{ML}}(\mathrm{se}_{\bar{\lambda}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$
$\delta > 1$	0.140 (0.022)	0.140 (0.022)	1.251 (0.044)	1.251 (0.040)
$\delta \approx 1$	0.011 (0.022)	0.010 (0.022)	0.840 (0.033)	0.840 (0.033)
$\delta < 1$	-0.130 (0.023)	-0.131 (0.022)	0.710 (0.031)	0.711 (0.030)
SP(α, θ)	**α**		**θ**	
	$\bar{\alpha}_{\mathrm{MM}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{MM}}})$	$\bar{\alpha}_{\mathrm{ML}}(\mathrm{se}_{\bar{\alpha}_{\mathrm{ML}}})$	$\bar{\theta}_{\mathrm{MM}}(\mathrm{se}_{\bar{\theta}_{\mathrm{MM}}})$	$\bar{\theta}_{\mathrm{ML}}(\mathrm{se}_{\bar{\theta}_{\mathrm{ML}}})$
$\delta > 1$	0.812 (0.024)	0.811 (0.019)	1.788 (0.067)	1.789 (0.057)
$\delta \approx 1$	0.984 (0.049)	0.982 (0.043)	0.869 (0.052)	0.870 (0.048)
$\delta < 1$	1.554 (0.133)	1.551 (0.124)	0.410 (0.039)	0.410 (0.037)

In the GP case, all three estimation methods yield similar results, although FF (not shown here) compared to MM has bigger standard errors. Nevertheless, the relative standard errors of $\bar{\theta}$, calculated as $\mathrm{RSE}_{\bar{\theta}} = \mathrm{se}_{\bar{\theta}}/\bar{\theta}$, are much smaller than those of $\bar{\lambda}$, regardless of the estimation method and the dispersion case. The parameter estimates of θ in the SP case are of similar precision as those obtained for the GP model in the over- and equi-dispersed data case, but become more imprecise as soon as the under-dispersed data case is at hand. As to the RSE of the parameter α, we found out that

they are more precise compared to those of the parameter λ. For $\delta > 1$ we obtain 2-4% accuracy, $se_{\bar{\alpha}}$ is 2-3% of α, for $\delta \approx 1$ approximately 4-5%, whereas for $\delta < 1$ around 8-8.5% estimation precision for α.

6 Application to Slovenian and Russian Texts

The results of fitting the Singh-Poisson model to 120 Slovenian texts are displayed in Figure 4, left panel. The solid black line in the graph is the reference bound C = 0.02, whereas the dashed line refers to C = 0.05. Obviously, the SP model provides a good fit for the majority of the texts. It seems to be unsuitable, however, for the three poems of Gregorčić, namely "Kesanje" (text no. 66), "Na sveti večer" (text no.69) and "Pri zibelki" (text no. 76). A closer look at their structure shows that all of them are indeed short texts, having length of 181, 117 and 114 words respectively[1].

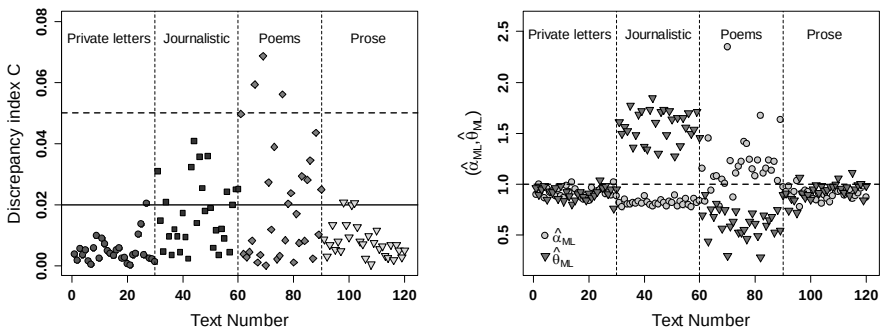

Fig. 4. Results of fitting the Singh-Poisson model to 120 Slovenian texts: discrepancy index (left) and estimated parameter regions (right)

For all 120 Slovenian texts maximum likelihood (ML) estimates of both parameters are computed and each pair of parameters $(\hat{\alpha}_{ML}, \hat{\theta}_{ML})$ is plotted versus the corresponding text, as shown in Figure 4, right panel. It is evident that each group of texts leads to a different pattern of parameters. In the case of private letters both parameters are very close to each other, the same holds for prose texts, although reversed in respect to the order. Contrary to this, in journalistic texts and poems parameters are quite distant from each other. The $\hat{\alpha}_{ML}$ outlier in Figure 4 (right panel) refers to Gregorčić's poem "Njega ni!" (text no. 70). This text has only 106 words, $\hat{\alpha}_{ML}$ = 2.35 and $\hat{\theta}_{ML}$ =

[1] Text length is a general problem in quantitative linguistics and cannot be discussed in detail here. Using texts which are "*complete*" from a lexical and paradigmatical point of view, is in some cases accompanied with the problem that some very short texts (e.g. private letters, poems etc.) have to be analyzed. Obviously the shortness of the text can cause problems in modeling them. In another context it has been shown that text and word length are interrelated in a systematic way [9].

0.3 ($\alpha_{\max}$ = 3.88). Surprisingly, the value of C = 0.0002 indicates an extremely good fit here. The results of fitting the Singh-Poisson model to 120 Russian texts are shown in Figure 5. Notice that for five texts the values of C are beyond the reference line. These texts include again four extremely short private letters by Achmatova written to Brodsky, Chardžiev and Maksimov (texts no. 2, 8, 13 and 20) consisting of 130, 63, 48 and 75 words, respectively, and a poem by Nekrasov "Muza" (text no. 73) with 302 words. The ($\hat{\alpha}_{ML}, \hat{\theta}_{ML}$) parameter regions (cf. Figure 5, right panel) are different from those of the Slovenian texts shown in Figure 4 (right), mostly for private letters, poems and prose.

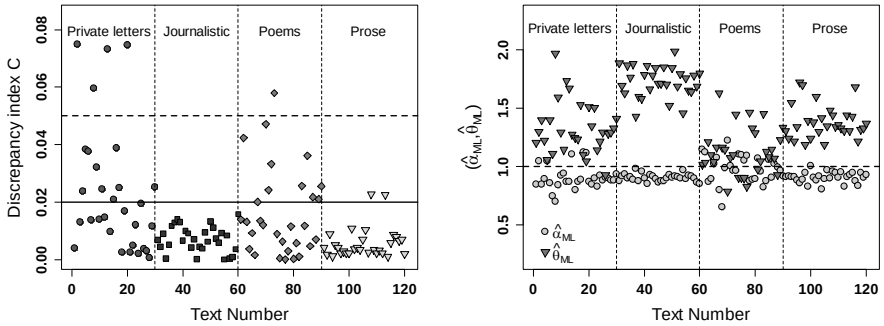

Fig. 5. Results of fitting the Singh-Poisson model to 120 Russian texts: discrepancy index (left) and estimated parameter regions (right)

Plotting the parameters of the SP model versus each other leads to a good discrimination of the three Slovenian text groups, as can be seen in Figure 6, left. Although some Russian texts are located in clearly defined areas, there are many overlappings. However, some general tendency of journalistic texts to build a separate category can still be observed (cf. Figure 6, right). One possible explanation can be found in the fact that the journalistic texts originate from a Russian quality newspaper being as such free of colloquial speech, for which the use of nouns and compound words is quite typical. Also, these types of words are featured by longer word forms, and hence the location of the journalistic texts in the upper area of Figure 6, right, seems plausible. Although determined by the use of poetic meters, the Russian poems cannot be as clearly distinguished from letters and prose texts, as the Slovenian poems do. This quite surprising phenomenon happens possibly due to higher heterogeneity of the Russian poems compared to the Slovenian ones.

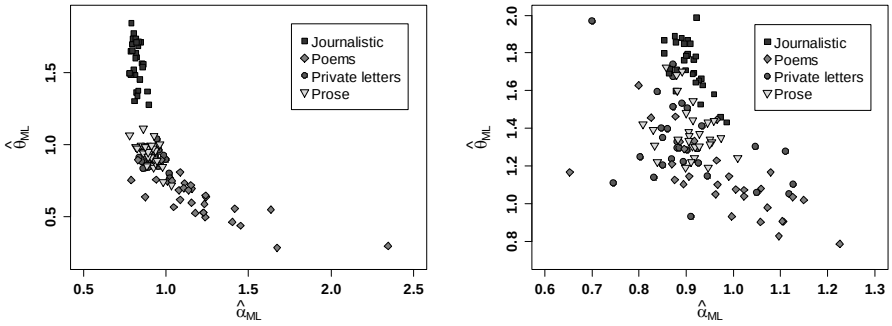

Fig. 6. Text types discrimination by (α, θ) parameter range of Singh-Poisson model of Slovenian (left) and Russian (right) language

Figure 7, left panel, clearly shows that the generalized Poisson model is not appropriate for almost half (46.67%) of the Slovenian journalistic texts. However, the existence of three different parameter patterns, namely for journalistic, poems and joint group of letters and prose texts, is evident from Figure 8, left, regardless of the fact that the model fit for journalistic texts is not satisfactory.

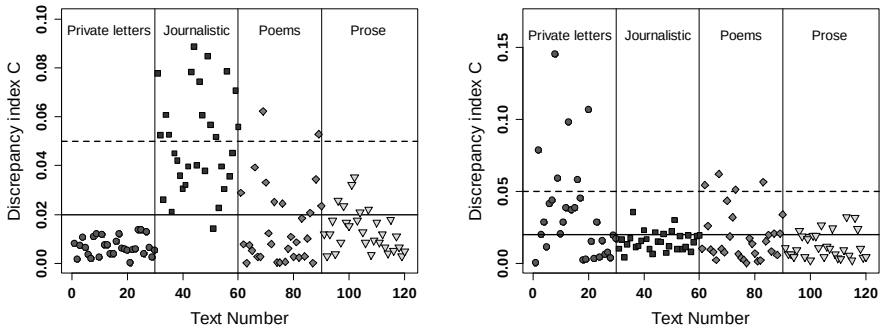

Fig. 7. Results of fitting the Generalized Poisson model to texts under study: Slovenian, left and Russian, right

As compared to this, the Poisson model provides more or less good fits for Russian texts (cf. Figure 7, right panel), with exception of Achmatova short letters no. 2, 8, 9, 13, 16, and 20 of 130, 63, 151, 48, 201, and 75 words respectively, as well as poems no. 62, 73 and 83 of length 157, 302, and 109 respectively. Yet, the discrimination of Russian texts is again not as obvious as it is in the case of the Slovenian text material (cf. Figure 8, right).

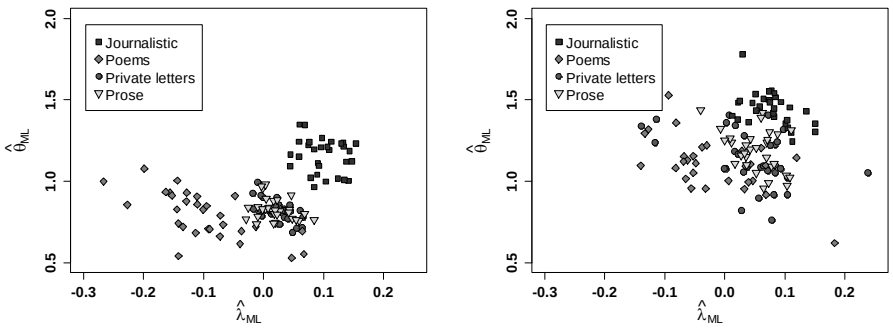

Fig. 8. Text types discrimination by (α, θ) parameter range of the Generalized Poisson model of Slovenian (left) and Russian (right) language

7 Summary

The analyzed Slavic languages, Slovenian and Russian, although belonging to the same family of Indo-European languages, are members of two different subgroups, namely the East Slavic (Russian) and the South Slavic (Slovenian) branch. Throughout history they have shown, however, a rather different development on all linguistic levels, in particular on the phonological, morphological and lexical level. Slovenian, on the contrary to Russian, was for instance lexically and syntactically heavily influenced by other non-Slavic languages, such as German or Italian. Despite these obvious disparities we have proven that the word length frequency distribution of our sampled texts can be described by one theoretical model. This model, known as the Singh-Poisson model, is a simple two-parametric generalization of the Poisson distribution with parameter θ. The additional parameter α tunes the type of dispersion. It allows to model under-dispersion ($1 < \alpha \leq \alpha_{\max}$), equi-dispersion (Poisson case $\alpha = 1$) and over-dispersion ($0 < \alpha < 1$). Therefore, the proposed model offers a unified approach for all dispersion cases. An additional benefit is that the maximum likelihood estimation leads, in case of the Singh-Poisson distribution, to the same estimates as the method based on the sample mean and the first frequency class. For this reason, the calculation of maximum likelihood estimates is a very simple task. In a simulation study we have demonstrated the usefulness of the parameter estimates under three data-driven dispersion scenarios. Finally, the Singh-Poisson model was applied to 120 Slovenian and 120 Russian texts, and in all cases we obtained reasonable and stable estimates.

8 References

1. Antić, G., Kelih, E., and Grzybek, P. (2006). Zero-syllable words in determining word length. In Grzybek, P. (Ed.), *Contributions to the Science of Language*. Dordrecht: Springer, 117 -156.

2. Consul, P.C. (1989). Generalized Poisson Distributions. Properties and Applications. New York: Marcel Dekker, Inc.
3. Consul, P.C., and Famoye, F. (2006). *Lagrangian Probability Distributions*. Boston: Birkhäuser.
4. Đuraš, G. and Stadlober, E. (2010). Modeling word length frequencies by the Singh-Poisson distribution. In Grzybek, P., Kelih, E., and Mačutek, J. (Eds.), *Text and Language. Structures · Functions · Interrelations · Quantitative Perspectives*. Wien: Praesens, 37-48.
5. Đuraš, G. (2012). *Generalized Poisson Models for Word Length Frequencies in Texts of Slavic Languages*. Dissertation, TU Graz.
6. Grzybek, P., Stadlober, E., Kelih, E., and Antić, G. (2005). Quantitative text typology: The impact of word length. In Weihs, C. and Gaul, W. (Eds.), *Classification - The Ubiquitous Challenge*. Heidelberg: Springer, 53-64.
7. Kelih, E. (2009). Slawisches Parallel-Textkorpus: Projektvorstellung von "Kak zakaljalas' stal' (KZS)". In Kelih, E., Levickij, V. V., and Altmann, G. (Eds.), *Methods of Text Analysis*. Chernivtsi: ČNU, 106-124.
8. Kelih, E. (2011). Zum Analytismus und Synthetismus in slawischen Sprachen: Morphologische Wortstruktur in slawischen Sprachen. In Bente Fischer, K., Krumbholz, G., and Lazar, M. (Eds.), *Beiträge der Europäischen Slavistischen Linguistik (Polyslav). Band 14*. München/Berlin: Sagner (Die Welt der Slaven. Sammelbände, Sborniki, 43), 99-107.
9. Kelih, E. (2012). On the dependency of word length on text length. Empirical results from Russian and Bulgarian parallel texts. In Grzybek, P., Naumann, S., and Vulanović, R. (Eds.), *Festschrift für Reinhard Köhler*. Wien: Praesens, 67–80.
10. Puig, P., and Valero, J. (2006). Count data distributions: Some characterizations with applications. *Journal of the American Statistical Association*, 101 (473), 332-340.
11. Stadlober, E. (1989). *Sampling from Poisson, binomial and hypergeometric distributions: Ratio of uniforms as a simple and fast alternative* (Bericht No. 303). Graz: Mathematisch-Statistische Sektion, Forschungsgesellschaft Joanneum.