Journal of Quantitative Linguistics
2002, Vol. 9, No. 3, pp. 275–279

0929-6174/02/0903-275$16.00

CONFERENCE REPORT

"Word Length in Texts. An International Symposium on Quantitative Text Analysis." Institute for Slavic Studies, Graz University, June 21–23, 2002*

Gordana Anti, Emmerich Kelih, and Peter Grzybek
Institute for Slavic Studies, Graz University, Graz, Austria

INTRODUCTION

The beginning of the research project "Word length (frequencies) in Slavic Language Texts" under the leadership of *Peter Grzybek* (Graz University) and *Ernst Stadlober* (Technical University of Graz), supported by the Austrian Fund for Scientific Research (cf., Grzybek & Stadlober, 2002), gave the opportunity for an international conference on "Word Length in Texts. An International Symposium on Quantitative Text Analysis."

The opening and the presentations by the participants from Graz took place in the facilities of the Institute for Slavic Studies, June 21st, 2002; the second part of the conference, from June 22nd to June 23rd, took place in the nearby castle Seggau, which proved to be an ideal place for all other lectures and discussions, in particular, because one could find enough time for the exchange of thoughts and mutual information.

The introductory speech by *Peter Grzybek* (Graz, Austria) clearly showed that the conference was conceived as an open interdisciplinary forum of scholars involved in Quantitative Linguistics, Mathematics, Statistics, Corpus- and Psycholinguistics, general Speech and Text-Science as well as Information Technology. Apart from a presentation of the FWF project,[1] Grzybek himself, in his introductory speech, presented the historic

*Address correspondence to: Peter Grzybek, Institute for Slavic Studies, Graz University, Merangasse 70, A-8010 Graz, Austria. Tel.: +43 316 380 2526. Fax: +43 316 380 9773. E-mail: grzybek@gewi.kfunigraz.ac.at
[1]See http://www.gewi.kfunigraz.ac.at/quanta

developments of word length studies: beginning with Augustus de Morgan, who was the first to point at word length as a specific characteristic of different authors in the middle of the 19th century, up to the contemporary approach by Altmann, Grotjahn, Köhler and Wimmer. Based on this historic context, the intentions and targets of the Graz Project have been implemented in the present state of research: accordingly, it is necessary to conduct systematic studies order to study how authorship, text type, temporal factors, etc., may influence the frequency distribution of the word length in texts.

As far as the contextual aspect of the symposium is concerned, the lectures were scheduled so that there was enough time to treat and discuss various, contextually independent, yet mutually related topics.

MATHEMATICS/STATISTICS

The paper "Unified Derivation of Some Linguistics Laws", written by *Gejza Wimmer* (Bratislava, Slovakia) und *Gabriel Altmann* (Lüdenscheid, Germany, not personally present at the conference), is of great importance for the theoretical base of Quantitative Linguistics. In this significant paper it is proved that well-known linguistic laws may be traced back to a common root. Wimmer and Altmann trace back well-known linguistic laws (Menzerath, Piotrovskij, Zipf-Mandelbrot and others) to a common root. This proves that various language processes are obviously part of a common mechanism.

In his paper "A Word Length Model Based on the Generalized One-Displaced Poisson-uniform Distribution", *Viktor Kromer* (Novosibirsk, Russia) showed that for the description of the frequency distribution of word length in various languages, a common model, precisely the one-displaced, generalized Poisson distribution might be appropriate. Extending this distribution with one further parameter, specific statements about some other text types can be made.

The third paper, whose authors are *Ernst Stadlober* (Graz, Austria) and *Mario Djuzelic* (Graz, Austria), associates of the Graz Word Length Project, approached the topic from a statistical-methodological point of view. Their paper "Multivariate Statistic Methods of Quantitative Text Analysis" was related to the question whether word length as a potential variable might be included in a quantitative text-typology.

STATISTICS/LINGUISTICS

Gordana Anti (Graz, Austria) and *Emmerich Kelih* (Graz, Austria), also associates in the above mentioned project, tried to bridge the gap between statistics and "traditional linguistics". Apart from the general problem of defining linguistic units and related issues of the research field (phoneme, grapheme, syllable, morpheme), their paper "On the Question of the so-called 0-syllable Words in Determining Word Length" discusses the question how to treat words without vowels (0-syllable words), often occurring in Slavic languages.

In his paper "Word Length as a Basis for a Typological Study of the Slavic Languages", *Otto Rottmann* (Hagen, Germany) presented empirical results with regard to the modeling of the distribution of word length in Slavic language texts. In this respect, he also presented a possible new approach in the quantitative typology of the Slavic language.

There were two papers concerning the Menzerathian laws and appropriate empirical verification and modification: *Werner Lehfeldt* (Göttingen, Germany) presented the genuine Slavic issues of the Slavic "Jerchange" (historic loss of vowels in week positions) in the paper "The Slavic Jerchange in Light of the Menzerathian Law" in connection with the Menzerath-Altmann laws. This law is also valid – as can be seen in the paper "The Length of Affixes as a Basic Menzerathian Regularity" presented by *Anatolij A. Polikarpov* (Moscow, Russia) – in the process of word formation: According to the Russian language material, the regularity between morpheme length and the position of morphemes (affixes, prefixes, root) can be proved within a word.

QUANTITATIVE LINGUISTICS/PSYCHOLINGUISTICS

Within the broad framework of quantitative linguistics and the explicitly interdisciplinary concept of the conference, two significant papers as to the synthesis of Linguistics, Statistics and Psychology were presented.

Gertraud Fenk-Oczlon (Klagenfurt, Austria), in her paper "Frequency, Length and Variability of Case-forms", discussed factors possibly having impact on the frequency, or the length on grammatical categories – as, for example, case-forms in Russian can have: the most frequent case-forms tend to accumulate and to vary. In a further paper by Gertraud Fenk-Oczlon – in co-operation with *August Fenk* (Klagenfurt, Austria) – entitled: "The Decay of

Function Words in the Recall of Sentences of Different Size", and based on an experimental test, the authors pointed out the relation between reproduction frequency of content and function words in real sentences.

INFORMATICS/QUANTITATIVE LINGUISTICS

The paper "Software Technical Considerations of a Multifunctional Corpus Interface for Quantitative Research" by *Reinhard Köhler* (Trier, Germany), was, although theoretically oriented, practically very important for automation-supported text-analyses: The theoretically fundamental requirements for building text data bases and the connection of software tools for analyses were considered.

Next, *Rudolf Schlatte* (Graz, Austria), in his contribution "Data Bases and Meta Data Bases: Quantitative Text Analysis and Text Administration", reported about the data base structure in the Graz project. This data base is not yet completed and, at present, planned for internal use, only; it will contain texts from three Slavic languages: Croatian, Russian, Slovenian.

CORPUS LINGUISTICS/QUANTITATIVE LINGUISTICS

The question of automatic text analysis and the construction of text corpora was considered by four scholars, who argued in favor of the analysis of (national) corpora, following the idea of Czech national corpus (Český narodní korpus").

Marko Tadi (Zagreb, Croatia), in his presentation "Developing the Croatian National Corpus (HNK)", except from presenting the Croatian national corpus, pointed out that Corpus Linguistics is to be seen as a discipline directly related to Quantitative Linguistics, and that in further interdisciplinary work, desirable synergetic effects should result from this.

Primo Jakopin (Ljubljana, Slovenia), on the one hand presented existing Slovenian text corpora (Fida, Nova Beseda); on the other hand, he discussed his own results in the field of "Word Frequencies in Slovenian".

Cvetana Krstev (Belgrade, Yugoslavia) presented the first steps in building a Serbian text corpus ("Corpora in Serbia and their Exploitation"), and *Duško Vitas* (Belgrade, Yugoslavia), in his lecture "Word Statistics in Serbian",

reflected problems of the formal classification of lexical units in the construction of text corpora.

CONCLUSION

On one hand, the conference showed the wide range of contemporary questions and research results in the field of Quantitative Linguistics: from a mathematical-statistical theoretical background, empirical tests of regularities in languages and texts, up to the technical realization and setting of corpus and text data bases. On the other hand, it became clear that such a wide spectrum of issues indicated the necessity of interdisciplinary and international cooperation.

Finally a remark as to the organization: The event intended to bring together not only participants from the fields of Quantitative Linguistics, Corpus and Psycholinguistics, Linguistics, Mathematics, Statistics, Informatics and "traditional" Slavic Sciences, but also scholars from the fields of German and Slavic Languages, in order to organize a meeting, realizing a significant transfer of knowledge.

Organizing a conference of this kind would not have been possible without the financial support provided by the following institutions: the Graz University (Vice Rector Office for Research and Teaching, Bureau for International Relations, Dean Office of the Faculty of Humanities, Institute for Slavic Studies), The Austrian East and South European Institute at Vienna (with an office in Bratislava), the country of Styria, as well as the city of Graz.

All papers presented at the conference will be published at the beginning of 2003 under the title "Word Length Studies", edited by Peter Grzybek.

The organizers plan expressed their hope that this conference should initiate further meetings.