Classification of Author and/or Genre? The Impact of Word Length

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

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

Abstract. 190 Russian texts – letters and poems by three different authors – are analyzed as to their word length. The basic question concerns the quantitative classification of these texts as to authorship or as to text sort. By way of multivariate analyses it is shown that word length is a characteristic of genre, rather than of authorship.[1]

1 Word Length and the Quantitative Description of Text(s) and Author(s)

This study focuses on word length. Word length is a central characteristic in the framework of quantitatively oriented linguistics. In fact, the study of word length can be traced back to a hundred year long tradition (as to a historical and methodological survey of these studies, cf. Grzybek 2004). Knowing this historical background, it is evident that word length, as it is studied today, is no isolated characteristic.[2]

The basic question of the present study is to what degree word length may contribute to the discrimination of authors and genres. An answer to this question will not only shed light on specific factors influencing word length; it will also provide an argument if word length is an appropriate variable to describe an author's individual style, or the stylistic traits of specific genres.

The discussion of these questions has a history of its own: as opposed to the field of *quantitative typology of texts* (cf. Alekseev 1988, Pieper 1979), approaches in the realm of *stylometry* (cf. Martynenko 1988) assume that the individual style of texts and/or authors can be quantitatively described. Part of this research has concentrated on the question of authorship attribution, particularly applying quantitative methods to decide doubtful cases of authorship (cf. Marusenko 1990). In a way, these approaches have paved the

[1] This study has been conducted in context of research project # 15485 («Word Length Frequencies in Slavic Texts»), financially supported by the Austrian Research Fund (FWF); cf.: http://www-gewi.uni-graz.at/quanta.

[2] Within a synergetic approach, word length is closely interrelated with other linguistic levels and units, and it is well known that word length interacts, e.g., with the number of phonemes (in a given inventory), with lexicon size (cf. Köhler 1986), with polysemy (cf. Altmann et al. 1982), or word length and word frequency (Strauss et al. 2004, with a survey of the Zipfian tradition).

way for contemporary research in the field of computer linguistics, where related problems are being discussed under the heading of automatic authorship attribution and text categorization. The status of this contemporary research may be characterized by two tendencies. On the one hand, word length is not at all taken into consideration; in this case, researchers assume word length to be a "low-level phenomenon" (cf. Stamatatos et al. 2001: 195), which leads to no reliable results, neither for text categorization nor for authorship attribution. On the other hand, word length is taken into account as one possible variable among others (such as, e.g., sentence length, lexical type-token ratio, adverb counts, etc.) for multivariate discriminant analyses (vgl. Karlgren and Cutting 1994). As to this line of research, there are a number of methodological problems which have not been sufficiently reflected:

1. More often than not, word length has been measured as the number of characters per word; it is a well-known fact, however, that for most languages, measuring word length as the number of characters (letter, graphemes) per word is no appropriate procedure leading to erroneous results due to the instability of the graphemic system (cf. Kelih/Grzybek 2004);
2. Most of the studies in this field do not analyze the impact of word length as a variable in its own right, but only as part of some undifferentiated pool of variables.

This situation gives rise to a new systematic study of word length as a possible discriminating variable for authorship attribution and/or text categorization, paying due attention to and avoiding the methodological flaws of the studies mentioned above.

2 A Case Study: Text Basis and Analytical Options

With regard to the problems discussed above, the present study proceeds as follows:

a. Word length is measured as the number of syllables per word; 'word' is thus understood as an orthographical-phonological unit, the systematic changes of which, depending on linguistic definitions, are well known as well (cf. Antić et al. 2004).
b. Discriminant analyses are undertaken, taking into consideration only variables which are directly related to or derived from the frequency distribution of x-syllable words in a given text.

In the present study, the word length of 190 Russian texts is analyzed. These texts are systematically chosen in order to design a balanced study, based on an approximately equal number of two different text types, written by three different authors. By way of multivariate methods, the role of word length as a characteristic of authorship or of text type shall be studied.[3]

[3] The text basis is part of the text data base developed in the research project mentioned above.

In order to study the relevance of word length on the level of text sorts and authors, respectively, ca. 30 texts written by three well-known Russian authors each (A. S. Puškin, D. Charms, and A.A. Achmatova) in two different sorts of text (poems and letters), are considered. Table 1 represents the composition of the sample.

Table 1. 190 Russian Texts

AUTHOR(S)	**TEXT TYPE(S)**	**AMOUNT**
A.A. Achmatova	Letters	30
	Poems	30
D. Charms	Letters	29
	Poems	30
A.S. Puškin	Letters	36
	Poems	35
Total		190

On the basis of this text sample, a number of different analytical options are at our disposal (cf. Figure 1). These options include the discrimination

- of authors within a given genre (i.e., studying only letters or poems, respectively);
- of different texts sorts written by different authors (e.g., Charms' private letters in contrast to Achmatova's poems);
- of text sorts without paying attention to authorship.

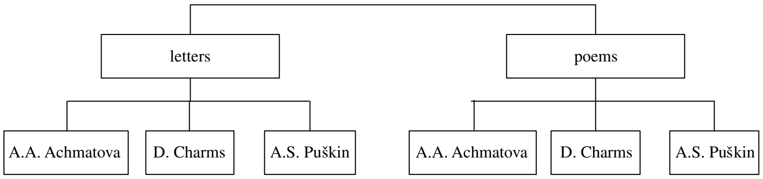

Fig. 1. Graphical Representation of the Text Data Base

3 Methods of Text Discrimination

As to the discrimination of author and/or text, we want to concentrate on the impact of word length, only. Therefore, from our pool of 30 possible discrimination variables, all those variables which are related to other characteristics of a text (such as, e.g., text length), will be excluded, as well as variables which, though primarily characterizing word length, have such factors as indirect components.[4]

3.1 Quantitative Measures for Text Analysis

Each text contains N words (w_i for $i = 1, 2, \ldots, N$). Word length (x_i) is measured in the number of syllables per word ($x_i = j$ where $i = 1, 2, \ldots N; j = 1, 2, \ldots K$). Actually we are dealing with words of $1, 2, 3, \ldots$, or K syllables. Words are divided into K frequency classes; f_j refers to the number of elements that belong to the same class (absolute frequencies). Texts will be quantitatively described by a number of measures reflecting the moments of the word length frequency distribution.

Not all variables which possibly describe the distribution are equally important for our study; our aim was to find a minimal set of variables, relevant for discriminant analyses (thus having the strongest classification power). On the basis of our empirical tests, we obtained a set of six variables, which are appropriate for our purposes. The definitions of these six variables are listed in Table 2.

Table 2. Six statistical measures characterizing 190 Russian texts

Variable	Formula	Explanation
m_2	$= s_0^2 = 1/N \cdot \sum\limits_{i=1}^{N} (x_i - \bar{x})^2$	empirical variance of the word length
m_4	$= 1/N \cdot \sum\limits_{i=1}^{N} (x_i - \bar{x})^4$	fourth central moment
v	$= s_0/m_1$	coefficient of variation
d	$= m_2/(m_1 - 1)$	quotient of dispersion
o_i	$= m_2/m_1$	first criterion of Ord
p_4	$= f_4/N$	relative proportion of four-syllable words

Every text, now, can be seen as a statistical object incorporating its information in the six variables listed in Table 2. Thus, the quantitative description of a given text j, belonging to group i, is given by an observation vector of dimension 6 (for $i = 1, 2\,; j = 1, \ldots, 95$):

$$\mathbf{x}_{ij} = (m_2\,(i,j)\,, m_4\,(i,j)\,, v\,(i,j)\,, d\,(i,j)\,, o_i\,(i,j)\,, p_4\,(i,j))$$

[4] Text length is, of course, an important characteristic of a text, and has well been used in other studies on authorship or genre discrimination (cf. Djuzelic 2002). Although in our case, the average text length of the letters ($\bar{x} = 238.20, s = 170.37$) does not significantly differ from that of the poems ($\bar{x} = 204.37, s = 178.59$) – as can be shown by a Mann/Whitney U-Test ($z = -1.56, p = 0.12$) – we have focused on word length, only, in order to strictly control the impact of this variable.

For each group, the mean values of the six variables are combined in the mean vector of the same dimension (for $i = 1, 2$):

$$\bar{\mathbf{x}}_i = \left(\bar{m}_2\,(i)\,, \bar{m}_4\,(i)\,, \bar{v}\,(i)\,, \bar{d}\,(i)\,, \bar{o}_i\,(i)\,, \bar{p}_4\,(i)\right)$$

Table 3 represents one example, including two Russian texts with all six statistical values discussed above. Actually, there are 95 texts from both genres in our text corpus. $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ denote the mean vector for the text groups, i.e., *letters* and *poems*, respectively, and they are calculated for all 95 texts of each group.

Table 3. Six statistical measures of two Russian texts for both text types

Text type	m_2	m_4	v	d	o_i	p_4
Letter #1	1.26	6.53	0.55	1.23	0.62	0.07
Letter #2	1.37	m7.07	0.50	1.01	0.58	0.16
$n_1 = 95\,;\ \ \bar{\mathbf{x}}_1 = ($	1.47	7.86	0.53	1.17	0.64	0.11)
Poem #96	0.81	2.04	0.45	0.83	0.41	0.04
Poem #97	0.86	2.92	0.49	0.97	0.46	0.04
$n_2 = 95\ \ \ \bar{\mathbf{x}}_2 = ($	0.92	2.57	0.47	0.88	0.45	0.06)

3.2 Discriminant Analysis

In a first step, the texts are discriminated along the category of ‘author’, only. In this case, each of our three authors – A.A. Achmatova {A}; D. Charms {C}; A.S. Puškin {P}) – is treated as a separate class, and no genre distinction is taken into consideration. As can be seen from Table 4[1], this results in a percentage of only 38.4% correctly discriminated texts (cf. Figure 2, which illustrates the finding that one cannot recognize any clearly separated group).

As can also be seen from Table 4[2], this poor result can be improved up to a percentage of 56%, if ‘genre’ is additionally taken into consideration. In the next step concentrating on one particular text group (i.e., either letters or poems), and testing each combination of two authors, one obtains definitely better results between 63% and 77% (cf. Table 4[3,4]). Finally concentrating on one individual author, only, and juxtaposing letters vs. poems, one gets even better results up to a percentage of 82% to 93% correctly classified texts (cf. Table 4[5]).

This overall result is a clear indication for word length being dependent on the type of text, rather than on authorship (i.e. being a good variable for text categorization, rather than authorship attribution).

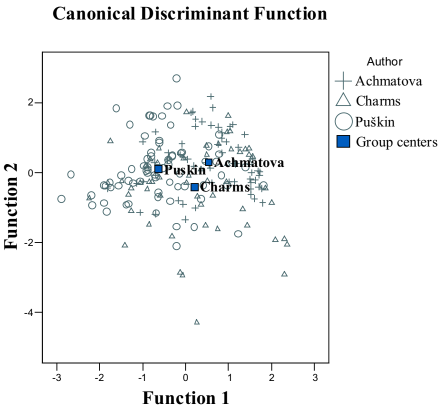

Fig. 2. Canonical discriminant function regarding three Russian authors

Table 4. Discriminant Analyses: Author vs. Genre

	Text Type	Author	Classification
$\boxed{1}$		{A}{C}{P}	38.40%
$\boxed{2}$	{Letters}{Poems}	{A}{C}{P}	46.30%
	{Letters}	{A}{C}{P}	55.80%
	{Poems}	{A}{C}{P}	54.70%
$\boxed{3}$	{Letters}	{A}{C}	62.70%
		{A}{P}	71.20%
		{C}{P}	67.70%
$\boxed{4}$	{Poems}	{A}{C}	76.70%
		{A}{P}	0.00%
		{C}{P}	73.80%
$\boxed{5}$	{Letters}{Poems}	{A}	81.70%
		{C}	93.00%
		{P}	93.20%

3.3 Statistical Distance as a Measure for Data Discrimination

Given these findings, it is important to see which relevant variables are appropriate for discriminant analyses. The univariate distance is an important measure for separating data corpora into two different text groups. Let us assume that the texts are independent samples $(x_{1_1}, \ldots, x_{1_{95}})$,$(x_{2_1}, \ldots, x_{2_{95}})$ of two distributions, which have possibly different theoretical means μ_i and the same variance σ^2. The theoretical means will be estimated by the arithmetic mean $\bar{x}_i$ of the sample, and the variance by pooling the two empirical variances s_i^2 of the sample as follows:

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

The univariate statistical distance D is given as:

$$D\left(\bar{x}_1, \bar{x}_2\right) = \frac{|\bar{x}_1 - \bar{x}_2|}{s_{pool}}$$

The distance D between two groups is thus defined as the distance between the group centers (means), standardized by the pooled variance. Table 5 contains mean values, standard deviations and univariate statistical distances for all six variables; also, results are given for all pairwise comparisons between these two text groups.

Table 5. Means, standard deviations and univariate statistical distances for pairwise comparisons (letters vs. poems)

Variable	**Text type**	$\bar{x}_1 \vert \bar{x}_2$	$s_1 \vert s_2$	$D(\bar{x}_1, \bar{x}_2)$
m_2	Letter	1.47	0.43	5.20
	Poem	0.92	0.17	
m_4	Letter	7.86	6.75	0.23
	Poem	2.57	1.09	
v	Letter	0.53	0.06	24.87
	Poem	0.47	0.03	
d	Letter	1.17	0.15	16.53
	Poem	0.88	0.11	
o_i	Letter	0.64	0.11	23.66
	Poem	0.45	0.06	
p_4	Letter	0.11	0.04	**36.17**
	Poem	0.06	0.03	

Table 5 shows the highest distance value D, based on the variable p_4 (i.e., the relative frequency of 4-syllable words). This means that variable p_4 has the strongest power for the separation of our text corpus into two groups: p_4 thus is the best discriminator for these two text groups.

The fourth central moment (m_4) has the lowest discrimination power, what implies a bad separation. The reason for this is the fact that although variable m_4 has the highest mean value, it has as large statistical deviation, which keeps the distance relatively small. Knowing that these two text groups remarkably differ as to the proportion of 4-syllable words, this result was to be expected. With variable p_4 alone, up to 76.3% of our texts can be correctly classified: combining p_4 with variable d, the percentage of correctly classified items improves to 89.5%. In Figure 3, variable p_4 is plotted against variable d for the two categories *letters* and *poems*.

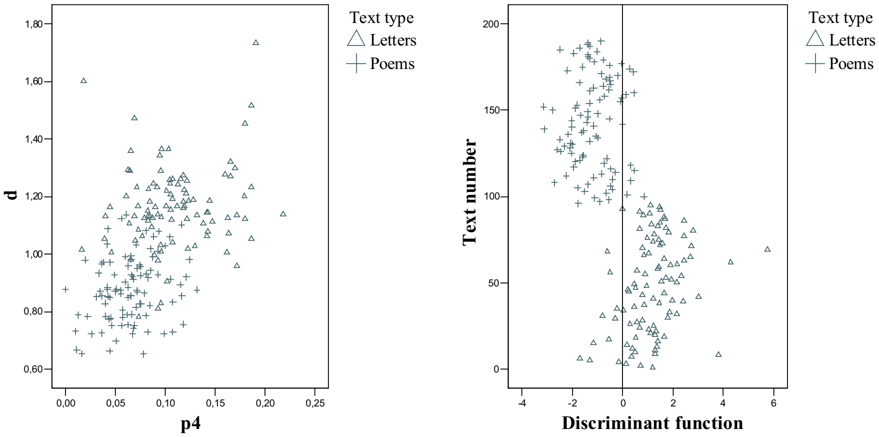

Fig. 3. Left scatter plot p_4 vs. d; right separation of letters and poems

Figure 3 illustrates the fact that it is possible to separate *letters* from *poems*. The linear discriminant function is calculated as a linear combination of relevant variables. In our case, the set of six variables is reduced to a set of two relevant variables, namely, p_4 and d. Figure 3 also shows the good separation power of the discriminant function. The cut point between the two groups is represented by the vertical line in 0, which marks the separation. Each point represents a text; the text numbers can be seen on the y-axis. Every text has different values of p_4 and d, so the value of the discriminant function is also different for each text: we can see two clearly separated groups. We can notice that only nine *poems* and eleven *letters* are misclassified. This corresponds to a high percentage of correct classifications, which sum up to 90.5%, or 88%, respectively.

4 Summary

Our study clearly shows that word length, if properly defined as the number of syllables per word, has a strong discriminating power for text categorization: with only two variables, a percentage of up to 90% correctly discriminated texts can be obtained. As opposed to this, word length does not seem to play an important role as to questions of authorship attribution.

References

ALEKSEEV, P.M. (1988):*Kvantitativnaja lingvistika teksta.* Leningrad.

ALTMANN, G., BEÖTHY, E., and BEST, K.H. (1982): Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 35, 537-543.*

ANTIĆ, G., KELIH, E., and GRZYBEK, P. (2004): Zero-syllable Words in Determining Word Length. In: P. Grzybek (Ed.): *Contributions to the Science of Language. Word Length Studies and Related Issues.* [In print]

DJUZELIC, M. (2002): *Einflussfaktoren auf die Wortlänge und ihrer Häufigkeitsverteilung am Beispiel von Texten slowenischer Sprache.* Dipl. Arbeit, TU Graz.

GRZYBEK, P. (2004): History and Methodology of Word Length Studies: The State of the Art. In: P. Grzybek (Ed.): *Contributions to the Science of Language: Word Length Studies and Related Issues.* [In print]

KARLGREN, J., and CUTTING, D. (1994): Recognizing text genres with simple metrics using discriminant analysis. In: *Proceedings of COLING 94, Kyoto, 1994.* [`http://www.sics.se/~jussi/Papers/1994_Coling_Kyoto_l/cmplglixcol.ps`]

KELIH, E., and GRZYBEK, P. (2004): Wortlänge in Silben und Graphemen. [In prep.]

KÖHLER, R. (1986): *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Studienverlag Brockmeyer, Bochum. [= Quantitative Linguistics; 31]

MARTYNENKO, G.Ja. (1988): *Osnovy stilemetrii.* Leningrad.

MARUSENKO, M.A. (1990): *Atribucija anonimnych i psevdoanonimnych literaturnych proizvedenij metodami raspoznavanija obrazov.* Leningrad.

PIEPER, U. (1979): *Über die Aussagekraft statistischer Methoden für die linguistische Stilanalyse.* Tübingen, Narr. (= Ars linguistica, 5)

STAMATATOS, E.; FAKOTAKIS, N., and KOKKINAKIS, G. (2001): Computer-Based Authorship Attribution Without Lexical Measures In: Computers and the Humanities, 35,193-214.

STRAUSS, U., GRZYBEK, P., and ALTMANN, G. (2004): Word length and word frequency. In: P. Grzybek (Ed.): *Contributions to the Science of Language. Word Length Studies and Related Issues.* [In print]