Grzybek, Peter (ed.):
Contributions to the Science of Text and Language.
In print.

Multivariate Statistical Methods in Quantitative Text Analyses[1]

Ernst Stadlober & Mario Djuzelic

1 Introduction

For the statistical investigation of texts we need the partition of texts into smaller units, their generating elements. These elements may be very heterogenous. The basic, simple elements at the lowest (first) level may be summarized to units which are the elements of the next higher (second) level and these elements of the second level can be unified to elements at the third level and so on. Following this order of ranking, the generating elements at the first level are the graphemes, at the second level we have phonemes, at the third level we get the syllables which constitute the words. From words we come to sentences, then to sections and from sections to chapters. So, in this example we defined seven levels and the question is at which level the quantitative analysis should be carried out.
The outcome of a concrete text appearing as a specific combination of its generating elements is usually the result of a complex decision process influenced by numerous factors. These factors may be authorship and time epoch, genre, functional style and others. For the characterization of texts we use the *word length* (elements of the fourth level) defined by the *number of syllables* (elements of the third level). Each text will be quantitatively described by a number of measures reflecting the moments of the distribution of its word length (mean value m_1, variance m_2, third moment m_3, and the quotients $I = m_2/m_1$ and $S = m_3/m_2$). Additionally, the number of syllables of the text will be defined as the *text length*. Our study was done within the framework of the Graz Project on Word Length (Frequencies) as described in

[1] This work was financially supported by the Austrian Science Foundation (FWF), contract # P-15485

Table 1: Six statistical measures characterizing Slovenian texts

m_1	average word length where word length is the number of syllables per word
m_2	empirical variance of the word length
$I = m_1/m_2$	first criterion of Ord (see Ord, 1967 [**?**])
$S = m_3/m_2$	second criterion of Ord with m_3 the third moment
TLS	text length as number of syllables
$\log(TLS)$	natural logarithm of text length

Grzybek/Stadlober (2002) and is specifically based on the diploma thesis of Djuzelic (2002) considering the following approach. A collection of three categories of texts (literary prose, journalistic prose, and poetry) will be analyzed by means of discriminant analysis to give answers to the following questions. Is it possible to discriminate the texts with the help of the measures mentioned above such that most of the texts can be assigned to the original category? Which measures are the most important ones for suitable discrimination and classification?

The following case study is based on 153 Slovenian texts: 52 (50) texts represent literary (journalistic) prose and 51 texts are poetic texts. Note that we use the same text base as the paper Antić et al. (2004), except one additional journalistic text in our data collection. The appendix of the paper mentioned contains details of these texts in Tables 7, 8 (author, title, chapter, year) and in Tables 9, 10 (statistical measures).

2 Quantitative Measures for the Analysis of Texts

The distribution of the word length of the texts are described by the four variables m_1, m_2, I and S, and the text length is characterized by the two variables TLS which is the length of text in syllables and its logarithm $\log(TLS)$. These two variables will act as *control variables* for our statistical procedures, because the texts were chosen from three groups which differ remarkably according their text length; e.g. the mean text length of literary texts is 4 times longer than the mean text length of journalistic texts, which again is 4 times longer than the mean text length of poetic texts (see Table 2). The definition of the variables used in our analysis are listed in Table 1 below.

Every text in our context is a statistical object carrying its information on $p = 6$ variables. In this way the quantitative description of text j from group i is

Table 2: Statistical values of two Slovenian texts for each group

	Text category	TLS	m_1	m_2	$\log(TLS)$	I	S
1	literary prose	4943	1.89	1.02	8.51	0.54	0.95
2	literary prose	2791	1.93	1.06	7.93	0.55	0.86
	$n_1 = 52, \overline{\mathbf{x}}_1 = ($	4000	1.84	0.96	8.05	0.52	0.90)
1	journalistic prose	1537	2.21	1.75	7.34	0.79	1.09
2	journalistic prose	1200	2.31	1.62	7.09	0.70	0.74
	$n_2 = 50, \overline{\mathbf{x}}_2 = ($	1084	2.25	1.59	6.78	0.71	0.85)
1	poetry	312	1.81	0.72	5.74	0.40	0.50
2	poetry	402	1.75	0.91	6.00	0.52	1.27
	$n_3 = 51, \overline{\mathbf{x}}_3 = ($	270	1.74	0.68	5.41	0.39	0.69)

given by an observation vector of dimension 6

$$\begin{aligned} \mathbf{x}_{ij} &= (TLS(i,j), m_1(i,j), m_2(i,j), \log(TLS)(i,j), I(i,j), S(i,j)) \quad (1) \\ &\quad \text{where} \quad j = 1, \ldots, n_i \,;\, i = 1, 2, 3\,. \end{aligned}$$

For each group i the mean values of the six variables are collected to a mean vector of same dimension:

$$\overline{\mathbf{x}}_i = \left(\overline{TLS}(i), \overline{m_1}(i), \overline{m_2}(i), \overline{\log(TLS)}(i), \overline{I}(i), \overline{S}(i)\right) \,,\, i = 1, 2, 3\,. \qquad (2)$$

An outline of the data with two texts of each category is given in the following Table 2.

2.1 Variance–Covariance Structure of the Variables

The variability of the data is measured by the symmetric variance–covariance matrix S of dimension 6×6. The diagonal elements s_{jj} of this matrix are the empirical variances of the variables and the non-diagonal elements s_{jk}, $j \neq k$, constitute the empirical co-variances between the variables j and k. The elements r_{jk} of the correlation matrix R are obtained from the variance–covariance matrix by the standardization $r_{jk} = s_{jk}/\sqrt{s_{jj}s_{kk}}$. It follows that $-1 \leq r_{jk} \leq 1$ where values near ± 1 (high negative or high positive correlation) indicate a nearly linear relationship between the two variables, and values $r_{jk} \approx 0$ signify that the variables are uncorrelated. The variance-covariance matrix S_1 and the correlation matrix R_1 of the texts in group 1 (literary prose)

Table 3: Variance-covariance and correlation matrix for text category 1: literary prose

$$S_1 = \begin{pmatrix} & TLS & log(TLS) & m_1 & m_2 & I & S \\ TLS & 8664007.55 & 1961.689 & 80.350 & 75.170 & 18.007 & 27.434 \\ log(TLS) & 1961.69 & 0.504 & 0.019 & 0.017 & 0.004 & 0.005 \\ m_1 & 80.35 & 0.019 & 0.004 & 0.006 & 0.002 & 0.001 \\ m_2 & 75.17 & 0.017 & 0.006 & 0.009 & 0.003 & 0.003 \\ I & 18.01 & 0.004 & 0.002 & 0.003 & 0.001 & 0.001 \\ S & 27.43 & 0.005 & 0.001 & 0.003 & 0.001 & 0.007 \end{pmatrix}$$

$$R_1 = \begin{pmatrix} & TLS & log(TLS) & m_1 & m_2 & I & S \\ TLS & 1 & 0.94 & 0.41 & 0.27 & 0.17 & 0.11 \\ log(TLS) & 0.94 & 1 & 0.41 & 0.25 & 0.14 & 0.09 \\ m_1 & 0.41 & 0.41 & 1 & 0.92 & 0.82 & 0.17 \\ m_2 & 0.27 & 0.25 & 0.92 & 1 & 0.98 & 0.33 \\ I & 0.17 & 0.14 & 0.82 & 0.98 & 1 & 0.39 \\ S & 0.11 & 0.09 & 0.17 & 0.33 & 0.39 & 1 \end{pmatrix}$$

are listed in Table 3. There are high correlations between the pairs *average word length* m_1 and quotient $I = m_2/m_1$ ($r = 0.98$) and moments m_1 and m_2 ($r = 0.92$). Rather low correlations appear between the second criterion of Ord (1967), $S = m_3/m_2$ and all other variables.

2.2 Statistical Distance and Linear Discriminant Function

2.2.1 Univariate Statistical Distance

The univariate statistical distance is an important measure for separating data of two different groups of text. It will be assumed that the texts are independent samples $(x_{11}, \ldots, x_{1n_1})$ and $(x_{21}, \ldots, x_{2n_2})$ of two distributions having possibly different theoretical means μ_i, but the same variance σ^2. The theoretical means are estimated by the arithmetic means $\overline{x}_i$ of the samples and the common variance can be estimated by pooling together the two empirical variances s_i^2 of the samples as

$$s_{pool}^2 = \frac{1}{n_1 + n_2 - 2} \left((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 \right) . \tag{3}$$

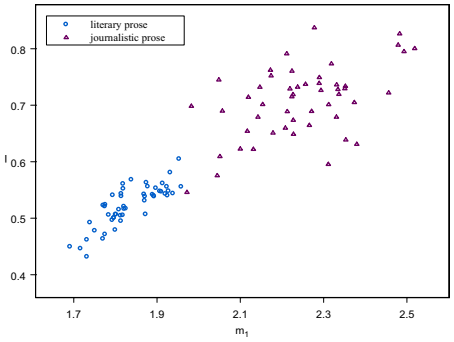
Table 4: Literary prose and journalistic prose: mean values, standard deviations, univariate statistical distances

Variable	**Text type**	$\overline{x}_j^{(1)} \vert \overline{x}_k^{(2)}$	$s_j^{(1)} \vert s_k^{(2)}$	$D(\overline{x}_j^{(1)}, \overline{x}_k^{(2)})$
TLS	literary prose	4000.00	2943.47	1.342
	journalistic prose	1084.20	784.47	
$\log(TLS)$	literary prose	8.05	0.71	1.869
	journalistic prose	6.78	0.64	
m_1	literary prose	1.84	0.07	**3.994**
	journalistic prose	2.25	0.13	
m_2	literary prose	0.96	0.96	0.900
	journalistic prose	1.59	0.20	
I	literary prose	0.52	0.04	**3.606**
	journalistic prose	0.71	0.06	
S	literary prose	0.90	0.09	0.328
	journalistic prose	0.85	0.22	

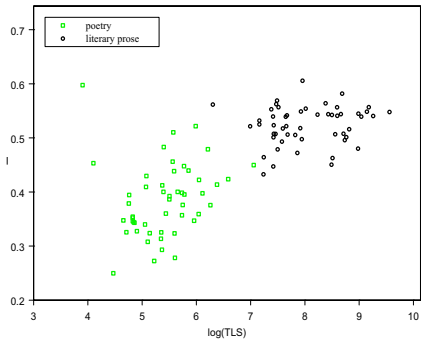
The univariate statistical distance D and the t-value $|t|$ are given as

$$D(\overline{x}_1, \overline{x}_2) = \frac{|\overline{x}_1 - \overline{x}_2|}{s_{pool}}\,, \quad |t| = \sqrt{\frac{n_1\, n_2}{n_1 + n_2}}\, D(\overline{x}_1, \overline{x}_2)\,. \tag{4}$$

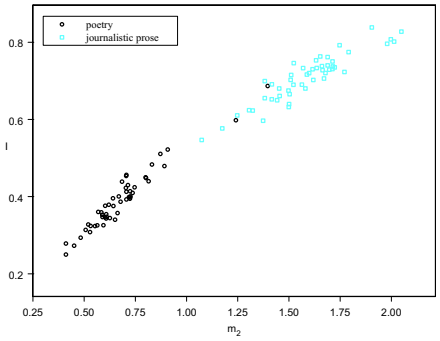
Tables 4, 5 and 6 contain the mean values, standard deviations and univariate statistical distances for all six variables giving the results of all pairwise comparisons according the three categories of text.
The comparison of literary prose and journalistic prose in Table 4 below shows the highest distance values $D \geq 3.6$ according the variables m_1 and I which are also highly correlated. However, the mean values of TLS differ at most, but the large empirical standard deviations keep the statistical distance between the two categories at a lower level.
The scatter plot in Figure 1(a) shows a very high correlation between m_1 and I for texts of type *literary prose* (lower part on the left) and also a high correlation for *journalistic texts* (upper part, right). However, the combination of these two variables results in a good discrimination of the two categories based on the larger values of both m_1 and I for journalistic texts.
Literary prose and *poetry* are discriminated best by the variable $\log(TLS)$ resulting in $D \approx 3.9$. Here a large difference of the mean values is combined with similar standard deviations having low order of magnitude compared to the means (see Table 5). Because of its better distributional properties, the



(a) Scatter plot of the pair (m_1, I) for literary prose and journalistic prose



(b) Scatter plot of the pair $(\log(TLS), I)$ for literary prose and poetry



(c) Scatter plot of the pair (m_2, I) for journalistic prose and poetry

Fig. 1: Scatterplots

Table 5: Literary prose and poetry: mean values, standard deviations, univariate statistical distances

Variable	**Text type**	$\overline{x}_j^{(1)} \vert \overline{x}_k^{(2)}$	$s_j^{(1)} \vert s_k^{(2)}$	$D(\overline{x}_j^{(1)}, \overline{x}_k^{(2)})$
TLS	literary prose	4000.0	2943.47	1.780
	poetry	269.86	1917.46	
$\log(TLS)$	literary prose	8.05	0.71	**3.943**
	poetry	5.41	0.62	
m_1	literary prose	1.84	0.07	1.045
	poetry	1.74	0.12	
m_2	literary prose	0.96	0.96	0.403
	poetry	0.68	0.17	
I	literary prose	0.52	0.04	**2.147**
	poetry	0.39	0.08	
S	literary prose	0.90	0.09	1.126
	poetry	0.69	0.25	

variable $\log(TLS)$ is a more significant measure for discrimination than the untransformed text length TLS. According to this, the only possible discriminator with respect to word length is the first criterion of Ord $I = m_1/m_2$ yielding $D \approx 2.1$.
The scatter plot of $\log(TLS)$ and I in Figure 1(b) illustrates the situation described above: the categories *literary prose* and *poetry* can be discriminated by $\log(TLS)$, but looking at the distribution of the variable I one can observe similar values in both text categories corresponding to a lower value of the statistical distance.
The most interesting results appear in the comparison of *journalistic prose* and *poetry*. Table 6 lists three measures of similar performance ($4.1 \leq D \leq 4.8$) for univariate discrimination where all three are based on word length variables: the variance m_2, the first criterion of Ord $I = m_1/m_2$ and the mean value m_1. For our comparison in Figure 1(c) we selected the most discriminating variables m_2 and I. The perfect linear relationship between these two variables is combined with a good discriminating power for the categories *journalistic prose* and *poetry*.

2.2.2 Multivariate Statistical Distance and Discriminant Function

In the following we will study multivariate observations looking at all $p = 6$ variables simultaneously. The theoretical background of discriminant analy-

Table 6: Journalistic prose and poetry: mean values, standard deviations, univariate statistical distances

Variable	**Text type**	$\overline{x}_j^{(1)} \vert \overline{x}_k^{(2)}$	$s_j^{(1)} \vert s_k^{(2)}$	$D(\overline{x}_j^{(1)}, \overline{x}_k^{(2)})$
TLS	journalistic prose	1084.16	784.47	1.432
	poetry	269.86	191.75	
$\log(TLS)$	journalistic prose	6.78	0.64	2.173
	poetry	5.41	0.62	
m_1	journalistic prose	2.25	0.13	**4.149**
	poetry	1.74	0.12	
m_2	journalistic prose	1.59	0.20	**4.795**
	poetry	0.68	0.17	
I	journalistic prose	0.71	0.06	**4.417**
	poetry	0.39	0.08	
S	journalistic prose	0.85	0.22	0.660
	poetry	0.69	0.25	

sis may be found in the books of Flury (1997) and Hand (1981). A distance measure between two groups of texts based on multivariate observations is a generalization of the univariate case given in (4). It will be assumed that the texts are independent samples of observation vectors $(\mathbf{x}_{j1}, \ldots, \mathbf{x}_{jn_j})$ and $(\mathbf{x}_{k1}, \ldots, \mathbf{x}_{kn_k})$ of two p–dimensional distributions having possibly different theoretical mean vectors $\boldsymbol{\mu}_j$ and $\boldsymbol{\mu}_k$ and the same $p \times p$ variance-covariance matrix Σ. The mean vectors are estimated by the vectors of the arithmetic means $\overline{\mathbf{x}}_j$ and $\overline{\mathbf{x}}_k$. The variance–covariance matrix Σ is estimated by the common empirical variance–covariance matrix S_{jk} obtained by pooling together the two variance-covariance matrices S_k and S_j of the groups as

$$S_{jk} = \frac{1}{n_j + n_k - 2} \cdot ((n_j - 1) \cdot S_j + (n_k - 1) \cdot S_k) \; . \tag{5}$$

The multivariate statistical distance $D(\overline{\mathbf{x}}_j, \overline{\mathbf{x}}_k)$ between the mean vectors $\overline{\mathbf{x}}_j$ and $\overline{\mathbf{x}}_k$ is defined as

$$D_{jk} = D\left(\overline{\mathbf{x}}_j, \overline{\mathbf{x}}_k\right) = \sqrt{(\overline{\mathbf{x}}_j - \overline{\mathbf{x}}_k)' S_{jk}^{-1} (\overline{\mathbf{x}}_j - \overline{\mathbf{x}}_k)} \, , \tag{6}$$

where S_{jk}^{-1} is the inverse of matrix S_{jk} and $\mathbf{x}'$ the transposed vector of $\mathbf{x}$. So, the distance D_{jk} between two groups is defined as the distance between the group centers (means) standardized by the pooled variance-covariance struc-

ture. As numerical values of the distances we get

$$D_{12} = 5.5167\,, \quad D_{13} = 4.7661 \quad D_{23} = 5.4022 \tag{7}$$

which are remarkably higher than the maximal values 3.99, 3.94 and 4.80 of the corresponding univariate distances given in Tables 4, reftab:est5 and 6 above. The variance-covariance matrices S_j, $j = 1, 2, 3$, and the pooled variance–covariance matrices S_{jk} may be found in the diploma thesis of Djuzelic (2002). The discrimination function Y_{jk} is introduced as a linear combination of the p–variables and can be calculated for each p–dimensional observation $\mathbf{x}_{lm}$ of the two groups as

$$Y_{jk}(\mathbf{x}_{lm}) = \mathbf{b}'_{ij}\mathbf{x}_{lm} \quad \text{with vector of coefficients} \quad \mathbf{b}_{ij} = S_{jk}^{-1}(\overline{\mathbf{x}}_j - \overline{\mathbf{x}}_k)\,. \tag{8}$$

The mean values $\overline{Y}^j_{jk}$, $\overline{Y}^k_{jk}$ of the groups, the center m_{jk} of the two groups and the standardized discriminant function Z_{jk} are defined as

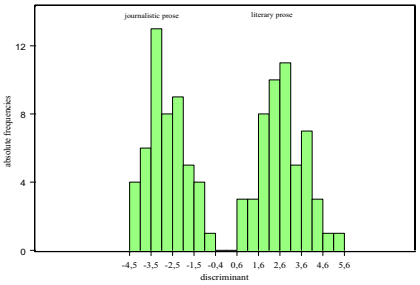
$$\overline{Y}^j_{jk} = Y_{jk}(\overline{\mathbf{x}}_j)\,, \quad \overline{Y}^k_{jk} = Y_{jk}(\overline{\mathbf{x}}_k)\,, \quad m_{jk} = \frac{1}{2}\left(\overline{Y}^j_{jk} + \overline{Y}^k_{jk}\right)\,, \tag{9}$$

$$Z_{jk}(\mathbf{x}_{lm}) = \left(Y_{jk}(\mathbf{x}_{lm}) - m_{jk}\right)/D_{jk}\,. \tag{10}$$

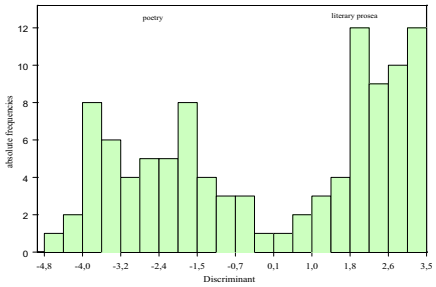
Now each observation vector $\mathbf{x}_{lm}$ can be classified according its value of Z_{jk}. For our data we get the following classification rules:

1. A text is classified as *literary prose* if $Z_{12} > 0$ and $Z_{13} > 0$.
2. A text is classified as *journalistic prose* if $Z_{12} < 0$ and $Z_{23} > 0$.
3. A text is classified as *poetry* if $Z_{13} < 0$ and $Z_{23} < 0$.

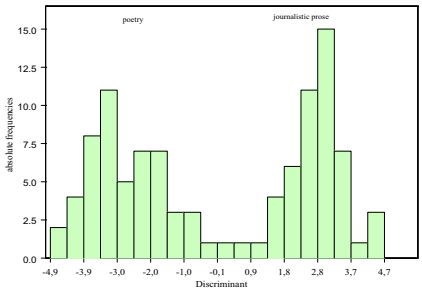
The specific situation is best explained by the histograms of the standardized discriminating variables Z_{12}, Z_{13} and Z_{23} exhibited as Figures 2(a), 2(b) and 2(c). With this graphical displays it is possible to judge the separation power of the discriminant functions. The cut point between two groups is zero as given above. The largest statistical distance $D_{12} = 5.5167$ appears between *journalistic prose* and *literary prose* resulting in a good discrimination by the variable Z_{12} (see Figure 2(a)). The lowest statistical distance of $D_{13} = 4.7661$ is between *poetry* and *literary prose* yielding a weaker potential of Z_{13} for separation (see Figure 2(b)). A slightly better result is obtained in the comparison between *poetry* and *literary prose* where the rather large distance $D_{23} = 5.4022$ implies a good separation of these two groups as can be observed in Figure 2(c).



(a) Separation of journalistic prose and literary prose: histogram of the discriminant Z_{12} with multivariate statistical distance $D_{12} = 5.517$



(b) Separation of poetry and literary prose: histogram of the discriminant Z_{13} with multivariate distance $D_{13} = 4.766$



(c) Separation of poetry and journalist prose: histogram of the discriminant Z_{23} with multivariate statistical distance $D_{23} = 5.402$

Fig. 2: Separations

3 Relevant and Redundant Variables in Linear Discriminant Functions

The linear discriminant functions as defined in (8) are calculated as linear combinations of all $p = 6$ variables. However, there may be some redundancy because of the correlation structure of the variables. Some pairs of variables have high correlations as presented in the correlation matrix of Table 3 for literary prose. It is possible to locate redundant variables in the linear combination by testing the significance of each variable in a stepwise manner. Starting with the whole set of $p = 6$ variables, each variable in the set is tested by calculating the corresponding test statistic which is a Student t statistic with $n_k + n_j - p - 1$ degrees of freedom. If there is at least one redundant variable in the set, i.e. having value $|t| < 2$, then the variable with the smallest $|t|$ value (this is also the variable with the smallest reduction of the statistical distance) is removed from the set. In the next stage the same procedure is carried out on the reduced set with $p' = p - 1$ variables. The procedure terminates when all variables in the remaining set are relevant. This test procedure is demonstrated in Table 7 comparing *literary prose* with *journalistic prose* where the variables S and TLS are identified as redundant variables. Hence the set of 6 variables is reduced to a set of four relevant variables, and this reduction has no impact on the distance function (marginal reduction from 5.5167 to 5.5131).
In the following the reduced linear discriminant functions for all three pairwise combinations are listed. Each combination contains $\log(TLS)$ as relevant variable which had to be expected.

Literary prose and journalistic prose

Reduced linear discriminant function with 4 variables

$$Y_{12}^{red} = 4.52910 \cdot \log(TLS) - 116.36175 \cdot m_1 + 126.8984 \cdot m_2 - 308.88416 \cdot I$$
$$D_{12(red)} = 5.5131 \text{ vs. } D_{12} = 5.5167$$

Literary prose and poetry

Reduced linear discriminant function with 3 variables

$$Y_{13}^{red} = -0.0014 \cdot TLS + 9.0437 \cdot \log(TLS) + 13.6011 \cdot m_2$$
$$D_{13(red)} = 4.7311 \text{ vs. } D_{13} = 4.7661$$

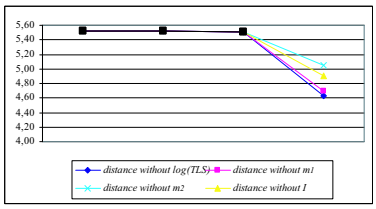
Journalistic prose and poetry

Reduced linear discriminant function with 3 variables

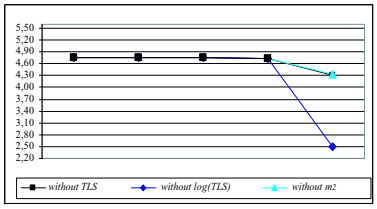
$$Y_{23}^{red} = 3.0937 \cdot \log(TLS) + 22.9766 \cdot m_1 + 39.6065 \cdot I$$
$$D_{23(red)} = 5.3366 \text{ vs. } D_{23} = 5.4022$$

Table 7: Redundant variables S and TLS in Y_{12} (first block), redundant variable TLS in $Y_{12}^{-\{S\}}$ (second block) and no redundant variable in $Y_{12}^{-\{S,TLS\}}$(third block)

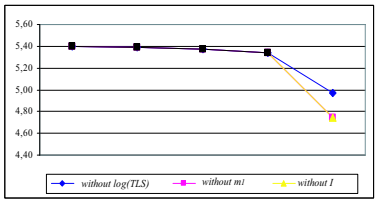
Variable	**coeff.** $b_{12(k)}$	**std.error** $se(b_{12(k)})$	**t-statistic** $t_{12(k)}$-values	**red.distance** $\hat{D}_{12(-k)}$
TLS	0.0002	0.0005	*0.3897*	*5.513*
$\log(TLS)$	4.0731	1.5774	2.5822	5.309
m_1	-117.3995	22.2230	-5.2828	4.757
m_2	129.0193	32.5310	3.9660	5.055
I	-314.3848	68.9248	-4.5613	4.926
S	0.6883	4.7043	**0.1463**	**5.516**
TLS	0.0002	0.0005	**0.3135**	**5.513**
$\log(TLS)$	4.1049	1.5533	2.6427	5.301
m_1	-118.0241	21.6579	-5.4495	4.724
m_2	128.8789	32.3504	3.9838	5.055
I	-312.4976	67.4393	-4.6338	4.914
$\log(TLS)$	4.5291	0.7755	5.8405	4.633
m_1	-116.3618	20.9648	-5.5759	4.697
m_2	126.8984	31.6495	4.0095	5.051
I	-308.8842	66.2722	-4.6608	4.911



(a) Distances for literary prose and journalistic prose



(b) Distances for literary prose and poetry



(c) Distances for journalistic prose and poetry

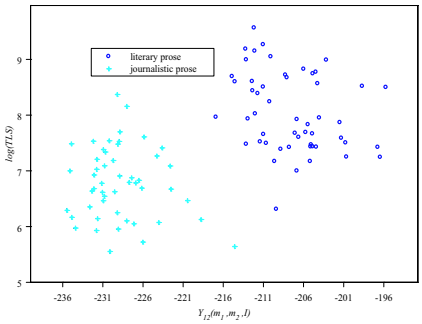
Fig. 3: Distances

Figures 3(a), 3(b) and 3(c) demonstrate the importance of relevant variables for all pairs of categories by comparing the multivariate distances before and after removing the respective variable. The pair *literary prose* and *journalistic prose* may be separated by the variables $\log(TLS)$ and m_1. *Literary prose* and *poetry* can not be discriminated without $\log(TLS)$; *Journalistic prose* and *poetry* differ at most with respect to the word length variables m_1 and I.

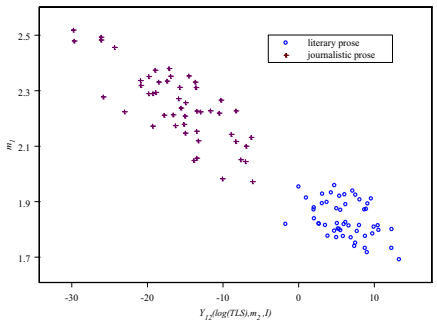
The scatter plots in Figures 4(a) and 4(b) show the values of the relevant variables $\log(TLS)$ and m_1 against the values of reduced discriminant functions (without the variable compared) for the categories *literary prose* and *journalistic prose*. The positive correlation in Figure 4(a) corresponds with a positive coefficient of $\log(TLS)$ in the discriminant function, i.e. the text lengths of the journalistic texts are rather shorter than the text lengths of the literary texts.

Figure 4(b) exhibits strong negative correlation, i.e. the coefficient of m_1 in the discriminant function is negative, and the mean word length of journalistic texts is longer than the mean word length literary texts.

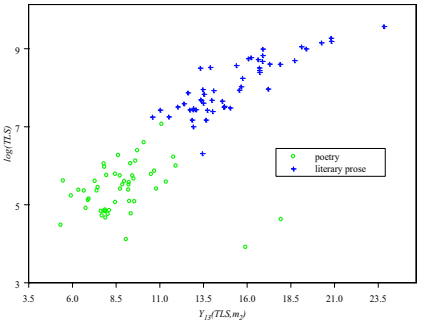
The categories *poetry* and *literary prose* are compared in Figure 4(c) where $\log(TLS)$ is plotted against the reduced discriminant function. The positive correlation implies a positive coefficient for $\log(TLS)$ in the discriminant func-



(a) Scatter plot of the relevant variable $\log(TLS)$ against the discriminant $Y_{12}(m_1, m_2, I)$

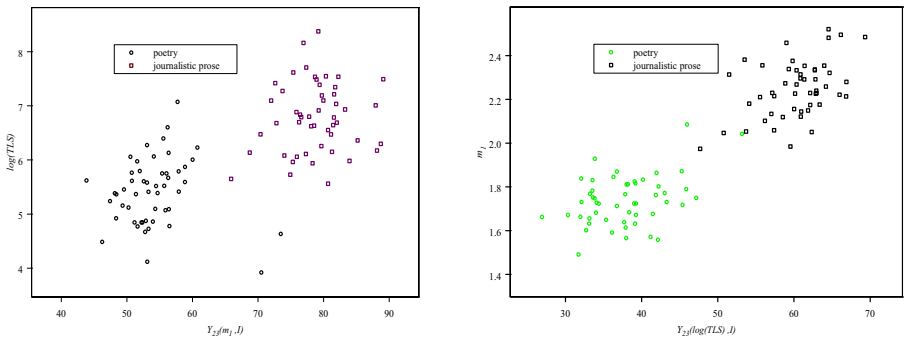


(b) Scatter plot of the relevant variable m_1 against the discriminant $Y_{12}(\log(TLS), m_2, I)$



(c) Scatter plot of the relevant variable $\log(TLS)$ against the discriminant $Y_{13}(TLS, m_2)$

Fig. 4: Scatter Plots



(a) Scatter plot of the relevant variable $\log(TLS)$ against the discriminant $Y_{23}(m_1, I)$

(b) Scatter plot of the relevant variable m_1 against the discriminant $Y_{23}(\log(TLS), I)$

Fig. 5: Scatter Plots

tion. The scatter plot expresses the obvious fact that the poetic texts are shorter than the literary texts.
Figures 5(a) and 5(b) display the values of the relevant variables $\log(TLS)$ and m_1 against the values of the reduced discriminant functions in terms of *journalistic prose* and *poetry*. Positive correlation in Figure 5(a) is connected with a positive coefficient for $\log(TLS)$ in the discriminant function. However, more than 50% of the texts in both categories do not differ according text length.
The effect of m_1 is also positive with a much better separation as before: all, but two poetic texts have smaller values of m_1 than journalistic texts.

3.1 Canonical Discrimination

Our approach of comparing two categories of text can be generalized to a simultaneous comparison of all three categories of text. For this we used a so-called canonical discriminant analysis with the three variables $\log(TLS)$, m_1 and I establishing canonical discriminant functions Z_1 and Z_2. Details of this procedure, together with an SPLUS program may be found in Djuzelic (2002). For a description of statistics with SPLUS we refer to the book of Venables/Ripley (1999).
The first block in Table 8 lists the coefficients of the discriminant functions which are also the components of the eigenvectors of Z_1 and Z_2. The second block contains the mean values and variances of the discriminants Z_1 and Z_2

Table 8: Canonical coefficients for discriminants Z_1 and Z_2

Variable	Z_1	Z_2
$\log(TLS)$	0.33752	-1.40306
m_1	4.66734	4.47832
I	9.51989	-1.82010
text category	**group means ǀ variances of Z_1 and Z_2**	
literary prose	16.25733 ǀ 0.52973	-4.02454 ǀ 0.83067
journalistic prose	19.49542 ǀ 1.20942	-0.74287 ǀ 1.09144
poetry	13.64796 ǀ 1.27444	-0.51754 ǀ 1.08310

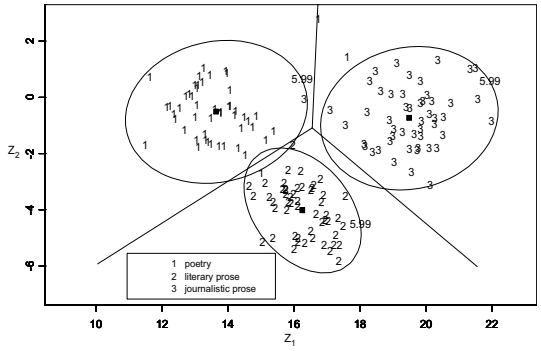


Fig. 6: Canonical discriminant functions with regions of classification for the three categories of text

for each text category.

The eigenvalues $\lambda_1 = 5.77386$ of Z_1 and $\lambda_2 = 2.64693$ of Z_2 express quotients of variances, i.e. the variance between the groups is 5.8 times, respectively 2.6 times higher than the variance within the groups. Hence, both variables Z_1 and Z_2 are good measures for the separation of the categories as can be observed in the scatter plot of Z_1 against Z_2 in Figure 6. The imposed lines partition the (Z_1, Z_2)–plane into the three regions of classification resulting in an excellent discrimination of the text categories: 150 (=98%) of 153 texts are classified correctly.

In detail we have the following. All 52 literary texts are classified correctly (category 3). One of the 50 journalistic texts (category 2) is assigned to cate-

gory 1 (poetry). Only two of the 53 poetic texts are misclassified: one text is classified as journalistic text and one as literary text.
Figure 6 contains also three ellipses of concentration each defined by a quadratic distance of 5.99 from the corresponding group means given in Table **??**.

4 Conclusions

Our case study on three categories of Slovenian texts was a first attempt to study the usefulness of discriminant analysis for the problem of text classification. The major results of our analysis may be summarized as follows.

1. In the univariate setting we calculated for all three pairwise comparisons the univariate statistical distances of six variables: two variables based on text length and four variables based on word length. This gave us first hints of the overall order of discrimination and the order of influence of specific variables.
2. The corresponding analysis of multivariate distances and discrimination functions demonstrated that the correlation structure of the variables may change the role of the variables, e.g. comparing literary prose and poetry the univariate analysis listed variable I as important, but variable m_2 as unimportant. In the multivariate analysis we ended up with m_2 as relevant and I as redundant. (This special effect is caused by the high correlation of the variables.)
3. We established a linear discriminant function for the pair (literary prose | journalistic prose) with four relevant variables. For the two other pairs (literary prose | poetry) and (journalistic prose | poetry) only three relevant variables appear in each discriminant function.
4. Both types of variables were relevant for discrimination: variables for text length as well as variables for word length.
5. Canonical discrimination of all three text categories with the variables $\log(TLS)$, m_1 and I was able to classify 98% of the texts correctly.
6. Our future research will be concentrated on the following considerations. Different categories of texts from various Slavic languages will be studied by classification methods to find combinations of discriminating variables based on word length only. For this we prepared a large collection of variables, i.e. statistical parameters describing word length. Our hope is to establish suitable classification rules for at least some interesting categories of texts.

References

Antić, G.; Kelih, E.; Grzybek, P. (2004): "Zero-syllable Words in Determining Word Length". In: Grzybek, P. (ed.), *Contributions to the Science of Text and language. Word Length Studies and Related Issues*.

Djuzelic, M. (2002): *Einflussfaktoren auf die Wortlänge und ihre Häufigkeitsverteilung am Beispiel von Texten slowenischer Sprache.* Diplomarbeit, Institut für Statistik, Technical University Graz.

Flury, B. (1997): *A First Course in Multivariate Statistics.* New York.

Grzybek, P.; Stadlober, E. (2002): "The Graz Project on Word Length (Frequencies)", in: *Journal of Quantitative Linguistics, 9(2)*; 187–192.

Hand, D. (1981): *Discrimination and Classification.* New York.

Ord, J.K. (1967): "On a System of Discrete Distributions", in: *Biometrika, 54*, 649–659.

Venables, W.N.; Ripley, B.D. (1999): *Modern Applied Statistics with S-Plus.* 3^{rd} Edition, New York.