Text and Language

Structures · Functions · Interrelations
Quantitative Perspectives

Edited by

Peter Grzybek
Emmerich Kelih
Ján Mačutek

prae
sens

Peter Grzybek
Emmerich Kelih
Ján Mačutek
(eds.)

Advisory Editor
Eric S. Wheeler

Text and Language

Structures · Functions · Interrelations.
Quantitative Perspectives

SONDERDRUCK

prae
sens

Statistical reduction of the feature space of text styles

Vasilij V. Poddubnyj, Anastasija S. Kravcova

1 Introduction

The style of a text is characterized by a random feature set that can include syntactic words, high frequency words, bigrams, etc. Every feature is measured by a relative frequency of the occurrence in the text. These frequencies specify the feature space of text styles. Every frequency set can be presented geometrically as a point in a multidimensional feature space. A number of different texts form a point "cloud", or a text scatter plot. However, these features are not of the same value. Some features describe better the style of the author or genre: they have greater frequency variance and better distinguish texts of different authors or genres. Others have smaller frequency variance and less discrimination. Besides there are some "noise" features. In most cases, these features are statistically related to each other. This means that a random feature set has redundancy. This paper considers the transformation of the feature space that allows one to find a minimal set of statistically independent latent features.

2 Principal component analysis

A widely used statistical method of feature space transformation is that of Principal Components Analysis – PCA (Afifi and Azen 1979). This method consists of the orthogonal linear transform of data to a new coordinate system in which the greatest variance of any projection of the data lies on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. As a result, new factors (principal components) are uncorrelated, and the first few components almost completely define the whole scatter of points; so the components with small variances can be omitted. In the space of the first two principal components, the scatter of text-points is maximal. New features (factors) are defined by factor loadings which are the coefficients at the initial factors. Principal component analysis requires only regularity of the correlation matrix of frequency features. The frequency distribution may be arbitrary and not necessarily Gaussian. However, the probabilistic approach to principal component analysis is substantially based on normal feature distribution (Lawrence 2005). Normalization of features requires a nonlinear transformation of the initial feature space.

3 Discriminant analysis

Another method of dimensional reduction is that of Discriminant Analysis (cf. Klecka 1980, Kendall and Stuart 1979. This method consists of a linear transformation of the coordinates of a feature space which leads to the maximization of the discrepancy of the average values of new features in different classes. The deviations of new features from their average values are uncorrelated and have equal variances within the classes. In the case of text analysis, the classes are the groups of texts that differ either in author, or in genre, or in gender of the author, or in age of the author, etc. Hence, the number of classes equals the number of authors, genres, etc. The direction of the first axis of the new feature space (coordinate axis of the first discriminant functions – DF) is chosen so that the centers of classes have maximum difference from each other on this axis (for the first DF). The second axis (coordinate axis of the second DF) is directed at a right angle to the first axis so that centers of classes have maximum difference from each other on this axis (for the second DF). The third axis is directed at a right angle to the plane of the two above mentioned axes, etc. The dimension of the new feature space (of DF) is less than the lesser of the dimension of the initial space and the number of classes minus one. The discrimination property of discriminant function decreases monotonically as the number of DF grows (in the space of the first two DF, the centers of classes differ from each other in maximal degree).

4 Ranking and normalization of frequency features of text style

Formally, discriminant analysis does not require the feature distribution to be normal, the same as principal component analysis. But, it needs non-degeneracy of the correlation feature matrices within and between the classes. However, evaluation of the quality of the discriminant function method (statistical significance of DF) is based on normality of features distribution. The normalization of features presumes a proper nonlinear transformation of the initial feature space (reduction to the Gaussian distribution).

Most methods for solving discrimination, classification, and recognition problems (such as discriminant analysis, Bayes classification, recognition methods, etc.) are based on the normal (Gaussian) feature distribution (Klecka 1980, Kendall and Stuart 1979). At the same time, relative frequencies of the initial feature system of the text style not always correspond to normal distribution. By this reason the application of the well-known parametric methods of mathematical statistics to text analysis is questionable.

As for the implementation, these methods are not always mathematically correct. Therefore two approaches are possible. The first approach consist of developing non-parametric (distribution-free) methods of discriminant analysis, classification, and recognition. The second approach is to find a nonlinear transformation of the absolute and relative frequencies of initial features that ensures the normality of both the feature distribution and the principal component and discriminant functions related to them.

This paper proposes a method of the second approach. Let us consider an ordered series of the relative frequencies for each feature in the analyzed text. Let n texts of different (in general) volumes N_i, $i = 1, \dots, n$, be examined. We select m features of text style (for example, m syntactic words or bigrams). Each j-th feature ($j = 1, \dots, m$) occurs in the i-th text v_{ij} times. The numbers v_{ij} are absolute frequencies of the occurrence of the j-th feature in the i-th text and can be presented in a table where the columns correspond to the features and the rows to the texts. It is obvious that the sum of absolute frequencies v_{ij} gives the whole number v_i of occurrence of the features set in the i-th text: $\sum_{j=1}^{m} v_{ij} = v_i$, $i = 1, \dots, n$. Then $p_{ij} = v_{ij}/v_j$ is the relative frequency of the j-th feature in i-th text; $\sum_{j=1}^{m} p_{ij} = 1$ for all $i = 1, \dots, n$. Thus the relative frequencies show the relative parts of features and assume values in the interval from 0 to 1, so they cannot be modeled in general by the normal distribution. The set of frequencies in the i-th row (the i-th text) forms a vector-row that specifies the coordinates of the i-th point-text in the feature space. We order the relative frequencies of each j-th feature in all the texts (across the j-th column, the j-th sample) in ascending order. The place of each element of a sample in the ordered series is called its rank. Thus, the vector-column $p_j = (p_{1j}, p_{2j}, \dots, p_{ij}, \dots, p_{nj})^T$ of relative frequencies of the j-th feature corresponds to the column-vector $r_j = (r_{1j}, r_{2j}, \dots, r_{ij}, \dots, r_{nj})^T$ of their ranks, $j = 1, \dots, m$. It will be noted that equal frequencies must have the same rank which is the arithmetic mean value of ranks in a bunch of equal frequencies. In this case, the row-vector $p_i = (p_{i1}, p_{i2}, \dots, p_{ij}, \dots, p_{im})$ of relative features frequencies will be matched by the row-vector $r_i = (r_{i1}, r_{i2}, \dots, r_{ij}, \dots, r_{im})$ of their ranks, $i = 1, \dots, n$.

It is a well known fact (Hollander and Wolfe 1999) that, under general conditions, the ranks have the uniform probability distribution in the interval from one to the sample size n. It follows from the fact that the empirical integral distribution function of ranks is the uniformly increasing step function on the interval $[0, n]$.

Let us divide each element of the column-vector of ranks r_j by $n+1$. Then the range of ranks will be the unit interval $[0, 1]$ with step $1/(n+1)$, so ranks from 1 to n will be transformed to the relative ranks from $1/(n+1)$ to $n/(n+1)$. Nonexistent ranks 0 and $n+1$ will correspond to the boundary values 0 and 1 of the interval $[0, 1]$. A vector of ranks obtained in this way will be called a vector of relative ranks. Then every column-vector of relative ranks will be transformed by making use of the function inverse to the integral function of the standard normal distribution. As a result, we get the set of column-vectors $x_j = (x_{1j}, x_{2j}, \dots, x_{ij}, \dots, x_{nj})^T$, $j = 1, \dots, m$, that are correlated and have the standard normal distribution function. The column-vector

$x_i = (x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{im})$, $i = 1, \ldots, n$, will characterize the i-th text by the set of normally distributed new features that are correlated and have zero mean and unit variance. New features are normally distributed relative to the ranks of the relative frequencies of initial features.

As a result, we come to the nonlinear transformation of an initial feature space of non-normally distributed relative frequencies v in a new feature space of normally distributed relative ranks x. This allows one to use the parametric methods of discriminant analysis and classification (Klecka 1980, Kendall and Stuart 1979).

5 Mathematical tools of principal component analysis

Now we will find the n-column-vectors $y_l = (y_{l1}, y_{l2}, \ldots, y_{lm})^T$, $l = 1, \ldots, m$, of principal the components of the normalized data $\{x_j\}$ by the linear transformations $y_l = xU_l - y_{l0}$. Here $y_{l0} = x_{..}U_l$ are scalars (average values of principal components), $x_{..}$ – the m-row-vector of the average value n-column-vectors $\{x_j\}$, $x_{..} = \sum_{i=1}^{n} x_{ij}/n$, $j = 1, \ldots, m$; the m-column-vectors of coefficients $\{U_l\}$ are eigenvectors of the following symmetric positive definite empirical covariance $m \times m$-matrix K of vectors $\{x_j\}$. The coefficient vectors correspond to nonnegative eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_l, \ldots, \lambda_m$ that decrease monotonically with the growth of index l. These nonnegative eigenvalues define the variances of the principal components. Thus we have:

$$KU_l = \lambda_l U_l, K_{jj'} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - x_{.j})(x_{ij'} - x_{.j'}), \tag{1}$$

$$My_l = 0, Dy_l = \lambda_l,\ j, j', l = 1, \ldots, m, \tag{2}$$

where M is the mathematical expectation, D is the variance, eigenvalues $\{\lambda_l\}$ are the roots of characteristic equation $\det(K - \lambda I) = 0$. In this equation I is the identity diagonal matrix. We choose $k < m$ of the first principal components as new features from the principal component system. The new features correspond to the first eigenvalues, greater than unity. As a result, we get the nonlinear statistical reduction of the feature space of texts style. The space obtained has a smaller dimension than the initial one.

6 Mathematical tools of discriminant analysis

Now we will find the n-column-vectors of the discriminant functions (DF) $z_l = (z_{l1}, z_{l2}, \ldots, z_{ln})$, $l = 1, \ldots, q$, $q = \min(m, g-1)$, of the normalized data $\{x_j\}$ by applying the linear transformations $z_l = xV_l - z_{l0}$, where $z_{l0} = x_{..}V_l$

are scalars, $x_{..}$ is the m-row-vector of average values of n-column-vectors $\{x_j\}$; $x_{..}j = \sum_{i=1}^{n} x_{ij}/n,\ j = 1,\dots,m$; the m-row-vectors $\{V_l\}$ are eigenvectors that correspond to the matrices B and W and obey the equation $BV_l = \lambda_l WV_l$, $l = 1,\dots,q$. The set $\{\lambda_l > 0\}$ is composed of their first q eigenvalues, that satisfy the equation $\det(B-\lambda W) = 0$ (Klecka 1980). Here $B = T - W$, where $T/(n-1)$ is the total covariance $m \times m$-matrix of vectors $\{x_j\}$:

$$T_{jj'} = \sum_{k=1}^{g} \sum_{i_k=1}^{n_k} (x_{i_k j} - x_{..j})(x_{i_k j'} - x_{..j'}),\ j, j' = 1,\dots,m\ . \tag{3}$$

Inner summation is taken by the indices (rows) that correspond to the k-th class, $i_k = 1,\dots,n_k$, where n_k is the number of the elements (rows) of the k-th class; $\sum_{k=1}^{g} n_k = n$; $W/(n-g)$ is the within-group covariance $m \times m$-matrix of vectors $\{x_j\}$:

$$W_{jj'} = \sum_{k=1}^{g} \sum_{i_k=1}^{n_k} (x_{i_k j} - x_{.kj})(x_{i_k j'} - x_{.kj'}),\ j, j' = 1,\dots,m\ . \tag{4}$$

Here $x_{.kj} = \sum_{i_k=1}^{n_k} x_{i_k j}/n_k$, $k = 1,\dots,g$, $j = 1,\dots,m$ are elements of the $g \times m$-matrix of average values of vectors $\{x_j\}$ in the group (class). When average values of vectors $\{x_j\}$ for different classes (centers of classes) are equal ($x_{.kj} = x_{..}j$, $k = 1,\dots,g$), then matrices T and W coincide, and all elements of the matrix B are zero. But if the averages for different classes differ from each other, then the values of elements of matrix B specify the discrepancy measure between the groups (classes). The maximization of expression $\lambda_l = (V_l^T BV_l)/(V_l^T WV_l)$, $l = 1,\dots,q$, with respect to the weight vectors V_l provides the maximum discrimination ability of DF and leads to equation $BV_l = \lambda_l WV_l$, $l = 1,\dots,q$, that defines the eigenvectors of the matrix $W^{-1}B$. Variables $\{\lambda_l\}$ are eigenvalues of this matrix. They give the discrepancy measure between the classes for each DF, in the order of decreasing eigenvalues.

The utility of each l-th DF (for every new feature that is obtained in this way from the initial features) can be evaluated by means of the canonical correlation coefficient (Klecka 1980) $R_l = \sqrt{\lambda_l/(1+\lambda_l)}$, $0 \le R_l < 1$, $l = 1,\dots,q$. This coefficient expresses the level of statistical relationship of the l-th DF with its classes. The nearer the coefficient of canonical correlation is to 1, the higher is its relationship with its classes, and the greater and more secure is its discrimination of the class centers. This allows one to answer the question how many discriminant functions from the maximum number $q = \min(m,g-1)$ ensure the statistically significant discrimination of the class centers.

Let $j < q$ be the number of the first calculated DF. In discriminant analysis, Wilks' Lambda statistic Λ is used to estimate the total discriminative power of the remaining DF (“remainder discrimination”; cf. Klecka 1980):

$$\Lambda_j = \prod_{i=j+1}^{q} \frac{1}{1+\lambda_i},\ j = 0,\ldots,q-1. \tag{5}$$

If $j = 0$, one has the highest remainder discrimination because all $\{\lambda_l\}$ are nonnegative. The remainder discrimination is the lowest when $j = q-1$. So, Wilks' Λ-statistic is the "inverse" measure of class discrimination. A value of Λ close to zero indicates high discrimination of classes (it means that the centers of classes are well divided and differ greatly from each other with respect to the value of point scattering within the classes). As Λ increases to its maximum value (one) there is a gradual deterioration of class differentiation (the centers of classes fail to be significantly different with respect to the point scattering within classes).

For an estimation of the statistical significance of the discriminative power of the first j discriminant functions, Pearson's chi square test is used. It is based on the statistic

$$\chi_j^2 = -(n - \frac{m+g}{2}) \ln \Lambda_j,\ j = 0,\ldots,q-1. \tag{6}$$

This statistic has the probability density function χ^2 with $\nu_j = (m-j)(g-j-1)$ degrees of freedom under the condition that hypothesis H_0 is true (Klecka 1980). That means the remaining $q-j$ DF don't improve the discrimination ability of the first j DF (they don't increase the distance between the centers of classes). It allows one to calculate the significance level P (p-level) of the chi square test that has been reached (the actual probability of an error of the first kind to reject by mistake the null hypothesis when it is true): $P_j = 1 - F(\chi_j^2|\nu_j)$, where $F(\chi^2|\nu)$ is the integral function of the chi square distribution with ν degrees of freedom.

The interpretation of the discriminant functions as hidden parameters that determine the differences of classes can be achieved by correlation coefficient analysis (factor loadings analysis) of the column-vector z_l of the discriminant functions with column-vectors x_j of the normalized relative ranks:

$$\rho_{ij} = \frac{1}{n-1} \sum_{i=1}^{n} z_{il}(x_{ij} - x_{..j}) / \sqrt{Dz_l Dx_j},\ l = 1,\ldots,q,\ j = 1,\ldots,m\,. \tag{7}$$

It is well known (Sachs 1972) that statistic $t = \rho\sqrt{(n-2)/(1-\rho^2)}$ has Student's t-distribution with $\nu = n-2$ degrees of freedom, provided that z_l and x_j have normal distribution and the null hypothesis (the correlation coefficient $\rho = 0$) is true. This enables one to find the critical value of Student's statistics as quantile $t_{crit} = t_{n-2,1-P_{crit}/2}$ of level $1 - P_{crit}/2$ of this distribution. The critical significance level of Student's t-test should be fixed, e.g., $P_{crit} = 0.05$. From here one easily gets the critical value $\rho_{crit} = t_{crit}/\sqrt{(n-2)+t_{crit}^2}$ of the correlation coefficient which specifies $(1 - P_{crit}) \cdot 100\%$-th interval $[-\rho_{crit}, \rho_{crit}]$ of the statistical insignificance of the correlation coefficient. The values of the correlation coefficient outside this interval are statistically significant on $P \leq P_{crit}$ level of significance.

7 Example of feature space reduction on the basis of methods of principal components and discriminant analysis

The above described procedures of constructing m-row-vectors r_i, m-row-vectors x_i, m-row-vectors y_i, and q-row-vectors z_i ($i = 1, 2, \ldots, n$) from the original m-row-vectors p_i of relative frequencies of text style features for each textual work are implemented in the *StyleAnalyzer* software (Shevelyov and Poddubnyj 2010) which is intended for the complex statistical analysis of textual work styles of different authors, genres, etc. Figures 1–3 give examples of using the described approach to ranking, normalization and reduction of a feature space on the basis of the methods of principal components and discriminant analysis.

Textual material is represented by 80 large works of fiction by 11 Russian authors of the 19th century (11 works by N.V. Gogol’, 3 by I.A. Goncharov, 18 by F.M. Dostoevskij, 2 by I.A. Kuprin, 3 by M.Ju. Lermontov, 7 by N.S. Leskov, 9 by A.S. Pushkin, 2 by M.E. Saltykov-Shchedrin, 8 by L.N. Tolstoj, 13 by I.S. Turgenev, 4 by A.P. Chekhov).

We used 55 syntactic words as text style attributes.[1] Absolute frequencies of their occurrence in the text are the text style features. These frequencies are being presented in *StyleAnalyzer* in the form of a spreadsheet with indication of authors and texts in rows and that of style attributes in columns. Figure 1 shows the connection between the original attributes – the relative frequencies of one in 55 features (namely, the forth one) in 80 texts (*init-data*), the ranks of relative frequencies (*rank-data*) and the relative ranks (normally distributed after the non-linear transformations) of relative frequencies (*gauss-data*).

Eigenvalues of the covariance matrix K are the variances of the principal components. The calculation of them for *init-data* and *gauss-data* variables shows that several first principal components are responsible for the majority of text variability. For example, the first six principal components (10.1% of its total number) explain 51.4% of the feature variability for *gauss-data* and 49.6% for *init-data*.

Eigenvalues of matrix $W^{-1}B$ for *init-data* and *gauss-data* variables are the variances of the discriminant functions of these variables. One can see that only $q = \min(m, g-1) = 10$ of them are other then zero; here $m = 55$ is the number

[1] These syntactic words are: в, на, с, за, к, по, из, у, от, для, во, без, до, о, через, со, при, про, об, ко, над, из-за, из-под, под, и, что, но, а, да, хотя, когда, чтобы, если, тоже, или, то есть, зато, будто, не, как, же, даже, бы, ли, только, вот, то, ни, лишь, ведь, вон, то-есть, нибудь, уже, либо.

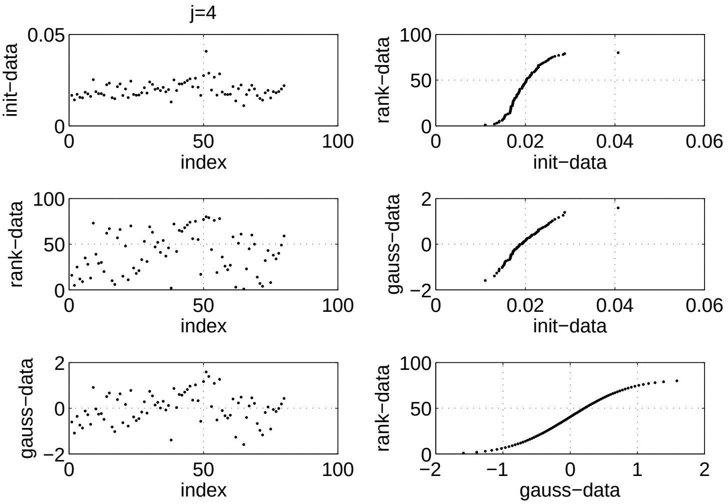


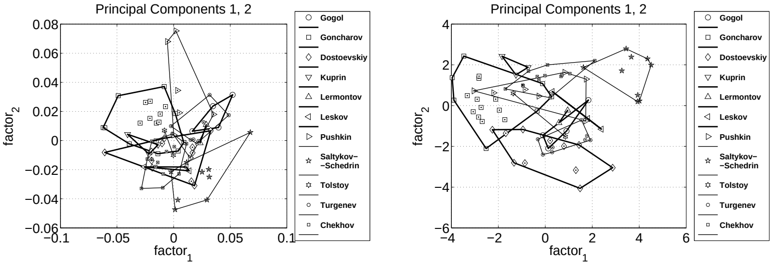
Figure 1: Nonlinear transformation of the data for the fourth feature

of original features (syntactic words), $g = 11$ is the number of classes (writers, authors of works). The calculation of the significance levels (p-levels) of the discriminant functions shows that almost all DF are statistically significant ($P < 0.05$ for the first nine DF).

The points with the markers of different types in Figure 2 represent 80 fiction works of 11 writers of the 19th century in the coordinates of the first two principal components (factors 1 and 2) for *init-data* (see Figure 2a) and *gauss-data* (see Figure 2b) variables. Convex hulls of sets of work-points for each author are shown by the closed broken lines. One can see that the normalized relative ranks of relative frequencies (*gauss-data*) distinguish between writers better than the relative frequencies.

The points with the markers of different types in Figure 3 refer to the same 80 fiction works of 11 writers of the 19th century in the coordinates of first two discriminant functions (factors 1 and 2) for *init-data* (see Figure 3a) and *gauss-data* (Figure 3a) variables. Convex hulls of sets of the work-points for each author are shown by the closed broken lines.

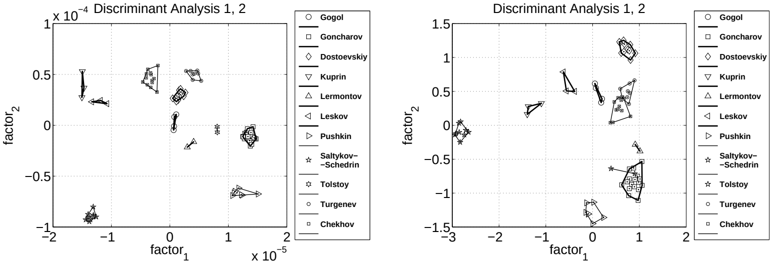
If one compares Figures 2 and 3, one sees that discriminant analysis provides full discrimination of classes by relative frequencies (*init-data*) and almost full discrimination by their normalized relative ranks (*gauss-data*), whereas the author classes overlap significantly in the course of principal component analysis.



(a) init-data

(b) gauss-data

Figure 2: Text representation in the coordinates of the first two principal components (features are 55 syntactic words)



(a) init-data

(b) gauss-data

Figure 3: Text representation in the coordinates of the first two discriminant functions (features are 55 syntactic words)

8 Conclusion

Thus, discriminant analysis ensures a considerably better discrimination of authors in terms of 55 syntactic words as compared with the analysis of principal components, though both methods provide graphical representation of the whole Russian fiction literature of the 19th century by sets of dots (representing texts) in the plane. This is to be expected since the discriminant analysis provides a transformation of the original attribute space of text styles that maximally increases the mean-square distance between the class centers fixing the distance variance between the elements (dots-texts) inside the classes on a constant level.

In other words, discriminant analysis makes author classes equally compact and maximally discriminated from each other. Residual overlapping of classes indicates the proximity of text styles of different authors that appears in the overlapping classes in the corresponding feature space.

In conclusion, it will be noted that close results could be obtained when the method of principal components and discriminant analysis is applied directly to ranks of frequencies rather than the normalized relative ranks of relative frequencies of attributes. This is due to the fact that gaussianity of data is no longer significant when the indicated methods are used for the multidimensional analysis of texts, though the calculations of statistical significance of the results may turn out to be incorrect.

References

Afifi, A.A.; Azen, S.P.
1979 *Statistical Analysis: A Computer Oriented Approach.* New York etc.: Academic Press.

Hollander, M.; Wolfe, D.A.
1999 *Nonparametric Statistical Methods.* New York etc.: Wiley.

Kendall, M.G.; Stuart, A.
1979 *The Advanced Theory of Statistics. Vol. 2: Inference and Relationship.* New York: Oxford University Press.

Klecka, W.R.
1980 *Discriminant analysis, Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-019.* Beverly Hills–London: Sage Publications.

Lawrence, N.
2005 "Probabilistic Non-Linear Principal Component Analysis with Gaussian Process Latent Variable Models", in: *Journal of Machine Learning Research,* 6; 1783–1816.

Sachs, L.
1972 *Statistische Auswertungsmethoden.* Berlin etc.: Springer.

Shevelyov, O.G.; Poddubny, V.V.
2010 "Complex investigation of texts with the system «StyleAnalyzer»". In: This volume, pp. 207–212.

Contents