Peter Grzybek
Editor

TEXT, SPEECH AND LANGUAGE TECHNOLOGY SERIES 31

Contributions to the Science of Text and Language

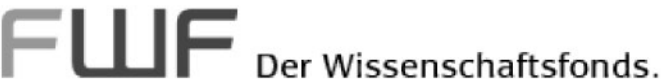
Word Length Studies and Related Issues

Springer

Contributions to the Science of Text and Language

Text, Speech and Language Technology

VOLUME 31



Veröffentlicht mit Unterstützung des Fonds zur Förderung der Wissentschaftlichen Forschung.

Contributions to the Science of Text and Language

Word Length Studies and Related Issues

Edited by

Peter Grzybek

University of Graz, Austria

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-1-4020-4069-6 (PB)
ISBN 978-1-4020-4067-2 (HB)
ISBN 978-1-4020-4068-9 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Printed on acid-free paper

Dedicated to all those pioneers in the field of quantitative linguistics and text analysis, who have understood that quantifying is not the aim, but a means to understanding the structures and processes of text and language, and who have thus paved the way for a theory and science of language

Contents

Preface

The studies represented in this volume have been collected in the interest of bringing together contributions from three fields which are all important for a comprehensive approach to the quantitative study of text and language, in general, and of word length studies, in particular: first, scholars from linguistics and text analysis, second, mathematicians and statisticians working on related issues, and third, experts in text corpus and text data bank design.

A scientific research project initiated in spring 2002 provided the perfect opportunity for this endeavor. Financially supported by the Austrian Research Fund (*FWF*), this three-year project, headed by Peter Grzybek (Graz University) and Ernst Stadlober (Technical University Graz) concentrates on the study of word length and word length frequencies, with particular emphasis on Slavic languages. Specifically, factors influencing word length are systematically studied.

The majority of contributions to be found in this volume go back to a conference held in Austria at the very beginning of the project, at Graz University and the nearby Schloss Seggau in June, 2002.[1] Experts from all over Europe were invited to contribute, with a particular emphasis on the participation of scholars from East European countries whose valuable work continues to remain ignored, be it due to language barriers, or to difficulties in the accessibility of their publications. It is the aim of this volume to contribute to a better mutual exchange of ideas.

Generally speaking, the aim of the conference was to diagnose and to discuss the state of the art in word length studies, with experts from the above-mentioned disciplines. Moreover, the above-mentioned project and the guiding ideas behind it should be presented to renowned experts from the scientific community, with three major intentions: first, to present the basic ideas as to the problem outlined, and to have them discussed from an external perspective in order to

[1] For a conference report see Grzybek/Stadlober (2003), for further details see `http://www-gewi.uni-graz.at/quanta`.

profit from differing approaches; second, to raise possible critical points as to the envisioned methodology, and to discuss foreseeable problems which might arise during the project; and third, to discuss, at the very beginning, options to prepare data, and analytical procedures, in such a way that they might be publicly useful and available not only during the project, but afterwards, as well.

Since, with the exception of the introductory essay, the articles appear in alphabetical order, they shall be briefly commented upon here in relation to their thematic relevance.

The introductory contribution by **Peter Grzybek** on the *History and Methodology of Word Length Studies* attempts to offer a general starting point and, in fact, provides an extensive survey on the state of the art. This contribution concentrates on theoretical approaches to the question, from the 19th century up to the present, and it offers an extensive overview not only of the development of word length studies, but of contemporary approaches, as well.

The contributions by **Gejza Wimmer** from Slovakia and **Gabriel Altmann** from Germany, as well as the one by **Victor Kromer** from Russia, follow this line of research, in so far as they are predominantly theory-oriented. Whereas Wimmer and Altmann try to achieve an all-encompassing *Unified Derivation of Some Linguistic Laws*, Kromer's contribution *About Word Length Distribution* is more specific, concentrating on a particular model of word length frequency distribution.

As compared to such theory-oriented studies, a number of contributions are located at the other end of the research spectrum: concentrating less on mere theoretical aspects of word length, they are related to the authors' work on text corpora. Whereas **Reinhard Köhler** from Germany, understanding a *Text Corpus as an Abstract Data Structure*, tries to generally outline *The Architecture of a Universal Corpus Interface*, the contributions by **Primož Jakopin** from Slovenia, **Marko Tadić** from Croatia, and **Duško Vitas, Gordana Pavlović-Lažetić, & Cvetana Krstev** from Belgrade concentrate on the specifics of Croatian, Serbian, and Slovenian corpora, with particular reference to word-length studies. Jakopin's contribution *On Text Corpora, Word Lengths, and Word Frequencies in Slovenian*, Tadić's report on *Developing the Croatian National Corpus and Beyond*, as well as the study *About Word Length Counting in Serbian* by Vitas, Pavlović-Lažetić, and Krstev primarily intend to discuss the availability and form of linguistic material from different text corpora, and the usefulness of the underlying data structure of their corpora for quantitative analyses. From this point of view their publications show the efficiency of co-operations between the different fields.

Another block of contributions represent concrete analyses, though from differing perspectives, and with different objectives. The first of these is the analysis by **Andrew Wilson** from Great Britain of *Word-Length Distribution*

in Present-Day Lower Sorbian. Applying the theoretical framework outlined by Altmann, Wimmer, and their colleagues, this is one example of theoretically modelling word length frequencies in a number of texts of a given language, Lower Sorbian in this case. **Gordana Antić, Emmerich Kelih, & Peter Grzybek** from Austria, discuss methodological problems of word length studies, concentrating on *Zero-Syllable Words in Determining Word Length*. Whereas this problem, which is not only relevant for Slavic studies, usually is "solved" by way of an authoritative decision, the authors attempt to describe the concrete consequences arising from such linguistic decisions. Two further contributions by **Ernst Stadlober & Mario Djuzelic** from Graz, and by **Otto A. Rottmann** from Germany, attempt to apply word length analysis for typological purposes: thus, Stadlober & Djuzelic, in their article on *Multivariate Statistical Methods in Quantitative Text Analyses*, reflect their results with regard to quantitative text typology, whereas Rottmann discusses *Aspects of the Typology of Slavic Languages Exemplified on Word Length*.

A number of further contributions discuss the relevance of word length studies within a broader linguistic context. Thus, **Simone Andersen & Gabriel Altmann** (Germany) analyze *Information Content of Words in Texts*, and **August Fenk & Gertraud Fenk-Oczlon** (Austria), study *Within-Sentence Distribution and Retention of Content Words and Function Words*.

The remaining three contributions have the common aim of shedding light on the interdependence between word length and other linguistic units. Thus, both **Werner Lehfeldt** from Germany, and **Anatolij A. Polikarpov** from Russia, place their word length studies within a Menzerathian framework: in doing so, Lehfeldt, in his analysis of *The Fall of the Jers in the Light of Menzerath's Law*, introduces a diachronic perspective, Polikarpov, in his attempt at *Explaining Basic Menzerathian Regularity*, focuses the *Dependence of Affix Length on the Ordinal Number of their Positions within Words*. Finally, **Udo Strauss, Peter Grzybek, & Gabriel Altmann** re-analyze the well-known problem of *Word Length and Word Frequency*; on the basis of their study, the authors arrive at the conclusion that sometimes, in describing linguistic phenomena, less complex models are sufficient, as long as the principle of data homogeneity is obeyed.

The volume thus offering a broad spectrum of word length studies, should be of interest not only to experts in general linguistics and text scholarship, but in related fields as well. Only a closer co-operation between experts from the above-mentioned fields will provide an adequate basis for further insight into what is actually going on in language(s) and text(s), and it is the hope of this volume to make a significant contribution to these efforts.

This volume would not have seen the light of day without the invaluable help and support of many individuals and institutions. First and foremost, my thanks goes to Gabriel Altmann, who has accompanied the whole project from its very beginnings, and who has nurtured it with his competence and enthusiasm

throughout the duration. Also, without the help of the Graz team, mainly my friends and colleagues Gordana Antić, Emmerich Kelih, Rudi Schlatte, and of course Ernst Stadlober, this book could not have taken its present shape.

Furthermore, it is my pleasure and duty to express my gratitude to the following for their financial support: first of all, thanks goes to the Austrian Science Fund (*FWF*) in Vienna for funding both research project # P15485 ("Word Length Frequencies in Slavic Language Texts"), and the present volume. Sincere thanks as well goes to various institutions which have repeatedly sponsored academic meetings related to this volume, among others: Graz University (Vice Rector for Research and Knowledge Transfer, Vice Rector and Office for International Relations, Faculty for Cultural Studies, Department for Slavic Studies), Technical University Graz (Department for Statistics), Office for the Government of the Province of Styria (Department for Science), Office of the Mayor of the City of Graz.

Finally, my thanks goes to Wolfgang Eismann for his help in interpreting some Polish texts, and to Bríd Ní Mhaoileoin for her careful editing of the texts in this volume.

Preparing the layout of this volume myself, using TEXor LATEX 2_ε, respectively, I have done what I could to put all articles into an attractive shape; any remaining flaws are my responsibility.

PETER GRZYBEK

INTRODUCTORY REMARKS: ON THE SCIENCE OF LANGUAGE IN LIGHT OF THE LANGUAGE OF SCIENCE

Peter Grzybek

The seemingly innocent formulation as to a *science of language* in light of the *language of science* is more than a mere play on words: rather, this formulation may turn out to be relatively demanding, depending on the concrete understanding of the terms involved – particularly, placing the term 'science' into a framework of a general theory of science. No doubt, there is more than one theory of science, and it is not the place here to discuss the philosophical implications of this field in detail. Furthermore, it has become commonplace to refuse the concept of a unique theory of science, and to distinguish between a general theory of science and specific theories of science, relevant for individual sciences (or branches of science). This tendency is particularly strong in the humanities, where 19th century ideas as to the irreconcilable antagony of human and natural, of weak and hard sciences, etc., are perpetuated, though sophisticatedly updated in one way or another.

The basic problem thus is that the understanding of 'science' (and, consequently, the far-reaching implications of the understanding of the term) is not the same all across the disciplines. As far as linguistics, which is at stake here, is concerned, the self-evaluation of this discipline clearly is that it fulfills the requirements of being a science, as Smith (1989: 26) correctly puts it:

> Linguistics likes to think of itself as a science in the sense that it makes testable, i.e. potentially falsifiable, statements or predictions.

The relevant question is not, however, to which extent linguistics considers itself to be a science; rather, the question must be, to which extent does linguistics satisfy the needs of a general theory of science. And the same holds true, of course, for related disciplines focusing on specific language products and processes, starting from subfields such as psycholinguistics, up to the area of text scholarship, in general.

Generally speaking, it is commonplace to say that there can be no science without theory, or theories. And there will be no doubt that theories are usually

P. Grzybek (ed.), Contributions to the Science of Text and Language: Word Length Studies and Related Issues, 1–14.

conceived of as models for the interpretation or explanation of the phenomena to be understood or explained. More often than not, however, linguistic understandings of the term 'theory' are less "ambitious" than postulates from the philosophy of science: linguistic "theories" rather tend to confine themselves to being conceptual systems covering a particular aspect of language. Terms like 'word formation theory' (understood as a set of rules with which words are composed from morphemes), 'syntax theory' (understood as a set of rules with which sentences are formed), or 'text theory' (understood as a set of rules with which sentences are combined) are quite characteristic in this respect (cf. Altmann 1985: 1). In each of these cases, we are concerned with not more and not less than a system of concepts whose function it is to provide a consistent description of the object under study. 'Theory' thus is understood in the descriptive meaning; ultimately, it boils down to an intrinsically plausible, coherent descriptive system (cf. Smith 1989: 14):

> But the hallmark of a (scientific) theory is that it gives rise to hypotheses which can be the object of rational argumentation.

Now, it goes without saying that the existence of a system of concepts is necessary for the construction of a theory: yet, it is a necessary, but not sufficient condition (cf. Altmann 1985: 2):

> One should not have the illusion that one constructs a theory when one classifies linguistic phenomena and develops sophisticated conceptual systems, or discovers universals, or formulates linguistic rules. Though this predominantly descriptive work is essential and stands at the beginning of any research, nothing more can be gained but the definition of the research object [...].

What is necessary then, for science, is the existence of a theory, or of theories, which are systems of specific hypotheses, which are not only plausible, but must be both deduced or deducible from the theory, and tested, or in principle be testable (cf. Altmann 1978: 3):

> The main part of a theory consists of a system of hypotheses. Some of them are empirical (= tenable), i.e. they are corroborated by data; others are theoretical or (deductively) valid, i.e. they are derived from the axioms or theorems of a (not necessarily identical) theory with the aid of permitted operations. A scientific theory is a system in which some valid hypotheses are tenable and (almost) no hypotheses untenable.

Thus, theories pre-suppose the existence of specific hypotheses the formulation of which, following Bunge (1967: 229), implies the three main requisites:

(i) the hypothesis must be *well formed* (formally correct) and *meaningful* (semantically nonempty) in some scientific context;

(ii) the hypothesis must be *grounded* to some extent on previous knowledge, i.e. it must be related to definite grounds other than the data it covers; if entirely novel it must be compatible with the bulk of scientific knowledge;

(iii) the hypothesis must be empirically testable by the objective procedures of science, i.e. by confrontation with empirical data controlled in turn by scientific techniques and theories.

In a next step, therefore, different levels in conjecture making may thus be distinguished, depending on the relation between hypothesis (h), antecedent knowledge (A), and empirical evidence (e); Figure1.1 illustrates the four levels.

(i) *Guesses* are unfounded and untested hypotheses, which characterize speculation, pseudoscience, and possibly the earlier stages of theoretical work.

(ii) *Empirical hypotheses* are ungrounded but empirically corroborated conjectures; they are rather isolated and lack empirical validation, since they have no support other than the one offered by the fact(s) they cover.

(iii) *Plausible hypotheses* are founded but untested hypotheses; they lack an empirical justification but are, in principle, testable.

(iv) *Corroborated hypotheses* are well-grounded and empirically confirmed; ultimately, only hypotheses of this level characterize theoretical knowledge and are the hallmark of mature science.

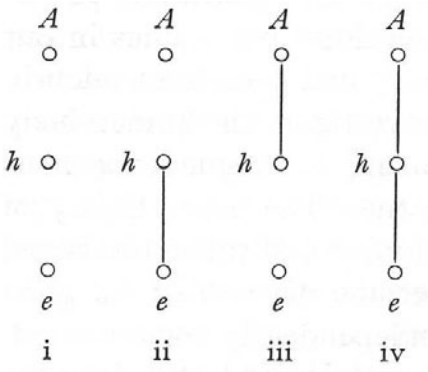


Figure 1.1: Levels of Conjecture Making and Validation

If, and only if, a corroborated hypothesis is, in addition to being well-grounded and empirically confirmed, general and systemic, then it may be termed a 'law'. Now, given that the "chief goal of scientific research is the discovery of patterns" (Bunge 1967: 305), a law is a confirmed hypothesis that is supposed to depict such a pattern.