Peter Zörnig, Emmerich Kelih* and Ladislav Fuks

Classification of Serbian texts based on lexical characteristics and multivariate statistical analysis

DOI 10.1515/glot-2016-0004

Abstract: We study lexical properties of different Serbian text types (poems, rock songs, different kinds of spoken language, prose, scientific and journalistic texts). We investigate characteristics and text parameters based on the frequency of word forms (relative frequencies, repeat rate, *h-point* and related indices). Relations between parameters are studied in order to identify the principal characteristics. We also apply techniques of multivariate analysis (cluster analysis, MDS) to classify the text types adequately. One additional objective of the present paper is to illustrate the functioning of these techniques in detail by explicit calculations of all programming steps.

Keywords: Text classification, multivariate statistical analysis, multidimensional scaling, Serbian, oral and written language, redundancy, h-point, lexical richness

1 Introduction

The purpose of the present article is a profound study of quantitative lexical characteristics of Serbian text types. We pay special attention to variations of written and oral varieties. Some of the analysed texts represent forms of spoken language, like rock songs, authentic examples of the spoken everyday language, whereas the other analysed text types are based on written varieties (like poems, prose, journalistic and scientific texts). In the first section we introduce relevant quantitative text characteristics, which are based on the rank frequency distribution of word forms. In particular we use the repeat rate of word forms, the lexical richness of texts (captured by indicator *a)*, the so-called *h-point*, and an indicator of word frequency distributions, which has recently been discussed at

***Corresponding author: Emmerich Kelih,** Institut für Slawistik, Universität Wien, Spitalgasse 2, Hof 3, A-1090 Wien, E-mail: emmerich.kelih@univie.ac.at
Peter Zörnig, Department of Statistics, Institute for Exact Sciences, University of Brasília, Asa Norte, 70910-900 Brasília, E-mail: peter@unb.br
Ladislav Fuks, E-mail: Ladislav.fuks@uni-graz.at

length in quantitative lexicology. Furthermore the frequency of the most common word forms in the analysed texts is taken into consideration. In the first section we examine the correlation between some pairs of these quantitative text properties (e. g. the correlation between the repeat rate and the most frequent word forms), since for the multivariate statistical analysis in the second part of the paper only non-correlated characteristics are used. In Section two we use a holistic approach to capture the relations between text types and all characteristics simultaneously by means of multivariate analysis techniques such as clustering and multidimensional scaling (MDS). Using these methods it will be shown in which way the similarity of texts on the lexical level, based on quantitatively determined characteristics, can be captured.

2 Pilot study: Repeat rate in language varieties

A basic quantitative text characteristic is the frequency of word forms and lexemes. Based on this quantitative property the so-called repeat rate can be calculated. This characteristic measures the degree of redundancy in the data under study (see Section 2.1 for a detailed description of the used corpus). Redundancy is a basic and constitutive property of language and text systems, and ensures the successful transmission of the informational context of a message. One possibility to study the redundancy on the text level is the analysis of word forms which occur repeatedly in texts. From information theory it is known that the more often a language sign (or entity or property) occurs, the lower is the amount of information provided. The more predictable a linguistic unit is, the more redundancy it contains (cf. Weber 2005: 217; Altmann 1973, 1972). According to this approach, redundancy can be operationalised and captured empirically, since the frequency of word forms in texts seems to be an adequate characteristic. An appropriate measure of the lexical structure is the repeat rate, which is based on the frequency of word forms in a text. It is defined by

$$R = \sum_{i=1}^{V} pr_i^2 \qquad [1]$$

where V is the number of word form types and pr_i the probability that type i occurs in the text (Zörnig and Altmann (1983), Popescu et al. (2009: 166)). Therefore, we can substitute pr_i in eq. [1] for f_i/N and obtain

$$R = \frac{1}{N^2} \sum_{i=1}^{V} f_i^2 \qquad [2]$$

where N is the sample size (text length in the number of word form tokens) and f_i the frequency of the respective word form in the text. It can be shown that the range of R is given by

$$1/V \leq R \leq 1 \quad [3]$$

where the lower limit $R = 1/V$ is assumed in the case when all probabilities are equal, i. e. $p_1 = \ldots = p_V = 1/V$. In this case all word forms occur with the same frequency, i. e. the repeat rate is minimal. The upper limit $R = 1$ is assumed when one of the probabilities p_i equals one and the others are zero. In this case all word forms are of the same type, i. e. the repeat rate is maximal. Obviously, the relation eq. [3] is equivalent to

$$0 \leq 1 - R \leq 1 - 1/V. \quad [4]$$

Thus the repeat rate can be normalised by defining the relative repeat rate

$$R_{rel} = \frac{1 - R}{1 - 1/V} \quad [5]$$

which has values between 0 and 1. The values of R_{rel} are close to 0 when R has a value close to 1, i. e. when the repeat rate is high; the values of R_{rel} are close to 1 when R has a value close to 1/V, i. e. when the repeat rate is low. Due to the normalising in eq. [5] it is possible to compare texts of different lengths, which is of great importance from a methodological and theoretical point of view.

2.1 Relative repeat rate in different Serbian varieties and text types

The corpus on which our analysis of word form frequencies is based consists of different varieties of spoken and written language: a) a sub-corpus of 194 rock songs (lyrics, released between 1979 and 2009) of the famous and popular Serbian (better ex-Yugoslavian) band Riblja Čorba, b) two sub-corpora of poems, written by the frontman of Riblja Čorba, Bora Đorđević (39 poems from the volume *Ravnodušan prema plaču* (Đorđević 1989) and 110 poems *Hej Sloveni* (Đorđević 1987), c) a comprehensive (over 30,000 tokens) corpus of authentic colloquial speech, taken from Savić and Polovina (1989), a "classical" Serbian prose text (Dnevnik o Čarnojeviću), written by Miloš Crnjanski (1893–1977), e) a selection of 90 poems written by M. Crnjanski, M. Dedinac and D. Maksimović, f) patient–doctor dialogues (3,999 tokens) and informal everyday communication (dialogues of families watching TV with 3116 tokens); both corpora taken from

Savić and Polovina (1989), g) three (longer) scientific articles from historical sciences and h) 30 journalistic texts from various authors.

This particular selection provides a possibility to gain inductively an impression about the amount of redundancy in different varieties, mainly focussing on speech and writing. For the quantitative analysis of the analysed text the number of word form types (V), the number of word form tokens (N), the relative repeat rate R_{rel}, the most frequent word (p_1) and the cumulated rang frequency $p_{cum10} = p_1 + ... + p_{10}$ of word forms with rank from 1 to 10 have been calculated. In Table 1 an illustration for the calculation of p_{cum10} is given, where the limit 10 has been chosen arbitrarily. This parameter illustrates the frequency behaviour in the initial part of the rank frequency of word forms, where preferentially only synsemantics, like particles, interjections, conjunctions etc. occur. In the analysed corpus of rock songs (cf. Table 2) the most frequent items is a conjunction (*i*), the second most frequent is *da* (either used as particle or as conjunction), followed by *je* (an auxiliary verb) etc. The ten most frequent items already cover 18.26 % of the whole corpus. As it can be seen from Table 2, the indicator p_{cum10} shows a quite broad range in the analysed varieties. It can be assumed that this particular characteristic seems to be a good "discriminator".

Table 1: Absolute and cumulated rank frequencies (rock songs).

Token	%	p_{cum10}
i	$p_1 = 3.02$	3.02
da	$p_2 = 2.94$	5.96
je	$p_3 = 2.60$	8.56
se	$p_4 = 2.04$	10.60
u	$p_5 = 1.73$	12.33
sam	$p_6 = 1.51$	13.85
ne	$p_7 = 1.24$	15.09
ja	$p_8 = 1.13$	16.21
na	$p_9 = 1.08$	17.29
ti	$p_{10} = 0.96$	18.26

The choice of written and oral varieties enables us to determine inductively the degree of redundancy. In particular one might expect all sub-corpora representing oral speech in a quite variant way to be the most redundant, e. g. to have the lowest relative repeat rate. A low repeat rate generally indicates a rather high usage of particular word forms, which in fact causes the high redundancy rate. Since the relative repeat rate is a rather abstract indicator one can also

Table 2: Used corpora and quantitative properties.

	Text type	Author/Source	N	V	R_{rel}	p_1	p_{cum10}
1	Modern poems I	Bora Đorđević	3,842	1,969	0.9963	2.9	16.2
2	Colloquial speech	Savić and Polovina (1989)	31,575	6,052	0.9936	3.9	21.3
3	Songs (rock)	Riblja Čorba	17,238	5,293	0.9951	3.0	18.3
4	("Classical") poems	M. Crnjanski, M. Dedinac, D. Maksimović	8,429	3,276	0.9944	5.1	18.1
5	Prose	Miloš Crnjanski	18,821	5,218	0.9923	5.5	22.5
6	Modern poems II	Bora Đorđević	3,428	1,768	0.9959	2.9	17.4
7	Science	Dragoslav Srejović	16,123	4,950	0.9933	5.4	20.49
8	Journalistic texts	various	16,286	5,437	0.9937	3.9	20.7
9	Patient–doctor dialogues	Savić and Polovina (1989)	3,999	1,034	0.9911	6.3	23.5
10	Everyday communication	Savić and Polovina (1989)	3,116	1,106	0.9935	4.01	22.75

consider p_1, which is relatively easy to interpret: the lower the repeat rate, the higher p_1 (cf. for details Section 2.2).

As already pointed out one would expect a rather low repeat rate for all oral varieties of the analysed corpus, namely the doctor–patient dialogues, everyday informal speech and generally the corpus of spoken Serbian language. This is explainable due to the high frequency of discourse markers (particles), interjections, conjunctions etc. in oral speech. This high frequency is thus representing an overall tendency to repeat particular word forms and accordingly particular parts of speech. However, a brief look at Figure 1, containing the analysed

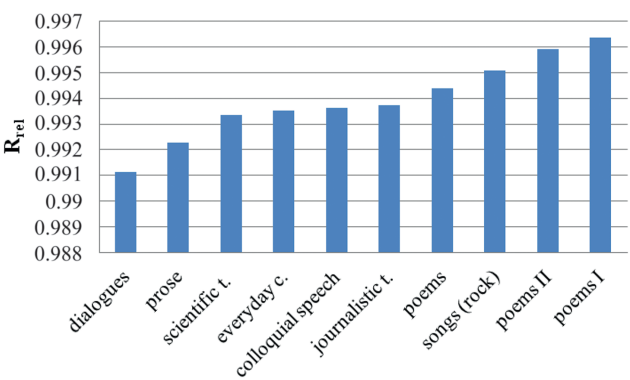

Figure 1: Relative repeat rate in various varieties of Serbian.

varieties and sorted by the increasing relative rate,[1] shows that not only oral speech is distinguished by a low repeat rate, but prose, and scientific and journalistic texts as well. Obviously these varieties have the same amount of redundancy as the oral speech (doctor–patient dialogues, everyday communication, and colloquial speech). The redundancy cannot be lower than a required comprehensibility/readability for a successful decoding process. Linguistically a high redundancy rate can be partly achieved by a high frequency of particular word forms, in particular of synsemantics (cf. Section 2.2 for the specific relation between the repeat rate and lexical structure of texts).

Based on Figure 1 the following results are obtained:

1. All analysed poems and rock songs are distinguished by a relatively high repeat rate. In this respect these varieties are characterised by a balanced lexical structure without any overexploitation of particular word forms. The latter would have a direct effect on the amount of redundancy and repeat rate. However, the results concerning the rock songs, which (in relation to the repeat rate) have to be treated as ordinary poems, clearly contradict the analysis by Fuks and Kelih (2012), where it has been shown that from the phonological, morphological and lexical point of view the rock songs are a mixture of Serbian slang, colloquial language and poetic language, with a predominant orientation towards oral speech and colloquial language. Now it clearly appears that according to the word frequency these songs clearly have to be classified as a poetic language. Although the analysed rock songs contain many lexical items from colloquial Serbian, there structural organisation is mostly orientated towards lyrics and poetics.
2. The repeat rate of journalistic and scientific texts and Serbian colloquial languages is approximately on the same level. The same holds true for prose texts, which can be understood as a heterogeneous mixture of colloquial speech, narrative sequences and descriptive style. Thus in this respect no clear-cut border of written and spoken language can be drawn based on the repeat rate, since dialogues, prose texts and scientific texts have more or less the same relative repeat rate.

The impression of the different levels of the relative repeat rate and a specific behaviour in some subgroups (on the one hand some kind of “poetic” language, including poems and rock songs, and on the other hand written and oral language) has to be tested by means of statistical methods. An appropriate test for obtaining significant differences between the relative repeat rate and

1 The calculated repeat rate gives only a preliminary idea. For an in-depth analysis of the repeat rate one should take the 95 % confidence interval and the variance into consideration.

Table 3: z-values for selected varieties.

Pairs of comparison		z-score
Dialogues	Poems I	9.17
	Rock songs	7.66
	Prose	2.06
Journalistic texts	Scientific texts	1.34

their variance is an asymptotic test, introduced and described in detail by Popescu et al. (2009: 165–169).

Using this method one obtains the z-scores, which are presented in Table 3. These scores have standard normal distribution. For $z > 1.96$ significant differences are stated at a level of significance of $\alpha = 5\,\%$.

There is no need for the presentation of all performed comparisons between pairs of varieties and text types, since in a first step the comparison of the varieties with the highest and lowest repeat rate is sufficient. The statistical difference between dialogues and poems is significant (at a level of $\alpha = 5\,\%$); statistically significant differences can be obtained between dialogues and the prose and between dialogues and rock songs (in all cases $z > 1.96$) too. However, no significant differences can be obtained between journalistic and scientific texts ($z = 1.3363$) and thus the selected varieties indeed represent a broad spectrum of the level of the repeat rate in Serbian written and oral texts. The further relevance of the relative repeat rate as a quantitative indicator of the lexical structure of texts will be examined in the next section in more detail, particularly regarding statistical interrelations with other quantitative characteristics R_{rel}, p_1, $p_{\mathrm{cum}10}$.

2.2 Interrelations: Relative repeat rate and lexical structure

In addition to its main function of being a descriptive indicator of the quantitative lexical structure, the relative repeat rate also appears to be a parameter which is directly interrelated with other quantitative text properties. A low relative repeat rate generally indicates that the front part of rank word frequencies is over-exploited. Empirically this can be demonstrated by a law-like interrelation between the relative repeat rate and p_1, e. g. the most frequent word form within one corpus or variety: the higher the relative repeat rate, the lower p_1, and the lower the repeat rate, the higher p_1. The stated interrelation between R_{rel} and p_1 can be captured by a simple power law $R_{rel} = 1.57^{\star}\ p_1^{-158.5}$. The value $R^2 = 0.77$ indicates at least the stated tendency (cf. Figure 2) regarding the interrelated behaviour of R_{rel} with respect to p_1. Since p_1 is interrelated with R_{rel}, a closer look at it seems to be appropriate.

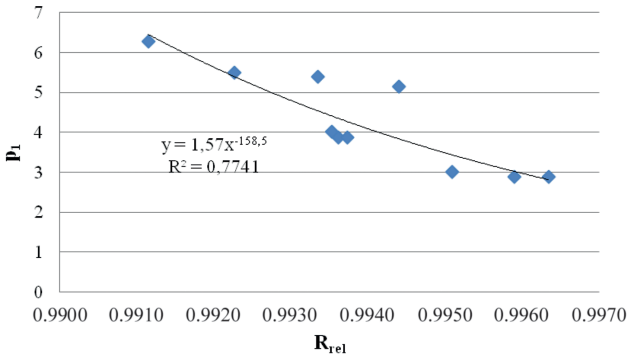

Figure 2: Interrelation between R_{rel} and p_1.

In our database a rather high p_1 (>5 %) can be obtained for the doctor–patient dialogues, for the scientific texts, the prose corpus and for the classical poems. In contrast, all other poems (written by Bora Đorđević) and especially rock songs have a slightly lower p_1. This rather simple text property indeed seems to be quite relevant and contains sufficient information about the redundancy amount of the analysed varieties. If one takes a closer look at the rank frequency distribution and the cumulated frequency of word forms in the rank from 1 to 10 (p_{cum10}), a quite different frequency coverage by the ten most frequent word forms can be observed: whereas only 16.5 % of the poems are covered by the ten most frequent word forms, in the dialogues already approximately a quarter of the whole corpus is covered by them. For details cf. Table 4 and 5 with the ten most frequent word forms of the analysed corpora, and p_{cum10}. The relevant

Table 4: Cumulated rank frequencies (p_{cum10}) and word form types.

Modern poems I		Colloquial speech		Songs (rock)		Classical poems		prose	
I	2.89	DA	3.88	I	3.02	I	5.14	I	5.49
DA	5.39	JE	7.04	DA	5.96	U	7.87	JE	9.29
SE	7.50	I	9.69	JE	8.56	DA	10.30	SE	11.97
JE	9.42	U	11.69	SE	10.60	SE	12.28	U	14.32
U	11.19	SE	13.67	U	12.33	JE	13.75	DA	16.14
MI	12.34	TO	15.38	SAM	13.85	ŠTO	14.79	SU	17.77
SAM	13.48	NE	16.91	NE	15.09	NE	15.76	SAM	19.06
NA	14.47	PA	18.43	JA	16.21	NA	16.70	NA	20.26
JA	15.33	A	19.85	NA	17.29	A	17.43	A	21.43
ZA	16.19	JA	21.26	TI	18.26	OD	18.12	ME	22.52

Table 5: Cumulated rank frequencies (p_{cum10}) and word form types.

Modern poems II		Scientific texts		Journalistic t.		Dialogues		Everyday comm.	
I	2.89	I	5.39	JE	3.86	DA	6.28	DA	4.01
DA	5.69	U	8.73	I	7.70	JE	9.53	JE	7.09
JE	8.11	SU	11.26	DA	11.01	TO	11.75	I	9.47
SE	10.36	SE	13.50	U	13.82	SE	13.93	U	11.68
U	12.16	NA	15.56	SE	15.87	I	16.05	TO	13.70
NA	13.62	JE	17.08	NA	17.40	NE	17.65	A	15.66
SAM	14.70	DA	18.23	SU	18.47	SAM	19.23	SE	17.55
SU	15.67	OD	19.23	ZA	19.25	PA	20.73	PA	19.35
KAD	16.54	A	19.86	OD	20.02	JA	22.18	NE	21.12
NE	17.39	ZA	20.49	NE	20.70	U	23.51	NA	22.75

word forms. It is obvious that p_{cum10} contains sufficient quantitative information about the initial part of the rank frequency distribution and the growth of synsemantics. Modelling the interrelation between R_{rel} and p_{cum10} leads as a matter of fact to quite satisfying results. Using a simple power model $R_{rel} = 12.70^* p_{cum10}^{-73.95}$ one obtains $R^2 = 0.87$, which indeed confirms the law-like interrelation of these two quantitative parameters of the lexical structure of texts (cf. Figure 3).

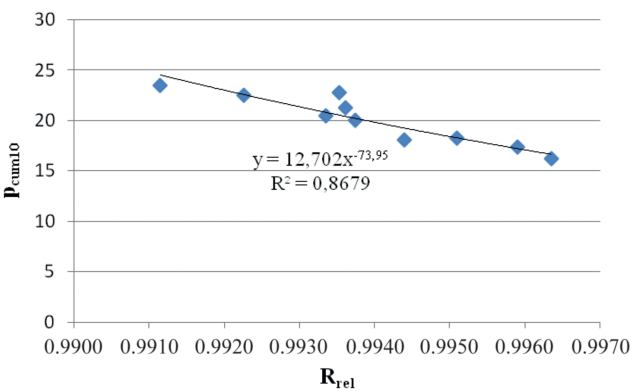

Figure 3: Interrelation of R_{rel} and p_{cum10}.

Finally the different growth of word forms in the initial part of the rank frequency distribution can be analysed.

Regarding the obtained word forms and parts of speech one has to state a quite specific mixture (cf. Tables 4 and 5):

1. The most frequent word form in all corpora (except journalism) is the conjunction *i*. Since the texts are not lemmatised it can't be excluded that *i* also appears as a particle, but in the most frequent cases *i* is used as a conjunction; it gives information about the grammatical and syntactical organisation of texts.
2. A similar high frequency as the conjunction *i* is obtainable for the word form *da*, which is a highly polysemous lexeme in Serbian, since it can be used as a particle or a conjunction. Moreover, as a conjunction it has a rich inner differentiation and can express causal, temporal and modal relations, which presumably causes the high usage in texts.
3. The third most frequent word form is the auxiliary verb *je* (infinitive form is *biti*), which is mainly used for the expression of temporal relations. A disambiguation and lemmatisation would surely lead to a much higher frequency.

Summarising this short overview of the "qualitative" input of the ranked word frequency clearly shows the dominance of synsemantic word forms. In Figure 4 the specific and individual growth of frequencies within the particular varieties can be seen. It is quite obvious that the most frequent word form determines the growth[2] of cumulated frequencies and no overlap of the cumulated frequencies can be obtained.

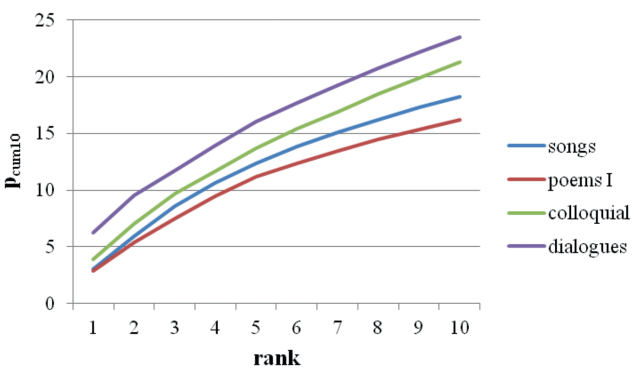

Figure 4: p_{cum10} for rank 1–10.

2 It is important to note, that there is no statistical correlation between text length (number of tokens and types) and p_{cum10}.

It is needless to say that in future this aspect of the frequency dynamics in texts must be explored much more systematically and in particular the statistical modelling of this growth has to be tackled. For the time being it is important to note that the most frequent word forms seem to determinate the further development of the rank frequencies.

The shown interrelations between the used parameters R_{rel}, p_1 and p_{cum10} are of great importance for the statistical classification of the analysed texts in the next section. Since the parameters are linked by law-like relations, there is no need to implement them all into multivariate analysis; one parameter seems to be sufficient. Since R_{rel} captures the whole word frequency distribution only this parameter will be used for the in-depth analysis of the similarity on the lexical level of the used Serbian texts. Furthermore some new parameters based on the rank word frequency will be introduced.

2.3 Parameter *a* and lexical richness: input from the *h-point*

The *h-point*, originally introduced into scientometrics and bibliometrics by Hirsch (2005), has been discussed intensively in quantitative linguistics and in word frequency studies (cf. Popescu and Altmann 2006, Mačutek, Popescu and Altmann 2008). The *h-point* is a fixed point on a rank-frequency distribution, where the rank r and the frequency $f(r)$ of a countable linguistic entity coincide. For special cases where one cannot obtain the point where $r = f_r$, one can determine the *h-point* by the point where the product of rank and frequency reaches its maximum (cf. Martináková et al. 2008: 93). In both above-mentioned cases the *h-point* can be obtained rather easily and mechanically. For the exact calculation of the *h-point* one can use the formula proposed by Popescu and Altmann (2008: 95):

$$h = \begin{cases} r & \text{if there is an } r = f_r \\ \dfrac{f_1 r_2 - f_2 r_1}{r_2 - r_1 + f_1 - f_2} & \text{if there is no } r = f_r \end{cases}$$

where f_r is the frequency of the element at rank r.

According to Popescu et al. (2009: 17–20), the *h-point* separates the vocabulary (V) of a text into two parts, namely into a class of magnitude h of frequent synsemantic auxiliaries (prepositions, conjunctions, pronouns, articles, particles, etc.) and a much greater class (the lexical items after the *h-point*) of autosemantics, which are not so frequent but form the lexicon and content of texts. Thus the *h-point* separates the "rapid" branch of synsemantics before the *h-point* from the "slow" branch of autosemantics after the *h-point*. Without a

doubt the separation of autosemantic and synsemantic word forms is not clear cut; sometimes there are autosemantics in the rapid branch and in other cases there are synsemantics in the slow branch. Thus the *h-point* is not an exact border between autosemantic and synsemantic word forms, but rather a fuzzy limit on a rank frequency distribution.

The main field of application of the *h-point* is quantitative text analysis and language typology. In cross-linguistic analysis and language typology the *h-point* can be interpreted as a sign of analytism, i. e. in analytic languages the number of word forms is smaller, and the synthetic elements are replaced by synsemantics. Furthermore, the *h-point* is considered to be a characteristic of individual texts within a given language and a sign of analytism/synthetism in cross-linguistic comparison. The *h-point* also can be used for the measurement of the lexical richness of texts. This seems to be valid only when one accepts the area below the *h-point* as being the relevant one for the lexical richness of a text. As shown above, this area is characterised by a large number of autosemantics and Popescu et al. (2009: 95 ff.) utilise this behaviour for their concept of *thematic concentration* of texts.

Despite the broad applicability of the *h-point* one has to take into account that the *h-point* systematically depends on the text lengths. Therefore, one has to analyse texts which are approximately of similar length or one uses indices which are based on the *h-point* but normalised by text length (cf. Popescu et al. 2009: 19).

Hirsch (2005) has argued that there is a relationship between the *h-point* and the text length N, represented by the total area below the rank-frequency curve, namely in the form of

$$N = ah^2.$$

However, as a dependence of a textological index on text length (sample size) is rather problematic, Popescu et al. (2009: 19) proposed the use of the parameter

$$a = \frac{N}{h^2}$$

as an adequate index, showing the partitioning of the texts into parts whose size is adapted to the text length. In this sense parameter a is a valuable parameter representing the area before the *h-point* in a proper way. As already pointed out, this area in rank frequency distributions is filled up with synsemantic auxiliaries, particles, interjections etc. (cf. Table 6 for further details). In this respect parameter a can be interpreted as an indicator of the synsemantic and pragmatic complexity of the texts analysed. Since the *h-point*

Table 6: Quantitative characteristics of used corpora.

No.	Variety	Author/source	N	V	*h-point*	*a*	*F(h)*	R_1
1	Modern poems I	Bora Đorđević	3,842	1,969	20	9.61	0.23	0.83
2	Colloquial speech	Savić and Polovina (1989)	31,575	6,052	38	6.91	0.45	0.63
3	Songs (rock)	Riblja Čorba	17,238	5,293	42,50	9.54	0.32	0.73
4	Classical poems	Crnjanski, Dedinac, Maksimović	8,429	3,276	30	9.37	0.28	0.77
5	Prose	Miloš Crnjanski	18,821	5,218	44	9.72	0.37	0.68
6	Modern poems II	Bora Đorđević	3,428	1,768	17.50	11.19	0.22	0.83
7	Science	Dragoslav Srejović	16,123	4,950	38.33	10.97	0.31	0.73
8	Journalistic texts	various	16,286	5,437	34	14.09	0.30	0.74
9	Doctor–patient dialogues	Savić and Polovina (1989)	3,999	1,034	24.5	6.66	0.35	0.72
10	Everyday communication	Savić and Polovina (1989)	3,116	1,106	20	7.79	0.31	0.76

depends directly on text length (this is a general and known drawback of the *h-point*) the parameter *a* is according to Popescu et al. (2009: 19) more adequate and more restrictive. Whereas the *h-point*, as already pointed out, directly depends on text length (using a simple linear model for the data yields $R^2 = 0.96$, which clearly indicates a strong interrelation), the parameter *a*, which is based on it, does not. Furthermore parameter *a* is much lower than the *h-point* and thus it represents the core of high-frequency words in the rank-frequency distribution.

Since in cross-linguistic studies the *h-point* and parameter *a* are interpreted as indicators of analytism (the number of word forms is smaller and synthetic elements are replaced by synsemantics), or synthetism, parameter *a* for text-based analyses within one language can only be understood as an indicator for the different degree of the grammatical and morphosyntactical organisation of texts.

2.4 Lexical richness – further possibilities

In addition to the grammatical organisation, the lexical richness of a text is considered to be a possibility for the determination of the similarity of texts. For the quantification of the lexical richness various indices have been discussed, but many of them are "suffering" from text length, e. g. the indices depend on the sample size (cf. Wimmer and Altmann 1996 for an comprehensive overview

on this problem and Hoover (2003) for general critics). Again, an idea by Popescu et al. (2009) deserves to be picked up. For defining an index of the lexical richness in a first step the cumulative relative frequency up to the *h-point* named $F(h)$ is taken into consideration. This area is covered mainly by auxiliaries, since the *h-point* is only a fuzzy border between auto- and synsemantic word forms. To account for this fuzziness, an empirical correction for the calculation of the indicator for lexical richness R_1 has been proposed by Popescu et al. (2009: 30) in the form of:

$$R_1 = 1 - \left(F(h) - \frac{h^2}{2N} \right).$$

In other words, the lexical richness is calculated based on the area with the main bulk of autosemantic word forms minus the pre-h-point area and a slight empirical correction. Indeed this seems to be a possibility of an operational recording of the lexical richness of texts. For the calculation of R_1 the *h-point* was rounded down and one gets the values presented in Table 6. First of all the ranked lexical richness based on R_1 leads to the following results. All analysed poems (regardless of the fact that they are written by a so-called "street poet" (Bora Đorđević) or by classical Serbian poets) have the highest lexical richness among the studied text corpora. Determined by its form (verse, rhythm, metre etc.) it is a rather specific kind of language, with its own poetic function. It is worth mentioning and partly astonishing that oral speech, which is analysed in different modulations (corpus of common colloquial speech, dialogues between doctor and patients, and informal everyday speech of people watching TV), displays a quite heterogeneous picture. Everyday speech of people ($R_1 = 0.77$) has almost the same lexical richness as classical poems ($R_1 = 0.77$) and doctor–patient dialogues have a slightly lower lexical richness ($R_1 = 0.72$), the corpus of common colloquial language has the lowest degree ($R_1 = 0.63$) of lexical richness. The obtained heterogeneity in oral speech can partly be explained by a text-internal specific regulation of phatic and expressive needs of oral language, combined with the omnipresent need of transmitting (partly new) information and content. In this respect it is quite interesting that journalistic and scientific texts and rocks songs seem to bear the same amount of lexical information. Regarding prose texts, they are generally closer to common oral speech ($R_1 = 0.68$), which is obviously determined by a high amount of mainly narrative sequences in oral speech. In the next section it will be shown how quantitative features of texts (lexical richness and grammatical structure) can be utilised for multidimensional scaling and cluster analysis.

3 Classification of text types via multivariate analysis

In the previous section we studied different characteristics of ten text varieties, i. e. R_{rel}, p_1, p_{cum10}, R_1, *h-point*, the parameter a etc. We also analysed some relations between pairs of these characteristics, in particular between R_{rel} and p_1, between R_{rel} and p_{cum10} and between the parameters a and R_1. It appears that

a.) the relative repeat rate R_{rel},
b.) the parameter a, and
c.) the lexical richness R_1

represent an appropriate set of different important text characteristics (generally the repetitions of words forms, grammatical/morphosyntactical structure and lexical richness) for the analysis of the similarities of the text types. The observed values of these three text characteristics are presented in Table 7.

Table 7: Observed parameter values.

No.	Author/source	R_{rel}	R_1	a
1	Poems I	0.9963	0.83	9.61
2	Colloquial speech	0.9936	0.63	6.91
3	Songs (rock)	0.9951	0.73	9.54
4	"Classical" poems	0.9944	0.77	9.37
5	Prose	0.9923	0.68	9.72
6	Poems II	0.9959	0.83	11.19
7	Science	0.9933	0.73	10.97
8	Journalistic texts	0.9937	0.74	14.09
9	Patient–doctor dialogues	0.9911	0.72	6.66
10	Everyday communication	0.9935	0.76	7.79

In the following we try to capture the relations between the above characteristics simultaneously. We make use of a clustering technique and a multidimensional scaling[3] (henceforth MDS) approach in which the text types are considered as

3 The application of multivariate statistical methods (clustering, multidimensional scaling, ANOVA etc.) in linguistics is becoming more and more popular. Cf. Janda (2013) about the application of these methods in cognitive linguistics, cf. Croft and Poole (2008) for application in language typology and universal research, Biber (1988) can be treated as pioneer of the application of multidimensional scaling in text classification and variational linguistics. Wheeler (2005) gives an overview of multidimensional scaling in linguistics and particularly emphasises dialectometry.

points in a geometric space. Instead of merely applying software (like, for example *R*) and presenting the calculated results we illustrate the functioning of the procedures step by step. The coordinates of the points correspond to the text characteristics (see Figure 6) and the distances between points represent the similarity/proximity between the corresponding text types. The similarities can be expressed by means of a general metric, e. g. the Minkowski distance, however we will confine ourselves to the specific case of the Euclidean distance. The mentioned techniques can be applied regardless of whether the individual characteristics are correlated or not.

Generally, in cluster analysis one searches for patterns in a data set by grouping the (multivariate) observations into clusters. The goal is to find a grouping such that the objects inside the clusters are similar (close) and objects in different clusters are dissimilar (distant).

It should be emphasised here that even with complex statistical methods one can usually not identify the "best" clustering. This is a question of definition, i. e. the goodness of a clustering depends on how the just mentioned objectives or other criteria are measured or weighted.

One hopes to detect "natural groupings" in the data which make sense for the respective application (see for example Rencher (2002: Chapter 14)). In our case we encounter groups of text types that can be well interpreted linguistically. For further applications of cluster analysis in quantitative linguistics see for example Jensen (2013).

MDS describes a family of techniques for the analysis, in particular visualisation of similarities on a set of objects (often called stimuli), to reveal a hidden structure underlying the data (see for example Steyvers (2006)). The main idea consists of embedding the given data points in a space of smaller dimension, preserving the distances between the points as well as possible. For example, it may turn out that the points lie on a circle, when they have been embedded into a plane. MDS techniques have already been applied in quantitative linguistics, e. g. in linguistic typology (studying structural similarities between languages), sociolinguistics (classifying varieties of English) and in semantic analysis (see Croft and Poole (2008) and Steyvers (2006)). In the present paper we use MDS to transform the three-dimensional data space of Table 7 into a two-dimensional one. Though the dimension reduction is not considerable, we use the example to illustrate how the statistical procedure works.

The objects of our analysis are the text types in Table 7. Since the ranges of the three observed characteristics differ considerably, it is reasonable to transform all ranges into the interval [0, 1], before illustrating the data geometrically (see for example De Souto et al. (2008: Section II)). This

Table 8: Normalised parameter values of the text types.

No.	Author/source	R_{rel}	R_1	*a*
1	Modern poems I	1.0000	1.0000	0.3970
2	Colloquial speech	0.4808	0.0000	0.0336
3	Songs (rock)	0.7692	0.5000	0.3876
4	"Classical" poems	0.6346	0.7000	0.3647
5	Prose	0.2378	0.2500	0.4118
6	Modern poems II	0.9231	1.0000	0.6097
7	Science	0.4231	0.5000	0.5801
8	Journalistic texts	0.5000	0.5500	1.0000
9	Patient–doctor dialogues	0.0000	0.4500	0.0000
10	Everyday communication	0.4615	0.6500	0.1521

rescaling guarantees that all text characteristics are equally reproduced in the geometric representation. For example, the first observation *r* of the parameter R_{rel} will be transformed linearly into $\bar{r} = \frac{r - r_{min}}{r_{max} - r_{min}}$, where r_{min} and r_{max} denote the minimal and the maximal observed value. In this way we obtain the data, presented in Table 8.

Each of these ten texts is now identified with the point given by its parameter values, e. g. text 5 (Prose) in Table 8 corresponds to the point $P_5 = (0.2378, 0.25, 0.4118)$ (see Figure 5). The Euclidean distance between the text types 5 and 6 for example is given by $\sqrt{(0.2308 - 0.9231)^2 + (0.25 - 1)^2 + (0.4118 - 0.6097)^2} \approx 1.0397$

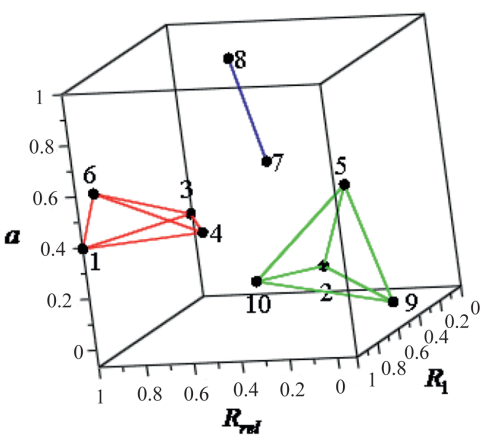

Figure 5: Three-dimensional representation of the text types in Table 8.

Figure 5 shows the localisation of the diverse text types based on the normalised parameters. One can visually recognise the three groups (clusters) which are indicated by connecting solid lines. One cluster consists of the poems (no. 1, 4, 6) and songs (no. 3), while another is formed by colloquial language in its various modalities (everyday language and diverse dialogues) and prose (no. 2, 5, 9, 10). Finally, a third group – though not as homogeneous as the others consist of journalistic and scientific texts (no. 7 and 8).

In the following section we show that this visual impression of clustering can be confirmed by the statistical method, presented in section 3.1.

3.1 The k-means clustering approach

We will divide the ten text types into clusters, by using one of the most common and simplest clustering techniques. This method is a heuristic, i. e. the result is not necessarily an optimal but usually a "good" solution (see Teknomo 2007 for simple examples).

The procedure consists of the following steps.

k-means clustering algorithm:

Initialisation: Specify the number k of clusters. Define k initial "centroids" $M_1,\ldots,M_k$.[4]
Iteration: Obtain k clusters by assigning each data point to the nearest centroid. Compute new centroids by the mean values of the current cluster elements.

Repeat the iteration until the clusters do not change any more.

The objective of this algorithm is to minimise the squares of distances inside of the k clusters, i. e. the objective function to be minimised is $z = \sum_{i=1}^{k} \sum_{P \in cluster\ i} d(P, M_i)^2$, where M_i is the centroid (mean value of the i-th cluster) and d(P,Q) denotes the distance between the points P and Q. The quantity z is a measure for the "quality" of the clustering.

We now apply the algorithm to the data in Table 9. Motivated by a look at Figure 5 we choose $k = 3$ and as initial centroids we use the texts/points P_1, P_9 and P_8, which are far away from each other. The distances between the ten data points are given by the following symmetric table.

4 In principle these points can be chosen arbitrarily, but with respect to the desired large distances between different clusters it is useful to choose k of the data points which are as far away as possible from each other.

Table 9: Distances between the investigated texts.

	$P_1 = M_1$	P_2	P_3	P_4	P_5	P_6	P_7	$P_8 = M_3$	$P_9 = M_2$	P_{10}
$P_1 = M_1$	0	1.18	0.551	0.474	1.07	0.226	0.785	0.903	1.21	0.687
P_2	1.18	0	0.677	0.789	0.518	1.24	0.743	1.11	0.659	0.661
P_3	0.551	0.677	0	0.242	0.594	0.568	0.396	0.671	0.863	0.416
P_4	0.474	0.789	0.242	0	0.606	0.483	0.362	0.666	0.773	0.279
P_5	1.07	0.518	0.594	0.606	0	1.04	0.357	0.713	0.513	0.53
P_6	0.226	1.24	0.568	0.483	1.04	0	0.708	0.731	1.24	0.738
P_7	0.785	0.743	0.396	0.362	0.357	0.708	0	0.43	0.72	0.455
$P_8 = M_3$	0.903	1.11	0.671	0.666	0.713	0.731	0.43	0	1.12	0.855
$P_9 = M_2$	1.21	0.659	0.863	0.773	0.513	1.24	0.72	1.12	0	0.525
P_{10}	0.687	0.661	0.416	0.279	0.53	0.738	0.455	0.855	0.525	0

A detailed study of Table 9 reveals that (1) the closest centroid for point P_1 is M_1, thus P_1 is assigned to the first cluster, (2) the closest centroid for point P_2 is M_2, thus P_2 is assigned to the second cluster etc.

Summarising the procedure, we obtain that the closest centroid for the point $P_1,\ldots,P_{10}$ is M_1, M_2, M_1, M_1, M_2, M_1, M_3, M_3, M_2, M_2, respectively. Thus the clusters of the first iteration are $C_1 = \{1,3,4,6\}$, $C_2 = \{2,5,9,10\}$ and $C_3 = \{7,8\}$. We now construct the *new centroids* by

$$M_1 = \frac{1}{4}(P_1 + P_3 + P_4 + P_6)$$
$$= \frac{1}{4}\left(\begin{pmatrix}1\\1\\0.397\end{pmatrix} + \begin{pmatrix}0.7692\\0.5\\0.3876\end{pmatrix} + \begin{pmatrix}0.6346\\0.7\\0.3647\end{pmatrix} + \begin{pmatrix}0.9231\\1\\0.6097\end{pmatrix}\right) = \begin{pmatrix}0.8317\\0.8\\0.4398\end{pmatrix}$$

and similarly $M_2 = \frac{1}{4}(P_2 + P_5 + P_9 + P_{10}) = \begin{pmatrix}0.2933\\0.3375\\0.1494\end{pmatrix}$ and $M_3 = \frac{1}{2}(P_7 + P_8) = \begin{pmatrix}0.4615\\0.525\\0.79\end{pmatrix}$.

The distances between the text types and the new centroids are as follows:

Table 10: Distances between texts and new centroids.

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
M_1	0.2648	0.9634	0.3108	0.2334	0.8151	0.2779	0.526	0.6974	1.004	0.4942
M_2	0.9999	0.4031	0.5565	0.5425	0.2836	1.0234	0.4783	0.9008	0.3478	0.3549
M_3	0.8185	0.9209	0.5072	0.4914	0.5215	0.6864	0.2149	0.2149	0.918	0.6501

As before, the closest centroid for the point $P_1,\ldots,P_{10}$ is M_1, M_2, M_1, M_1, M_2, M_1, M_3, M_3, M_2, M_2, respectively. Therefore the same clusters as in the previous iteration are determined. The algorithm ends, since it cannot improve the above clustering, which is identical to that found previously. By visual inspection of Figure 5 and comparing with Table 9 we verify that points inside of the same cluster are close (distance always ≤ 0.661) while the distance between points of different clusters is *usually* larger than that value. The clustering into the groups $C_1 = \{1,3,4,6\}$, $C_2 = \{2,5,9,10\}$ and $C_3 = \{7,8\}$ found by pure geometrical or numerical arguments can be interpreted linguistically quite reasonably.

Some numerical experiments have shown that the same clustering as above is found for diverse other selections of the initial centroids. The objective value of the given clustering is (see Table 10)

$$\begin{aligned} z &= d(P_1 - M_1)^2 + d(P_3 - M_1)^2 + d(P_4 - M_1)^2 + d(P_6 - M_1)^2 + d(P_2 - M_2)^2 + d(P_5 - M_2)^2 \\ &\quad + d(P_9 - M_2)^2 + d(P_{10} - M_2)^2 + d(P_7 - M_3)^2 + d(P_8 - M_3)^2 \\ &= 0.2648^2 + 0.3108^2 + 0.2334^2 + 0.2779^2 + 0.4031^2 + 0.2836^2 + 0.3478^2 + 0.3549^2 \\ &\quad + 0.2149^2 + 0.2149^2 = 0.8806. \end{aligned}$$

One could try to improve this clustering for example by removing point P_5 from cluster C_2 and assigning it instead to cluster C_3, since P_5 is also "close" to M_3. However the above objective value z would change to $z_{new} = z - d(P_5\text{-}M_2)^2 + d(P_5 - M_3)^2 = 0.8806 - 0.2836^2 + 0.5215^2 = 1.0721$, indicating a worse clustering.

3.2 A classical scaling approach

Usually, in MDS applications the objects under study have many characteristics, thus they correspond to points in a high-dimensional space. The main problem of MDS consists of mapping the given data points P_i into points Q_i in a space of "small" dimension, conserving the distances as far as possible. Through this reduction of dimensionality the complexity of the problem can be considerably simplified.

In this section we reduce the three-dimensional space used for the illustration of text types to a two-dimensional space and confirm the clustering found in Section 2.1. In the following we apply a classical scaling algorithm to reduce an n-dimensional space to a k-dimensional one ($k<n$). The procedure is called *double centring*, since the points Q_i are chosen such that their centre coincides with the origin of the coordinate system (see Borg and Groenen (1997)). The advantage of this method is that no iterative procedure is necessary. An analytical solution is provided, based on principles of linear algebra. The

algorithm which will be presented as follows can be realised for example by using the software MAPLE if the dimension of the original space is not too large.

Classical MDS algorithm:

1) Given a matrix $X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{pmatrix}$ whose lines correspond to characteristics of m objects, interpreted as points in the n-dimensional space.
2) Construct the matrix of squared distances $D^{(2)}$ between these points.
3) Construct the *double centre matrix* $B = -\frac{1}{2} J\, D^{(2)}\, J$, in which J is defined by

$$J = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix} - \frac{1}{m} \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \qquad [6]$$

(all matrices in eq. [6] are $m \times m$).

3) Determine the k largest positive eigenvalues $\lambda_1,\ldots,\lambda_k$ of B and corresponding orthonormal eigenvectors $e^1,\ldots,e^k$, i. e. these vectors have length 1 and e^i and e^j are orthogonal for $i \neq j$.
4) Calculate $Y = E \begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_k} \end{pmatrix}$ ($m \times k$ matrix)

where $E = (e^1|\ldots|e^k)$ is the matrix whose columns are the eigenvectors $e^1,\ldots,e^k$. The matrix Y has m lines, representing vectors in the k-dimensional space, the distances between which in general are similar to those between the lines of the matrix X.

We now apply the algorithm to the matrix in Table 8, so we have

$$X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,3} \\ \vdots & & \vdots \\ x_{10,1} & \cdots & x_{10,3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0.397 \\ \vdots & \vdots & \vdots \\ 0.4615 & 0.65 & 1.1521 \end{pmatrix} \qquad [7]$$

Furthermore we set $k = 2$, i. e. the three-dimensional space will be reduced to a two-dimensional one:

1) We calculate the symmetric 10×10 matrix

$$D^{(2)} = \begin{pmatrix} {d^{(2)}}_{1,1} & \cdots & {d^{(2)}}_{1,10} \\ \vdots & & \vdots \\ {d^{(2)}}_{10,1} & \cdots & {d^{(2)}}_{10,10} \end{pmatrix}.$$

For example, $d^{(2)}_{1,2} = 1.4017$ is the square of the Euclidean distance between lines 1 and 2 of the matrix eq. [7].

2) We determine the 10×10 matrix B.

3) The $k = 2$ largest eigenvalues of B are $\lambda_1 \approx 1.6015$ and $\lambda_2 \approx 0.5891$ with corresponding matrix of orthonormal eigenvectors

$$E \approx \begin{pmatrix} -0.4493 & 0.3678 \\ 0.44 & 0.1688 \\ -0.0801 & 0.0802 \\ -0.1059 & 0.1291 \\ 0.3038 & -0.2784 \\ -0.4866 & 0.0874 \\ 0.0231 & -0.292 \\ -0.1892 & -0.7278 \\ 0.4637 & 0.1822 \\ 0.0805 & 0.2826 \end{pmatrix},$$

i. e. the eigenvectors e^1 and e^2 are the first and second column of E, respectively. The eigenvectors have length 1 and are orthogonal.

4) We calculate the 10×2 matrix

$$Y \approx E \begin{pmatrix} \sqrt{1.6015} & 0 \\ 0 & \sqrt{0.5891} \end{pmatrix} \approx \begin{pmatrix} -0.5685 & 0.2823 \\ 0.5568 & 0.1296 \\ -0.1013 & 0.0615 \\ -0.134 & 0.0991 \\ 0.3845 & -0.2137 \\ -0.6158 & 0.0671 \\ 0.0292 & -0.2241 \\ -0.2394 & -0.5586 \\ 0.5868 & 0.1398 \\ 0.1018 & 0.2169 \end{pmatrix}$$

The lines of Y are the "new" points in the plane representing the ten text types in Table 8. They are illustrated in Figure 6. For example, the third text type in Table 8, "songs (rock)", is now represented by the point P_3 in Figure 6 with the coordinates −0.1013 and 0.0615.

The distances between any two points in Figure 6 correspond approximately to those in Figure 5. We obtain the same clusters as in Section 2.1, using the k-means clustering algorithm or simple visual inspection: $C_1 = \{1,3,4,6\}$: Poems and Songs, $C_2 = \{7,8\}$: Science and Journalism and $C_3 = \{2,5,9,10\}$: Colloquial speech, prose, patient–doctor dialogues, and everyday communication. Note that by means of the reduction of dimensionality a scaling algorithm also reduces the original dimension n to dimension k (which is usually much smaller than n). Hence, the original n characteristics are "compressed" into k main characteristics

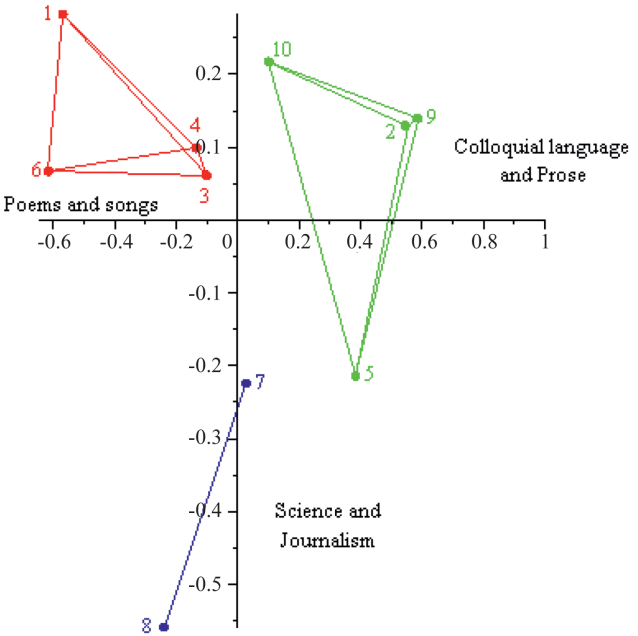

Figure 6: Two-dimensional representation of the ten text types/varieties.

and it is always most interesting for the application to interpret the latter. In our case $n = 3$ characteristics are compressed into $k = 2$, corresponding to the coordinate axes in Figure 6. We suggest the following interpretation of the axes. In the direction of the vertical axis (from bottom to top), the information content decreases along with the degree of goal-oriented and objective mediation of knowledge, news, facts and events. This is reflected in the gradual diversity of the information content of scientific, journalistic and prosaic texts. In the direction of the horizontal axis (from left to right) one can observe a gradual transition from written to oral language, since poems, songs and colloquial language use these different forms to transfer information, accompanied by a specific regulation of the redundancy. This is supported by the fact that these text types have different functions. In particular, the existence of the poetic group, which includes different poems and rock songs, shows indeed that on the level of the formal structure their classification is justified, even though rock songs are mostly designed for an oral, vocalised representation. Generally poems and rock songs are in this respect a specific form of a poetical language with explicit aesthetic and rhythmic qualities.

The clustering found above generally confirms the complex interrelation of redundancy and lexical richness in the analysed varieties and text types.

Bearing in mind the quantitative information with different rather simply obtainable quantitative facets of word frequencies (R_{rel}, R_1, and parameter a), the obtained results are satisfying from a linguistic point of view too. Without any doubt, the proposed method and obtained clustering of classification of text must also be corroborated and applied in many other languages before far-reaching generalisations can be made.

4 Conclusion

The main aim of the presented paper was to show the adequacy of the selected quantitative parameters, which are based on the frequency of lexical items (word-form tokens and types). In particular it appears that the relative repeat rate, parameter a and R_1 (as indicators based on the *h-point*, reflecting the vocabulary richness of texts) are appropriate measures for the classification and clustering of text types and varieties of Serbian spoken and written language. One main advantage of these measures is, that they are independent from the sample size and that they all can be normalised, i. e. transformed into the interval [0, 1]). The most appropriate and linguistically well interpretable classification (supported by means of techniques of multivariate analysis, namely cluster analysis and MDS) is based on the reduction of a three-dimensional to a two-dimensional one. Thereby we could identify two main text characteristics, namely the degree of goal- and fact-orientated mediation of information and the different degree of redundancy in varieties of spoken and written language. As a perspective a more in-depth analysis of the repeat rate on the lexical level and the "empirical" behaviour of the h-point and related indices is required.

Acknowledgment: The author are indebted to those who reviewed this article for many helpful comments and suggestions.

References

Altmann, Gabriel. 1973. Mathematische Linguistik [Mathematical linguistics]. In Koch, Walter A. (ed.), *Perspektiven der Linguistik*, 208–232. Stuttgart: Kröner.

Altmann, Gabriel. 1972. Status und Ziele der quantitativen Sprachwissenschaft [Status and goals of quantitative linguistics]. In Jäger Siegfried (ed.), *Linguistik und Statistik*, 1–9. Braunschweig: Vieweg.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge u.a.: Cambridge University Press.

Borg, I. and Groenen, P. 1997. *Modern multidimensional scaling: Theory and applications*. New York: Springer.

Croft, William and Poole, Keith. 2008. Multidimensional scaling and other techniques for uncovering universals. *Theoretical Linguistics* 34(1). 75–84.

De Souto Marcílio C.P. et al. 2008. *Comparative Study on Normalizing Procedures for Cluster Analysis of Gene Expression Datasets*, IJCNN, 2792–2798.

Đorđević, Bora. 1986. *Ravnodušan prema plaču* [Indifferent towards crying]. Beograd: Književne Novine.

Đorđević, Bora. 1987. *Hej Sloveni. Pesme* [Hey Slavs]. Beograd: Glas.

Fuks, Ladislav and Kelih, Emmerich. 2012. Lingvistička Analiza YU–Roka: Riblja čorba [Linguistic analysis of YU-rock: Riblja čorba]. In Ekaterina Kislova et al. (eds.), *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV 15)* (Die Welt der Slaven, Sammelbände/ Sborniki 46), 90–96. München: Sagner.

Hirsch, J.E. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America* 102(46). 16569–16572.

Hoover, David L. 2003. Another perspective on vocabulary richness. *Computers and the Humanities* 37(2). 151–178.

Janda, Laura A. 2013. Quantitative methods in cognitive linguistics: An introduction. In Laura A. Janda (ed.), *Cognitive linguistics. The quantitative turn. The essential reader*, 1–32. Berlin: De Gruyter Mouton.

Jensen, Kim Ebensgaard. 2013. *An R-based Cluster Analysis Program for Linguistic Analysis*. Aalborg: PhD thesis.

Mačutek, Ján, Popescu, Ioan-Iovitz and Altmann, Gabriel. 2007. Confidence intervals and tests for the h-point and related text characteristics. *Glottometrics* 15. 45–52.

Martináková, Zuzana et al. 2008. Some problems of musical texts. *Glottometrics* 16. 80–110.

Popescu, Ioan-Iovitz and Altmann, Gabriel. 2006. Some aspects of word frequencies. *Glottometrics*. 13. 23–46.

Popescu, Ioan-Iovitz and Altmann, Gabriel. 2008. On the regularity of diversification in language. *Glottometrics* 17. 94–108.

Popescu, Ioan-Iovitz, et al. 2009. *Word Frequency Studies*. (Quantitative Linguistics, 64). Berlin, New York: Mouton De Gruyter.

Rencher, Alvin C. 2002. *Methods of Multivariate Analysis*. New York: Wiley.

Savić, Svenka and Vesna Polovina. 1989. *Razgovorni srpskohrvatski jezik* [Colloquial Serbo-Croatian language]. Novi Sad: Institut za južnoslovenske jezike. Filozofski fakultet.

Steyvers, Mark. 2006. Multidimensional scaling. In Lynn Nadel (ed.), *Encyclopedia of Cognitive Science*. Chichester: Wiley & Sons. [=http://onlinelibrary.wiley.com/doi/10.1002/0470018860.s00585/abstract] (accessed February 23, 2016)

Teknomo, Kardi. 2007. K-means Clustering Tutorial, [=http:\\people.revoledu.com\kardi\tutorial\kMean] (accessed February 23, 2016)

Weber, Sabine. 2005. Zusammenhänge [Interrelations]. In Reinhard Köhler, Gabriel Altmann and Rajmund G. Piotrowski (eds.), *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook* (Handbücher zur Sprach- und Kommunikationswissenschaft, 27), 214–226. Berlin, New York: de Gruyter.

Wheeler, Eric 2005. Multidimensional scaling for linguistics. In Reinhard Köhler, Gabriel Altmann and Rajmund G. Piotrowski (eds.), *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook* (Handbücher zur Sprach- und Kommunikationswissenschaft, 27), 548–553. Berlin, New York: de Gruyter

Wimmer, Gejza and Altmann, Gabriel. 1996. On vocabulary richness. *Journal of Quantitative Linguistics*. 6(1). 1–9.

Zörnig, Peter and Gabriel Altmann. 1983. The repeat rate of phoneme frequencies and the Zipf-Mandelbrot law. In Reinhard Köhler and Joachim Boy (eds.), *Glottometrika 5* (=Quantitative Linguistics, 20), 205–211. Bochum: Brockmeyer.