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.[3]

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.[4]

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

[3] 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`.

[4] 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).

doubtful cases of authorship (cf. Marusenko (1990)). In a way, these approaches have paved the 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 and 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