Word Length and Frequency Distributions

Gordana Antić[1], Ernst Stadlober[1], Peter Grzybek[2], and Emmerich Kelih[2]

[1] Department of Statistics, Graz University of Technology, A-8010 Graz, Austria
[2] Department for Slavic Studies, Graz University, A-8010 Graz, Austria

Abstract. In this paper we study word length frequency distributions of a systematic selection of 80 Slovenian texts (private letters, journalistic texts, poems and cooking recipes). The adequacy of four two-parametric Poisson models is analyzed according their goodness of fit properties, and the corresponding model parameter ranges are checked for their suitability to discriminate the text sorts given. As a result we obtain that the Singh-Poisson distribution seems to be the best choice for both problems: first, it is an appropriate model for three of the text sorts (private letters, journalistic texts and poems); and second, the parameter space of the model can be split into regions constituting all four text sorts.

1 Text base

The relevance of word length studies in general, and for purposes of text classification particularly, has recently been studied in detail and is well documented – cf. Grzybek (ed.) (2005), Antić et al. (2005), Grzybek et al. (2005). On the basis of multivariate analyzes, convincing evidence has been obtained that word length may play an important role in the attribution of individual texts to specific discourse types, rather than to individual authors.

The present study continues this line of research, in so far as the word length frequency distributions of 80 Slovenian texts are analyzed. Yet, this study goes a step further in a specific direction. Most studies in this field, particularly the ones mentioned above, thus far have conducted discriminant analyzes on the basis of characteristics derived from the empirical frequency distributions; in this paper, however, an attempt is made to introduce an additional new aspect to this procedure, by carrying out discriminant analyzes based on the parameters of a theoretical discrete probability model fitted to the observed frequency distribution.

The texts which serve as a basis for this endeavor represent four different text types (private letters, journalistic texts, poems, and cooking recipes), twenty texts of each text type being analyzed. These texts have been chosen in a systematic fashion based on previous insight described in the studies mentioned above. The specific selection of the text sorts has been deliberately made in order to cover the broad textual spectrum, or its extreme realizations, at least. Table 1 represents the composition of the sample.

The paper aims at giving answers to the following questions.

Table 1. Text Sample: 80 Slovenian Texts

AUTHOR	TEXT TYPE	AMOUNT
Ivan Cankar	Private letters	20
Journal *Delo*	Journalistic text	20
Simon Gregorčić	Poems	20
anonymous	Cooking recipes	20
Total		80

a. Can the word length frequency distributions of our sampled texts be theoretically described, and if so, is one discrete probability model sufficient to describe them, or is more than one model needed?
b. Based on the answer to the first set of questions, it is interesting to find out whether one can discriminate the texts by using the parameters of the given model(s) as discriminant variables. In case of a positive answer, this would give us the possibility to attach a certain text to a text group by classifying the parameter values of the fitted model.

Before going into the details, it should be mentioned that word length is measured by the number of syllables per word. Since our texts are taken from a pre-processed corpus (Graz Quantitative Text Analysis Server QuanTAS), the length of a word, defined as an orthographic-phonological unit (cf. Antić et al. 2005) can be automatically analyzed, using specially designed programs.[1]

2 Searching For a Model

In finding a suitable model for word length frequency distributions, an ideal solution for future interpretations of the model parameters would be the existence of a unique model, appropriate for all analyzed texts of the text basis under study. The totality of all texts of a given natural language would be an extreme realization of this procedure.

Furthermore, it is important to find the simplest model possible, i.e., a model with a minimal number of parameters (model of low order). If more than one model is necessary for the description of a particular text sample, it may be important to establish the connections between these models, and to find out whether they can be derived as special cases of one unifying, higher-order model.

[1] The text base is part of the text database developed in the interdisciplinary Graz research project on »Word Length Frequencies in Slavic Texts«. Here, each text is submitted to unified tagging procedures (as to the treatment of headings, numbers, etc.). For details, see `http://www-gewi.uni-graz.at/quanta`.

Due to the fact that we are concerned with words that have at least one syllable, these models will be considered to be 1-displaced. In the subsequent discussion we restrict our study to generalizations of the 1–displaced Poisson distribution, having two parameters each. It is well known that the standard Poisson model with one parameter is able to describe special classes of texts only. So we investigate four different two parametric generalizations which proved to be adequate models for specific texts of several languages (cf. Best 1997): (a) Cohen-Poisson, (b) Consul-Jain-Poisson, (c) Hyper-Poisson, and (d) Singh-Poisson. In order to test the goodness of fit of these probability models, we apply the standardized discrepancy coefficient $C = \chi^2/N$, where N is the text length (number of words in the text). As an empirical rule of thumb we consider the fit of the model (i) as not appropriate in case of $C > 0.02$, (ii) as sufficient if $0.01 < C \leq 0.02$, and (iii) as extremely good if $C \leq 0.01$.

The result of fitting the four models to the 80 members of our text base is given in Figure 1, where geometrical symbols represent the different models. The horizontal line in the graphical display is the reference bound $C = 0.02$.

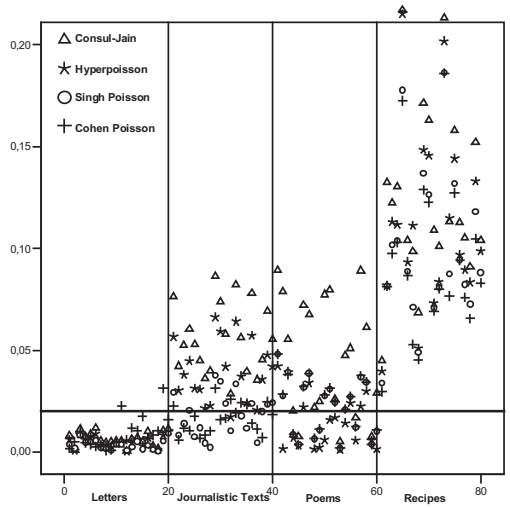


Fig. 1. Results of Fitting Four Two-Parameter Poisson Models to 80 Slovenian Texts

It can be observed that for the text group of recipes, the values of C are far beyond the reference line, in case of all probability models; therefore none of the models is appropriate for recipes.[2]

[2] As more detailed studies have shown, recipes are generally quite "resistent" to modelling, and cannot be described by other models either. This is in line with linguistic assumptions calling into doubt their textual quality.

Additionally, Fig. 1 shows that the Consul-Jain-Poisson model is not appropriate for both journalistic texts and poems. As compared to this, the Cohen-Poisson model provides more or less good fits for private letters, journalistic texts, and poems, but further analyses showed that this model is not able to discriminate journalistic texts from private letters. Consequently, we now restrict our attention to Hyper-Poisson and Singh-Poisson distributions only.

2.1 The 1-Displaced Hyper-Poisson (a,b) Distribution

This distribution has repeatedly been discussed as a model for word length frequency and sentence length frequency distributions. It is a generalization of the Poisson distribution with parameter a, by introducing a second parameter b. In its 1-displaced form, the Hyper-Poisson distribution is given as

$$P_x = \frac{a^{x-1}}{{}_1F_1\,(1;b;a)\,b^{(x-1)}},\quad x = 1,2,3,\ldots\quad a>0, b>0 \tag{1}$$

where ${}_1F_1\,(1;b;a)$ is the confluent hypergeometric series with first argument 1 and $b^{(x-1)} = b\,(b+1)\ldots(b+x-2)$ (cf. Wimmer/Altmann 1999, 281). The first raw and the second central moment of the 1-displaced Hyper-Poisson distribution are

$$\begin{aligned} \mu &= E(X) = a + (1-b)(1-P_1) + 1 \\ Var(X) &= (a+1)\mu + \mu(2-\mu-b) + b - 2\,. \end{aligned} \tag{2}$$

The estimates $\bar{x}$ and m_2' can be used for calculating the unknown parameters a and b as:

$$\begin{aligned} \hat{a} &= \bar{x} - (1-\hat{b})(1-\hat{P}_1) - 1 \\ \hat{b} &= \frac{\bar{x}^2 - m_2' + \bar{x}(1+\hat{P}_1) - 2}{\bar{x}\hat{P}_1 - 1}\,. \end{aligned} \tag{3}$$

Detailed analysis shows that the fits of the Hyper-Poisson distribution to some of the journalistic texts are not appropriate. As listed in Table 2, only five of twenty journalistic texts can be adequately described by the Hyper-Poisson model.

A closer look at the structure of the journalistic texts shows that the frequencies of 2- and 3-syllable words tend to be almost the same; however, a good fit of the Hyper-Poisson model demands rather a monotonic decreasing trend of these frequencies. This may be illustrated by the following two examples. Let us consider two typical journalistic texts from the journal *Delo* (# 29 and # 32). The observed word length frequencies of these two texts are represented in Fig. 2.

For one of the two texts (# 32), we obtain a a good fit ($C = 0.0172$), for the other one a bad fit ($C = 0.0662$). For each of these two texts, we independently simulated ten artificial texts from the Hyper-Poisson distribution

Table 2. Fitting the Hyper-Poisson Distribution to Journalistic Texts

Text	$\hat{a}$	$\hat{b}$	C	Text	$\hat{a}$	$\hat{b}$	C
21	2.14	2.12	0.06	31	2.95	3.17	0.04
22	2.81	3.09	0.03	32	2.85	3.66	**0.02**
23	2.60	3.31	0.04	33	2.06	1.81	0.06
24	2.16	2.40	0.05	34	2.53	2.68	0.04
25	3.02	3.25	0.03	35	3.10	3.67	**0.02**
26	2.11	2.36	0.03	36	2.58	2.77	0.06
27	2.82	3.07	**0.02**	37	2.75	3.52	**0.02**
28	3.04	3.46	**0.02**	38	1.81	1.68	0.04
29	2.26	2.33	0.07	39	2.53	2.52	0.05
30	2.09	1.87	0.06	40	1.82	1.82	0.04

with parameter combinations near to the estimated parameters of the given texts. These ten simulations are plotted in the same graph (see Figure 3) to exhibit the random effect and to study the distributional characteristic of an "ideal" text following the Hyper-Poisson distribution.

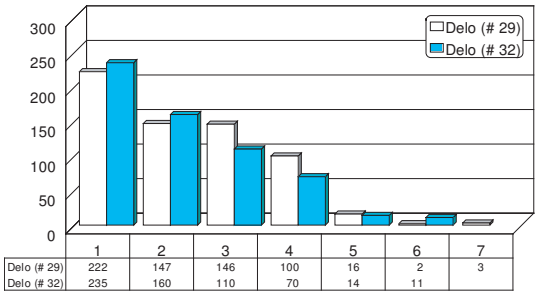


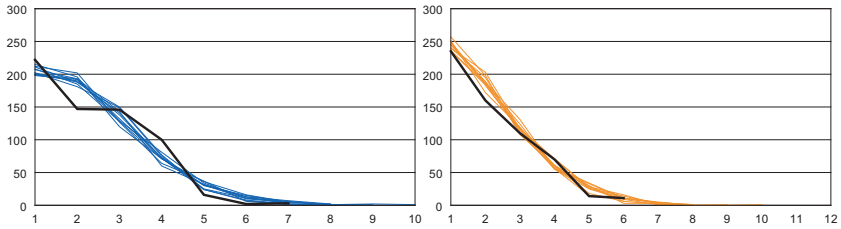
Fig. 2. Word Length Frequencies for Texts #29 and #32 (from *Delo*)

Figure 3 shows that the monotonic decreasing trend is essential for modelling texts with the Hyper-Poisson distribution, but this criterion is not satisfied in case of the text with bad fit ($C = 0.0662$).

2.2 The 1-Displaced Singh-Poisson (a,α) Distribution

The next model to be tested is the 1-displaced Singh-Poisson model, which introduces a new parameter α changing the relationship between the probability of the first class and the probabilities of the other classes. It is given as

$$P_x = \begin{cases} 1 - \alpha + \alpha e^{-a}, & x = 1 \\ \dfrac{\alpha a^{(x-1)} e^{-a}}{(x-1)!}, & x = 2, 3 \ldots \end{cases}$$



(a) Text # 29 ($C \approx 0.07$) (b) Text # 29 ($C \approx 0.02$)

Fig. 3. Simulating Hyper-Poisson Distributions; (a) : (2.26;2.33), (b) : (2.85;3.66)

where $a > 0$ and $0 \leq \alpha \leq 1/(1-e^{-a})$ (cf. Wimmer/Altmann 1999: 605). The first raw and the second central moment of the 1-displaced Singh-Poisson distribution are

$$\mu = E(X) = \alpha a + 1$$
$$Var(X) = \alpha a(1 + a - \alpha a)\,.$$

The estimated parameters $\hat{a}$ and $\hat{\alpha}$ are functions of the empirical moments of the distribution given as

$$\hat{a} = \frac{m_{(2)}}{\bar{x} - 1} - 2, \quad \hat{\alpha} = \frac{(\bar{x} - 1)^2}{m_{(2)} - 2\bar{x} + 2}$$

where $m_{(2)}$ is an estimation of the second factorial moment $\mu_{(2)}$.

The 1-displaced Singh-Poisson model proves to be appropriate for the majority of *private letters* and *journalistic texts*. In case of the *poems*, where the fitting results are less convincing, we obtain $\alpha \approx 1$ for all twenty texts analyzed; this is a clear indication that for poems, even the 1-displaced Poisson model seems to be satisfactory. On the other hand, for the group of *recipes*, this model is not appropriate, due to peculiar relationships between the frequencies: in some cases two or more frequency classes are nearly equal, in other cases there are tremendous ups and downs of frequency classes; the model, however, demands rather monotone relationships between frequency classes.

3 Interpretion of Parameters

Since the 1-displaced Singh-Poisson distribution turns out to be an appropriate model for three of the four text groups (private letters, journalistic texts, and poems), the next step includes an analysis of possible connections between the parameters of this model. Figure 4 represents the results of this analysis as scatter plot: the estimated parameter $\hat{\alpha}$ is represented by circles, the estimated parameter $\hat{a}$ by triangles.

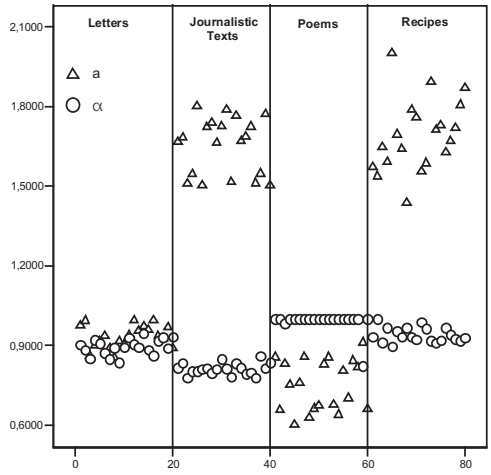


Fig. 4. Parameter Regions of the 1-Displaced Singh-Poisson Model

It is evident that each group of texts leads to a different pattern of the parameters: in case of private letters, both parameters are very close to each other in a very small interval $[0.88; 0.95]$; in case of journalistic texts, opposed to this, they are quite distant from each other, and for poems, their placement on the scatter plot is reversed with respect to the order in the previous two cases as can be seen in Table 3 and Fig. 4.

Table 3. Confidence Intervals for Both Singh-Poisson Parameters

Conf. interval	Letters	Journalistic	Poems	Recipes
$\hat{a}$	$[0.914; 0.954]$	$[1.602; 1.703]$	$[0.705; 0.796]$	$[1.629; 1.756]$
$\hat{\alpha}$	$[0.880; 0.909]$	$[0.801; 0.822]$	$[0.972; 1.009]$	$[0.926; 0.952]$

The parameter values for the recipes are also added in the same plot irrespective of the fact that there is a bad fit. One can observe that they are placed in a specific parameter region. According to a, there is an overlapping of the confidence intervals of journalistic texts and recipes; with respect to α, there is an overlapping of poems and recipes.

However, as shown in Figure 4, both parameters taken together lead to a good discrimination of all four text groups, regardless of the fact that the model fit for recipes is not appropriate.

4 Conclusions

In this study, 80 Slovenian texts from four different text types are analyzed: private letters, journalistic texts, poems, and cooking recipes. In trying to find a unique model within the Poisson family for all four groups, Poisson models with two parameters proved to be adequate for modelling three out of four text types. The relatively simple 1-displaced Singh-Poisson distribution yielded the best results for the first three text groups. However, texts belonging to the group of cooking recipes have a peculiar structure which cannot be modelled within the Poisson family, requiring a certain monotonic relationship between frequency classes. Different texts from a given language (in our case Slovenian) can thus be compared and distinguished on the basis of the specific model parameters.

As an additional result, we demonstrated that, at least in our case, the parameters of the 1-displaced Singh-Poisson distribution are suited to discriminate between all four text sorts. This discrimination yields better results than the other three Poisson models studied.

References

ANTIĆ, G., KELIH, E.; GRZYBEK, P. (2005): Zero-syllable Words in Determining Word Length. In: P. Grzybek (Ed.): *Contributions to the Science of Language. Word Length Studies and Related Issues.* Kluwer, Dordrecht, 117–157.

BEST, K.-H. (Ed.) 1997: *The distribution of Word and Sentence Length.* WVT, Trier. [= Glottometrika; 16]

GRZYBEK, P. (Ed.) (2005): *Contributions to the Science of Language. Word Length Studies and Related Issues.* Kluwer, Dordrecht.

GRZYBEK, P., STADLOBER, E., KELIH, E., and ANTIĆ, G. (2005): Quantitative Text Typology: The Impact of Word Length. In: C. Weihs and W. GAUL (Eds.), *Classification – The Ubiquitous Challenge.* Springer, Heidelberg; 53-64.

KELIH, E., ANTIĆ, G., GRZYBEK, P. and STADLOBER, E. (2005) Classification of Author and/or Genre? The Impact of Word Length. In: C. Weihs and W. GAUL (Eds.), *Classification – The Ubiquitous Challenge.* Springer, Heidelberg; 498-505.

WIMMER, G., and ALTMANN, G. (1999): *Thesaurus of Univariate Discrete Probability Distributions.* Essen.