Journal of Quantitative Linguistics
2002, Vol. 9, No. 2, pp. 187–192

0929-6174/02/0902-187$16.00

PROJECT REPORT

The Graz Project on Word Length (Frequencies)

Peter Grzybek[1,*] and Ernst Stadlober[2,*]
[1]Institut für Slawistik, Universität Graz, Graz, Austria, and [2]Institut für Statistik, Technische Universität Graz, Graz, Austria

Beginning with the year 2002, the Austrian Science Foundation (Fonds zur Förderung wissenschaftlicher Forschung, FWF, Vienna; project no.: P-15485) will financially support a 3-year project on the study of word length in texts from Slavic languages. Particular emphasis is laid on the frequency distribution (FD) of word length and on factors influencing word length (FDs). The interdisciplinary and inter-university project, in which specialists from linguistics, text scholarship, computer science and statistics are involved, is headed by Peter Grzybek (Institute for Slavic Studies, University of Graz) and Ernst Stadlober (Institute of Statistics, Graz University of Technology).

The general theoretical horizon of the project is rather broad and daring: it is claimed that the application of quantitative (in complementary addition to qualitative) methods in the field of the humanities is one way to overcome the still current myth of the 'two cultures' (as conceived by Snow and as upheld by a never ending number of adherents to this theory). As soon as one understands 'culture' and 'nature' as specific (cultural) constructs, however, the perspective changes, and two options emerge as how to deal with the alleged juxtaposition of these two concepts: on the one hand, the historically changing definitions of 'culture' and 'nature' (and reasons for these definitions) can be rendered a scholarly topic, and on the other hand, the convergences between them should be made a scholarly object in its own

*Address correspondence to: Peter Grzybek, Institut für Slawistik, Universität Graz, Merangasse 70, A-8010 Graz. E-mail: grzybek@uni-graz.at or Ernst Stadlober, Institut für Statistik, Technische Universität Graz, Steyrergasse 17/IV, A-8010 Graz. E-mail: stadlober@stat.tu-graz.ac.at (See also: http://www-gewi.uni-graz.at/quanta)

right. In fact, the latter approach characterizes the theoretical background of the Graz project: Whereas language and linguistic texts are regarded as particular cultural products and specific sign systems within a cultural system, statistics is considered to be an appropriate meta-language for cultural studies in general, and for linguistic studies as one specific instance of them.

As compared to these rather daring general assumptions, the concrete objectives of the Graz project must be classified as relatively modest, since 'only' word length, word length frequencies, and factors influencing them are considered. Yet the scope consequently pursued is innovative, if one takes into account the rather young history of word length studies.

A SHORT HISTORY OF WORD LENGTH STUDIES

The word, just like the sentence, is a central element for any (process of) text construction. Despite this central role, word length as a theoretical category in its own right has been largely neglected in linguistics and text-oriented disciplines. Only recently, accompanied by the rise of synergetic linguistics, the question of the frequency of occurrence of words of specific lengths ('word length frequencies') in texts (of a given language, a given author, a given genre, etc.) has been theoretically integrated in systematic contexts, and only recently a particular theory of word length distribution(s) has been developed.

Of course, there are quite a number of earlier approaches, mainly from the 1950s and 60s, when structuralism and information theory were in their heyday. At that time, quantitative aspects of the word were considered to be relevant for the stylistic study of author specific or discourse specific characteristics. Usually these approaches concentrated on the average length of words – as is well-known, a proposal brought forth as early as 1851 by the English mathematician and logician Augustus de Morgan (1806–71).

It is obvious that in these earlier approaches due attention has been paid to the fact that mean values can be based on quite different concrete frequency distributions and types of distributions, and, as a result, can vary to differing degrees. Consequently, the relevant works within so-called quantitative stylistics have considered not only mean length, but also variance as a specific textual characteristic. Still, from a contemporary point of view, there are at least two major objections to the earlier approaches mentioned:

1. Mean value and variance are only two specific characteristics of the distribution; without doubt, they offer a correct, but restricted look at the structure of the complete data set. To arrive at more solid results, it would therefore be necessary to analyze further measures, such as, to name but a few, the standard errors of mean length, of the standard deviation, or of the median; also, the (relative) variation coefficient and its standard error; the dispersion index and its standard error; skewness, kurtosis and their standard errors, and many more. Also, various measures based on entropy and its variance, such as (relative) redundancy, repeat rate, and so on should be analyzed in detail. Djuzelic, Grzybek, and Stadlober (2002) have listed many relevant measures; the importance of these various measures as well as theoretical interrelationships between them will have to be studied both theoretically and empirically.
2. More often than not, the analysis of word length has concentrated on mean length or variance, not (only) of individual texts, but of (more or less clearly defined) text corpora, be it from a given author, a given literary period, a particular functional style, or from other sources. Based on the (wrong) assumption that by accumulating as many texts as possible (of a given language or genre, by a particular author, etc.), one tried to find something like a 'norm', neglecting the fact that every text is the specific (individual) result of a text generating (producing) process, which is guided by particular linguistic and/or psycholinguistic regularities. Therefore, contemporary quantitative linguistics assumes that a text corpus is nothing but a heterogeneous text mixture (a 'quasi text'), and that, as a consequence, there can be no text accumulation which might be homogeneous enough to have constant parameters. For the concrete study of word length (frequency), the existence of 'local' influencing factors (such as authorship, genre, functional style, etc.) implies the separate analysis of individual texts, rather than of text corpora.
3. Any particular measure (or any combination of measures) characterizing a specific frequency distribution does not allow for an answer to the question of how these specific measures are motivated by the frequency distribution itself, that is for an answer to the question, how the individual frequencies of i-syllable words are combined in a specific type of frequency distribution. Empirical results thus far available indeed show that the frequency with which one-, two-, three-, (etc.) syllable words occur in texts, is organized not chaotically, but by specific laws; knowledge of these laws may allow deeper insights into text structure and processing.

As opposed to earlier assumptions that a single, unique law might be responsible for the frequencies of word length in texts (cf. Ebanov, Fucks, and others), one now takes into consideration a flexible system of a super-ordinate basic model and particular modifications due to text-, author-, genre-, time-specific and other factors (cf. Wimmer, Köhler, Grotjahn, & Altmann 1994; Wimmer & Altmann, 1996).

Still, the crucial question of which factors may have an influence on word length (frequencies), or how these factors may interact, has not yet been studied systematically. So this is the project's starting point. Theoretically speaking, there are two options as to how influencing factors might come into play:

a. According to the first assumption the impact of factors such as authorship, genre, diachrony, and so on will result in different frequency distribution models. If this plausible hypothesis should turn out to be true, it would be important to come to know whether the models relevant for texts of a given language may be interpreted as special cases of a common, super-ordinate model.
b. According to the second assumption the mentioned factors may lead to different models; in this case, these factors would only have a minor impact on the specific frequency distribution model, but a significant influence on the specific parameters of a given model.

Both variants will be (and, of course, have been) observed in real texts, but up to now, systematic studies of this particular question have never been undertaken. Hence the project's objectives are mainly concentrated on the following three points:

1. The systematic study of word length frequency distributions in texts from three different Slavic languages (Croatian, Russian, Slovenian) aims at the distinction between language specific vs. factors common across languages.
2. The systematic study of an individual author's style, of text typology, etc. aims at the isolation of factors possibly influencing word length and its frequency distribution in texts; the relevance of these factors and possible interactions between them will have to be studied in a second step.
3. Assuming that there are positive answers to (1) and (2), the direction of asking the question may be reversed, now aiming at text identification and text classification. In other words: Knowing the relevant measures to be studied in steps (1) and (2), will it be possible to attribute individual texts to specific authors, text types, and so on, with a given probability? Or, asked more modestly: What contribution do word length studies have when trying to find an answer to these questions?

As to the general 3-year schedule, the project is divided into three consecutive 1-year phases:

1. Initially, it will be necessary to construct a profound text data bank with approximately 1000 texts in each of the three Slavic languages (Russian, Croatian, Slovenian) under study, accompanied by the relevant meta-data. In composing the text corpus, particular emphasis will have to be put on questions of text typology in order to provide an adequate textual basis for the statistical analyses in a later stage. Optionally, the data bank shall be made available for external usage as well, provided an adequate corpus interface will be available. Also, specific software tools for text analysis will have to be constructed (if possible, allowing for further developments, possibly including non-Slavic languages).
2. In the next step, the accumulated texts will have to be prepared for the analyses, as far as the textual basis is concerned: This work not only includes a unitary treatment of abbreviations, headings, numbers, foreign words, and so on, but also the definition of a sentence, of how narrative, descriptive, dialogical, etc. sequences can be distinguished. Given these textual preparations, first statistical analyses can be run in order to provide the raw data for each text, which will have to be stored in adequate files for the relevant statistical programs to be used in the last stage.
3. The last phase includes quantitative and qualitative analyses: Now, it will be necessary to find an adequate distribution model (or various distribution models, depending on the results to be obtained). Further analyses will then have to concentrate on the factors influencing either the distribution's parameters or the distribution type as a whole. This work is supposed to result in discrimination and cluster analyses, yielding deep insight into text typology, text classification and text discrimination.

If in fact relevant structures and regularities will have to be observed, they will have to be understood as being of importance for information processing in general. In this case, the statistical methods applied will represent an optimal basis for subsequent inter-disciplinary studies further attempting to bridge the 'two cultures' of natural and cultural sciences.

REFERENCES

Djuzelic, M., Grzybek, P., & Stadlober, E. (2002). *Statistische Kenngrößen zur Beschreibung von Wortlängenhäufigkeitsverteilungen*. Unpublished manuscript.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics, 1*, 98–106.

Wimmer, G., & Altmann, G. (1996). The theory of word length distribution: Some results and generalizations. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 112–133). Trier: WVT.