Contributions to the Science of Text and Language

Text, Speech and Language Technology

VOLUME 31

FWF Der Wissenschaftsfonds.

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 $2_\varepsilon$, 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.

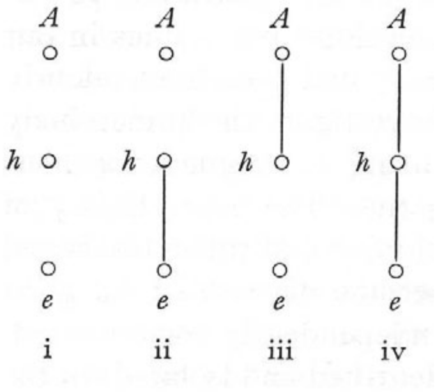

**Figure 1.1:** Levels of Conjecture Making and Validation

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.

Without a doubt, use of the term ‘law’ will arouse skepticism and refusal in linguists’ ears and hearts.[1] In a way, this is no wonder, since the term ‘law’ has a specific connotation in the linguistic tradition (cf. Kovács 1971, Collinge 1985): basically, this tradition refers to 19th century studies of sound laws, attempts to describe sound changes in the history of (a) language.

In the beginnings of this tradition, predominantly in the Neogrammarian approach to Indo-European language history, these laws – though of descriptive rather than explanative nature – allowed no exceptions to the rules, and they were indeed understood as deterministic laws. It goes without saying that up to that time, determinism in nature had hardly ever been called into question, and the formation of the concept of ‘law’ still stood in the tradition of Newtonian classical physics, even in Darwin’s time, he himself largely ignoring probability as an important category in science.

The term ‘sound law’, or ‘phonetic law’ [Lautgesetz] had been originally coined as a technical term by German linguist Franz Bopp (1791–1867) in the 1820s. Interestingly enough, his view on language included a natural-scientific perspective, understanding language as an organic physical body [organischer Naturkörper]. At this stage, the phonetic law was not considered to be a law of nature [Naturgesetz], as yet; rather, we are concerned with metaphorical comparisons, which nonetheless signify a clear tendency towards scientific exactness in linguistics. The first militant “naturalist-linguist” was August Schleicher (1821–1868). Deeply influenced by evolutionary theorists, mainly Charles Darwin and Ernšt Hackel, he understood languages to be a ‘product of nature’ in the strict sense of this word, i.e., as a ‘natural organism’ [Naturorganismus] which, according to his opinion, came into being and developed according to specific laws, as he claimed in the 1860s. Consequently, for Schleicher, the science of language must be a natural science, and its method must by and large be the same as that of the other natural sciences. Many a scholar in the second half of the 19th century would elaborate on these ideas: if linguistics belonged to the natural sciences, or at least worked with equivalent methods, then linguistic laws should be identical with the natural laws. Natural laws, however, were considered mechanistic and deterministic, and partly continue to be even today. Consequently, in the mid-1870s, scholars such as August Leskien (1840–1916), Hermann Osthoff (1847–1909), and Karl Brugmann (1849–1919) repeatedly emphasized the sound laws they studied to be exceptionless. Every scholar admitting exceptions was condemned to be addicted to subjectivism and arbitrariness. The rigor of these claims began to be heavily discussed from the 1880s on, mainly by scholars such as Berthold G.G. Delbrück (1842–1922), Mikołai Kruszewski

[1] Quite characteristically, Collinge (1985), for example, though listing some dozens of *Laws of Indo-European*, avoids the discussion of what ‘law’ actually means; for him, these “are issues better left to philosophers of language history” (ibd., 1).

(1851–87), and Hugo Schuchardt (1842–1927). Now, 'laws' first began to be distinguished from 'regularities' (the latter even being sub-divided into 'absolute' and 'relative' regularities), and they were soon reduced to analogies or uniformities [Gleichmäßigkeiten]. Finally, it was generally doubted whether the term 'law' is applicable to language; specifically, linguistic laws were refuted as natural laws, allegedly having no similarity at all with chemical or physical laws.

If irregularities were observed, linguists would attempt to find a "regulation for the irregularity", as linguist Karl A. Verner (1846–96) put it in 1876. Curiously enough, this was almost the very same year that Austrian physicist Ludwig Boltzmann (1844–1906) re-defined one of the established natural laws, the second law of thermodynamics, in terms of probability.

As will be remembered, the first law of thermodynamics implies the statement that the energy of a given system remains constant without external influence. No claim is made as to the question, which of various possible states, all having the same energy, is at stake, i.e. which of them is the most probable one. As to this point, the term 'entropy' had been introduced as a specific measure of systemic disorder, and the claim was that entropy cannot decrease in case processes taking place in closed systems. Now, Boltzmann's statistical re-definition of the concept of entropy implies the postulate that entropy is, after all, a function of a system's state. In fact, this idea may be regarded to be the foundation of statistical mechanics, as it was later called, describing thermodynamic systems by reference to the statistical behavior of their constituents.

What Boltzmann thus succeeded to do was in fact not less than deliver proof that the second law of thermodynamics is not a natural law in the deterministic understanding of the term, as was believed in his time, and is still often mistakenly believed, even today. Ultimately, the notion of 'law' thus generally was supplied with a completely different meaning: it was no longer to be understood as a deterministic law, allowing for no exceptions for individual singularities; rather, the behavior of some totality was to be described in terms of statistical probability. In fact, Boltzmann's ideas were so radically innovative and important that almost half a century later, in the 1920s, physicist Erwin Schrödinger (1922) would raise the question, whether not all natural laws might generally be statistical in nature. In fact, this question is of utmost relevance in theoretical physics, still today (or, perhaps, more than ever before). John Archibald Wheeler (1994: 293) for example, a leading researcher in the development of general relativity and quantum gravity, recently suspected, "that every law of physics, pushed to the extreme, will be found to be statistical and approximate, not mathematically perfect and precise."

However, the statistical or probabilistic re-definition of 'law' escaped attention of linguists of that time. And, generally speaking, one may say it remained unnoticed till today, which explains the aversion of linguists to the concept of law, at the end of the 19th century as well as today... Historically speaking, this aversion has been supported by the spirit of the time, when scholars like Dilthey (1883: 27) established the hermeneutic tradition in the humanities and declared singularities and individualities of socio-historical reality to be the objective of the humanities. It was the time when 'nature itself', as a research object, was opposed to 'nature *ad hominem*', when 'explanation' was increasingly juxtaposed to 'interpretation', and when "nomothetic law sciences" [nomothetische Gesetzeswissenschaften] were distinguished from "idiographic event sciences" [idiographische Ereigniswissenschaften], as Neokantian scholars such as Heinrich Windelband and Wilhelm Rickert put it in the 1890s. Ultimately, this would result in what Snow should term the distinction of *Two Cultures*, in the 1960s –a myth strategically upheld even today. This myth is well prone to perpetuating the overall skepticism as to mathematical methods in the field of the humanities. Mathematics, in this context, tends to be discarded since it allegedly neglects the individuality of the object under study. However, mathematics can never be a substitute for theory, it can only be a tool for theory construction (Bunge 1967: 467).

Ultimately, in science as well as in everyday life, any conclusion as to the question, whether observed or assumed differences, relations, or changes are essential, are merely chance or not, must involve a decision. In everyday life, this decision may remain a matter of individual choice; in science, however, it should obey conventional rules. More often than not, in the realm of the humanities, the empirical test of a given hypothesis has been replaced by the acceptance of the scientific community; this is only possible, of course, because, more often than not, we are concerned with specific hypotheses, as compared to the above Figure 1.1, i.e., with plausible hypotheses.

As soon as we are concerned with empirical tests of a hypothesis, we face the moment where statistics necessarily comes into play: after all, for more than two hundred years, chance has been statistically "tamed" and (re-)defined in terms of probability. Actually, this is the reason why mathematics in general, and particularly statistics as a special field of it, is so essential to science: ultimately, the crucial function of mathematics in science is its role in the expression of scientific models. Observing and collecting measurements, as well as hypothesizing and predicting, typically require mathematical models.

In this context, it is important to note that the formation of a theory is not identical to the simple transformation of intuitive assumptions into the language of formal logic or mathematics; not each attempt to describe (!) particular phenomena by recourse to mathematics or statistics, is building a theory, at least not in the understanding of this term as outlined above. Rather, it is important that there be a model which allows for formulating the statistical hypotheses in terms of probabilities.

At this moment, human sciences in general, and linguistics in particular, tend to bring forth a number of objections, which should be discussed here in brief (cf. Altmann 1985: 5ff.):

a. The most frequent objection is: “We are concerned not with quantities, but with qualities.” – The simple answer would be that there is a profound epistemological error behind this ‘objection’, which ultimately is of ontological nature: actually, neither qualities nor quantities are inherent in an object itself; rather they are part of the concepts with which we interpret nature, language, etc.

b. A second well-known objection says: “Not everything in nature, language, etc. can be submitted to quantification.” – Again, the answer is trivial, since it is not language, nature, etc., which is quantified, but our concepts of them.

In principle, there are therefore no obstacles to formulate statistical hypotheses concerning language in order to arrive at an explanatory model of it; the transformation into statistical meta-language does not depend so much on the object, as on the status of the concrete discipline, or the individual scholar’s education (cf. Bunge 1967: 469).

A science of language, understood in the manner outlined above, must therefore be based on statistical hypotheses and theorems, leading to a complete set of laws and/or law-like regularities, ultimately being described and/or explained by a theory. Thus, although linguistics, text scholarship, etc., in the course of their development, have developed specific approaches, measures, and methods, the application of statistical testing procedures must correspond to the following general schema (cf. Altmann 1973: 218ff.):

1. The formulation of a linguistic hypothesis, usually of qualitative kind.
2. The linguistic hypothesis must be translated into the language of statistics; qualitative concepts contained in the hypothesis must be transformed into quantitative ones, so that the statistical models can be applied to them. This may lead to a re-formulation of the hypothesis itself, which must have the form of a statistical hypotheses. Furthermore, a mathematical model must be chosen which allows the probability to be calculated with which the hypothesis may be valid with regard to the data under study.
3. Data have to be collected, prepared, evaluated, and calculated according to the model chosen. (It goes without saying that, in practice, data may stand at the beginning of research – but this should not prevent anyone from going “back” to step one within the course of scientific research.)
4. The result obtained is represented by one or more digits, by a particular function, or the like. Its statistical evaluation leads to an acceptance or refusal of the hypothesis, and to a statement as to the significance of the results.

Ultimately, this decision is not given a priori in the data, but the result of disciplinary conventions.

5. The result must be linguistically interpreted, i.e., re-translated into the linguistic (meta-)language; conclusions must be linguistically drawn, which are based on the confirmed or rejected hypothesis.

Now what does it mean, concretely, if one wants to construct a theory of language in the scientific understanding of this term? According to Altmann (1978: 5), designing a theory of language must start as follows:

> When constructing a theory of language we proceed on the basic assumption that language is a self-regulating system all of whose entities and properties are brought into line with one another in some way or other.

From this perspective, general systems theory and synergetics provide a general framework for a science of language; the statistical formulation of the theoretical model thus can be regarded to represent a meta-linguistic interface to other branches of sciences. As a consequence, language is by no means understood as a natural product in the 19th century understanding of this term; neither is it understood as something extraordinary within culture. Most reasonably, language lends itself to being seen as a specific cultural sign system. Culture, in turn, offers itself to be interpreted in the framework of an evolutionary theory of cognition, or of evolutionary cultural semiotics, respectively. Culture thus is defined as the cognitive and semiotic device for the adaption of human beings to nature. In this sense, culture is a continuation of nature on the one hand, and simultaneously a reflection of nature on the other – consequently, culture stands in an isologic relation to nature, and it can be studied as such.

Therefore culture, understood as the functional correlation of sign systems, must not be seen in ontological opposition to nature: after all, we know at least since Heisenberg’s times, that nature cannot be directly observed as a scientific object, but only by way of our culturally biased models and perspectives. Both ‘culture’ and ‘nature’ thus turn out to be two specific cultural constructs. One consequence of this view is that the definitions of ‘culture’ and ‘nature’ necessarily are subject to historical changes; another consequence is that there can only be a unique theory of ‘culture’ and ‘nature’, if one accepts the assumptions above. As Koch (1986: 161) phrases it: “ ‘Nature’ can only be understood via ‘Culture’; and ‘Culture’ can only be comprehended via ‘Nature’.”

Thus language, as one special case of cultural sign systems, is not – and definitely not *per se*, and not *a priori* – understood as an abstract system of rules or representations. Primarily, language is understood as a sign system serving as a vehicle of cognition and communication. Based on the further assumption that communicative processes are characterized by some kind of economy between the participants, language, regarded as an abstract sign system, is understood as the economic result of communicative processes.

Talking about economy of communication, or of language, any exclusive focus on the production aspect must result in deceptive illusions, since due attention has to be paid to the overall complexity of communicative processes: In any individual speech act, the producer's creativity, his or her principally unlimited freedom to produce whatever s/he wants in whatever form s/he wants, is controlled by the recipient's limited capacities to follow the producer in what s/he is trying to communicate. Any producer being interested in remaining understood (even in the most extreme forms of avantgarde poetry), consequently has to take into consideration the recipient's limitations, and s/he has to make concessions with regard to the recipient.

As a result, a communicative act involves a circular process, providing something like an economic equilibrium between producer's and recipient's interests, which by no means must be a symmetric balance. Rather, we are concerned with a permanent process of mutual adaptation, and of a specific interrelation of (partly contradictory) forces at work, leading to a specific dynamics of antagonistic interest forces in communicative processes. Communicative acts, as well as the sign system serving communication, thus represent something like a dynamic equilibrium.

In principle, this view has been delineated by G.K. Zipf as early as in the 1930s and 40s (cf. Zipf 1949). Today, Zipf is mostly known for his frequency studies, mainly on the word level; however, his ideas have been applied to many other levels of language too, and have been successfully transferred to other disciplines as well.

Most importantly, his ideas as to word length and word frequency have been integrated into a synergetic concept of language, as envisioned by Altmann (1978: 5), and as outlined by Köhler (1985) and Köhler/Altmann (1986). It would be going too far to discuss the relevant ideas in detail here; still, the basic implications of this approach should be presented in order to show that the focus on word length chosen in this book is far from accidental.

## Word Length in a Synergetic Context

Word length is, of course, only one linguistic trait of texts, among others. In this sense, word length studies cannot be but a modest *contribution to* an overall *science of language*. However, a focus on the word is not accidental, and the linguistic unit of the word itself is far from trivial.

Rather, word length is an important factor in a synergetic approach to language and text, and it is by no means an isolated linguistic phenomenon within the structure of language. Given one accepts the distinction of linguistic levels, such as (1) phoneme/grapheme, (2) syllable/morpheme, (3) word/lexeme, (4) clause, and (5) sentence, structurally speaking, the word turns out to be hierarchically located in the center of linguistic units: it is formed by lower-level units, and itself is part of the higher-level units. The question here cannot be, of course, in how far each of the units mentioned are equally adequate for linguistic models, in how far their definitions should be modified, or in how far there may be further levels, particularly with regard to specific text types (such as poems, for example, where verses and stanzas may be more suitable units).

At closer inspection (cf. Table 1.1), at least the first three levels are concerned with recurrent units. Consequently, on each of these levels, the re-occurrence of units results in particular frequencies, which may be modelled with recourse to specific frequency distribution models. To give but one example, the famous Zipf-Mandelbrot distribution has become a generally accepted model for word frequencies. Models for letter and phoneme frequencies have recently been discussed in detail. It turns out that the Zipf-Mandelbrot distribution is no adequate model, on this linguistic level (cf. Grzybek/Kelih/Altmann 2004). Yet, grapheme and phoneme frequencies seem to display a similar ranking behavior, which, in both cases depends on the relevant inventory sizes and the resulting frequencies with which the relevant units are realized in a given text (Grzybek/Kelih/Altmann 2005).

Moreover, the units of all levels are characterized by length; and again, the length of the units on one level is directly interrelated with those of the neighboring levels, and, probably, indirectly with those of all others. This is where Menzerath's law comes into play (cf. Altmann 1980, Altmann/Schwibbe 1989), and Arens's law as a special case of it (cf. Altmann 1983).

Finally, systematic dependencies cannot only be observed on the level of length; rather, each of the length categories displays regularities in its own right. Thus, particular frequency length distributions may be modelled on all levels distinguished.

Table 1.1, illustrating the basic interrelations, may be, *cum grano salis*, regarded to represent something like the synergetics of linguistics in a nutshell.

**Table 1.1:** Word Length in a Synergetic Circuit

| | | | |
|---|---|---|---|
| | SENTENCE | Length | Frequency |
| | | ↕ | |
| | CLAUSE | Length | Frequency |
| ↱ | | ↰ ↕ | |
| Frequency | WORD / LEXEME | Length | Frequency |
| ↕ ↱ | | ↰ ↕ | |
| Frequency | SYLLABLE / MORPHEME | Length | Frequency |
| ↕ ↱ | | ↰ ↕ | |
| Frequency | PHONEME / GRAPHEME | Length | Frequency |

Much progress has been made in recent years, regarding all the issues mentioned above; and many questions have been answered. Yet, many a problem still begs a solution; in fact, even many a question remains to be asked, at least in a systematic way. Thus, the descriptive apparatus has been excellently developed by structuralist linguistics; yet, structuralism has never made the decisive next step, and has never asked the crucial question as to explanatory models. Also, the methodological apparatus for hypothesis testing has been elaborated, along with the formation of a great amount of valuable hypotheses.

Still, much work remains to be done. From one perspective, this work may be regarded as some kind of "refinement" of existing insight, as some kind of detail analysis of boundary conditions, etc. From another perspective, this work will throw us back to the very basics of empirical study. Last but not least, the quality of scientific research depends on the quality of the questions asked, and any modification of the question, or of the basic definitions, will lead to different results.

As long as we do not know, for example, what a word is, i.e., how to define a word, we must test the consequences of different definitions: do we obtain identical, or similar, or different results, when defining a word as a graphemic, an orthographic, a phonetic, phonological, a morphological, a syntactic, a psychological, or other kind of unit? And how, or in how far, do the results change – and if so, do they systematically change? – depending on the decision, in which units a word is measured: in the number of letters, or graphemes, or of sounds, phones, phonemes, of morphs, morphemes, of syllables, or other units? These questions have never been systematically studied, and it is a problem *sui generis*, to ask for regularities (such as frequency distributions) on each of the levels mentioned. But ultimately, these questions concern only the first degree of uncertainty, involving the qualitative decision as to the measuring units: given, we clearly distinguish these factors, and study them systematically, the next questions concern the quality of our data material: will the results be the same, and how, or in how far, will they (systematically?) change, depending on the decision as to whether we submit individual texts, text segments, text mixtures, whole corpora, or dictionary material to our analyses? At this point, the important distinction of types and tokens comes into play, and again the question must be, how, or in how far, the results depend upon a decision as to this point.

Thus far, only language-intrinsic factors have been named, which possibly influence word length; and this enumeration is not even complete; other factors as the phoneme inventory size, the position in the sentence, the existence of suprasegmentals, etc., may come into play, as well. And, finally, word length does of course not only depend on language-intrinsic factors, according to the synergetic schema represented in Table 1.1. There is also abundant evidence that external factors may strongly influence word length, and word length frequency distributions, factors such as authorship, text type, or the linguo-historical period when the text was produced.

More questions than answers, it seems. And this may well be the case. Asking a question is a linguistic process; asking a scientific question, is a also linguistic process, – and a scientific process at the same time. The crucial point, thus, is that if one wants to arrive at a science of language, one must ask questions in such a way that they can be answered in the language of science.

## References

Altmann, Gabriel
1973 "Mathematische Linguistik." In: W.A. Koch (ed.), *Perspektiven der Linguistik.* Stuttgart. (208–232).

Altmann, Gabriel
1978 "Towards a theory of language." In: *Glottometrika 1.* Bochum. (1–25).

Altmann, Gabriel
1980 "Prolegomena to Menzerath's Law." In: *Glottometrika 2*. Bochum. (1–10).

Altmann, Gabriel
1983 "H. Arens' ≫Verborgene Ordnung≪ und das Menzerathsche Gesetz." In: M. Faust; R. Harweg; W. Lehfeldt; G. Wienold (eds.), *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik.* Tübingen. (31–39).

Altmann, Gabriel
1985 "Sprachtheorie und mathematische Modelle." In: *SAIS Arbeitsberichte aus dem Seminar für Allgemeine und Indogermanische Sprachwissenschaft 8.* Kiel. (1–13).

Altmann, Gabriel; Schwibbe, Michael H.
1989 *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen. Mit Beiträgen von Werner Kaumanns, Reinhard Köhler und Joachim Wilde.* Hildesheim etc.

Bunge, Mario
1967 *Scientific Research I. The Search for Systems.* Berlin etc.

Collinge, Neville E.
1985 *The Laws of Indo-European*. Amsterdam/Philadelphia.

Dilthey, Wilhelm
1883 *Versuch einer Grundlegung für das Studium der Gesellschaft und Geschichte.* Stuttgart, 1973.

Grzybek, Peter; Kelih, Emmerich; Altmann, Gabriel
2004 Graphemhäufigkeiten (Am Beispiel des Russischen) Teil II: Theoretische Modelle. In: *Anzeiger für Slavische Philologie, 32*; 25–54.

Grzybek, Peter; Kelih, Emmerich; Altmann, Gabriel
2005 "Häufigkeiten von Buchstaben / Graphemen / Phonemen: Konvergenzen des Rangierungsverhaltens." In: *Glottometrics, 9*; 62–73.

Koch, Walter A.
1986 *Evolutionary Cultural Semiotics.* Bochum.

Köhler, Reinhard
1985 *Linguistische Synergetik. Struktur und Dynamik der Lexik.* Bochum.

Köhler, Reinhard; Altmann, Gabriel
1986 "Synergetische Aspekte der Linguistik", in: *Zeitschrift für Sprachwissenschaft, 5*; 253–265.

Kovács, Ferenc
1971 *Linguistic Structures and Linguistic Laws.* Budapest.

Rickert, Heinrich
1899 *Kulturwissenschaft und Naturwissenschaft.* Stuttgart, 1986.

Schrödinger, Erwin
1922 "Was ist ein Naturgesetz?" In: Ibd., *Was ist ein Naturgesetz? Beiträge zum naturwissenschaftlichen Weltbild.* München/Wien, 1962. (9–17).

Smith, Neilson Y.
1989 *The Twitter Machine*. Oxford.

Snow, Charles P.
1964 *The Two Cultures: And a Second Look.* Cambridge, 1969.

Wheeler, John Archibald
1994 *At Home in the Universe.* Woodbury, NY.

Windelband, Wilhelm
1894 *Geschichte und Naturwissenschaft.* Strassburg.

Zipf, George K.
1935 *The Psycho-Biology of Language: An Introduction to Dynamic Philology.* Cambridge, Mass., $^{2}$1965.

Zipf, George K.
1949 *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology.* Cambridge, Mass.

# HISTORY AND METHODOLOGY OF WORD LENGTH STUDIES

## *The State of the Art*

Peter Grzybek

## 1. Historical Roots

The study of word length has an almost 150-year long history: it was on August 18, 1851, when Augustus de Morgan, the well-known English mathematician and logician (1806–1871), in a letter to a friend of his, brought forth the idea of studying word length as an indicator of individual style, and as a possible factor in determining authorship. Specifically, de Morgan concentrated on the number of letters per word and suspected that the average length of words in different Epistles by St. Paul might shed some light on the question of authorship; generalizing his ideas, he assumed that the average word lengths in two texts, written by one and the same author, though on different subjects, should be more similar to each other than in two texts written by two different individuals on one and the same subject (cf. Lord 1958).

Some decades later, Thomas Corwin Mendenhall (1841–1924), an American physicist and metereologist, provided the first empirical evidence in favor of de Morgan's assumptions. In two subsequent studies, Mendenhall (1887, 1901) elaborated on de Morgan's ideas, suggesting that in addition to analyses "based simply on mean word-length" (1887: 239), one should attempt to graphically exhibit the peculiarities of style in composition: in order to arrive at such graphics, Mendenhall counted the frequency with which words of a given length occur in 1000-word samples from different authors, among them Francis Bacon, Charles Dickens, William M. Thackerey, and John Stuart Mill. Mendenhall's (1887: 241) ultimate aim was the description of the "normal curve of the writer", as he called it:

> [...] it is proposed to analyze a composition by forming what may be called a 'word spectrum' or 'characteristic curve', which shall be a graphic representation of the arrangement of words according to their length and to the relative frequency of their occurrence.

*P. Grzybek (ed.), Contributions to the Science of Text and Language: Word Length Studies and Related Issues,* 15–90.

Figure 2.1, taken from Mendenhall (1887: 237), illustrates, by way of an example, Mendenhall's achievements, showing the result of two 1000-word samples from Dickens' *Oliver Twist*: quite convincingly, the two curves converge to an astonishing degree.

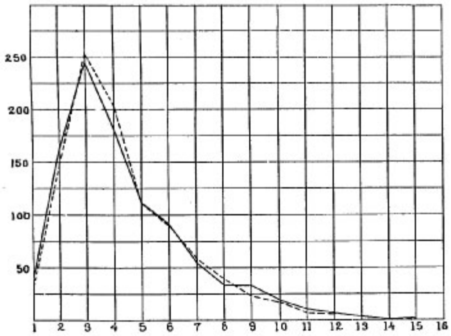

**Figure 2.1:** Word Length Frequencies in Dickens' *Oliver Twist* (Mendenhall 1887)

Mendenhall (1887: 244) clearly saw the possibility of further applications of his approach:

> It is hardly necessary to say that the method is not necessarily confined to the analysis of a composition by means of its mean word-length: it may equally well be applied to the study of syllables, of words in sentences, and in various other ways.

Still, Mendenhall concentrated solely on word length, as he did in his follow-up study of 1901, when he continued his earlier line of research, extending it also to include selected passages from French, German, Italian, Latin, and Spanish texts.

As compared to the mere study of mean length, Mendenhall's work meant an enormous step forward in the study of word length, since we know that a given mean may be achieved on the basis of quite different frequency distributions. In fact, what Mendenhall basically did, was what would nowadays rather be called a frequency analysis, or frequency distribution analysis. It should be mentioned, therefore, that the mathematics of the comparison of frequency distributions was very little understood in Mendenhall's time. He personally was mainly attracted to the frequency distribution technique by its resemblance to spectroscopic analysis.

Figure 2.2, taken from Mendenhall (1901: 104) illustrates the curves from two passages by Bacon and Shakespeare. Quite characteristically, Mendenhall's conclusion was a suggestion to the reader: "The reader is at liberty to draw any conclusions he pleases from this diagram."

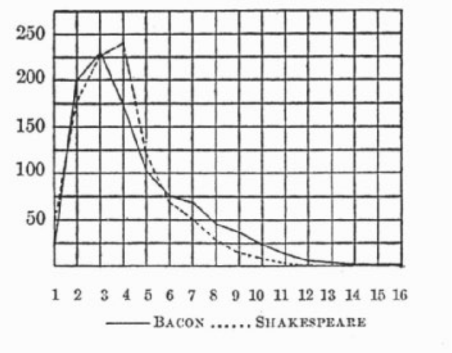

**Figure 2.2:** Word Length Frequencies in Bacon's and Shakespeare's Texts (Mendenhall 1901)

On the one hand, one may attribute this statement to the author's 'scientific caution', as Williams (1967: 89) put it, discussing Mendenhall's work. On the other hand, the desire for calculation of error or significance becomes obvious, techniques not yet well developed in Mendenhall's time.

Finally, there is another methodological flaw in Mendenhall's work, which has been pointed out by Williams (1976). Particularly as to the question of authorship, Williams (1976: 208) emphasized that before discussing the possible significance of the Shakespeare–Bacon and the Shakespeare–Marlowe controversies, it is important to ask whether any differences, other than authorship, were involved in the calculations. In fact, Williams correctly noted that the texts written by Shakespeare and Marlowe (which Mendenhall found to be very similar) were primarily written in blank verse, while all Bacon's works were in prose (and were clearly different). By way of additionally analyzing works by Sir Philip Sidney (1554–1586), a poet of the Elizabethan Age, Williams (1976: 211) arrived at an important conclusion:

> There is no doubt, as far as the criterion of word-length distribution is concerned, that Sidney's prose more closely resembles prose of Bacon than it does his own verse, and that Sidney's verse more closely resembles the verse plays of Shakespeare than it does his own prose. On the other hand, the pattern of difference between Shakespeare's verse and Bacon's prose is almost exactly comparable with the difference between Sidney's prose and his own verse.

Williams, too, did not submit his observations to statistical testing; yet, he made one point very clear: word length need not, or not only, or perhaps not even primarily, be characteristic of an individual author's style; rather word length, and word length frequencies, may be dependent on a number of other factors, genre being one of them (cf. Grzybek et al. 2005, Kelih et al. 2005).

Coming back to Mendenhall, his approach should thus, from a contemporary point of view, be submitted to cautious criticism in various aspects:

(a) *Word length is defined by the number of letters per word.–* Still today, many contemporary approaches (mainly in the domain of computer sciences), measure word length in the number of letters per word, not paying due attention to the arbitrariness of writing systems. Thus, the least one would expect would be to count the number of sounds, or phonemes, per word; as a matter of fact, it would seem much more reasonable to measure word length in more immediate constituents of the word, such as syllables, or morphemes. Yet, even today, there are no reliable systematic studies on the influence of the measuring unit chosen, nor on possible interrelations between them (and if they exist, they are likely to be extremely language-specific).

(b) *The frequency distribution of word length is studied on the basis of arbitrarily chosen samples of 1000 words.–* This procedure, too, is often applied, still today. More often than not, the reason for this procedure is based on the statistical assumption that, from a well-defined sample, one can, with an equally well-defined degree of probability, make reliable inferences about some totality, usually termed population. Yet, as has been repeatedly shown, studies along this line do not pay attention to a text's homogeneity (and consequently, to data homogeneity). Now, for some linguistic questions, samples of 1000 words may be homogeneous – for example, this seems to be the case with letter frequencies (cf. Grzybek/Kelih/Altmann 2004). For other questions, particularly those concerning word length, this does not seem to be the case – here, any selection of text segments, as well as any combination of different texts, turns out to be a "quasi text" destroying the internal rules of textual self-regulation. The very same, of course, has to be said about corpus analyses, since a corpus, from this point of view, is nothing but a quasi text.

(c) *Analyses and interpretations are made on a merely graphical basis.–* As has been said above, the most important drawback of this method is the lack of objectivity: no procedure is provided to compare two frequency distributions, be it the comparison of two empirical distributions, or the comparison of an empirical distribution to a theoretical one.

(d) *Similarities (homogeneities) and differences (heterogeneities) are unidimensionally interpreted.–* In the case of intralingual studies, word length frequency distributions are interpreted in terms of authorship, and in the case of interlingual comparisons in terms of language-specific factors, only; the possible influence of further influencing factors thus is not taken into consideration.

However, much of this criticism must then be directed towards contemporary research, too. Therefore, Mendenhall should be credited for having established an empirical basis for word length research, and for having initiated a line of research which continues to be relevant still today. Particularly the last point mentioned above, leads to the next period in the history of word length studies. As can be seen, no attempt was made by Mendenhall to find a formal (mathematical) model, which might be able to describe (or rather, theoretically model) the frequency distribution. As a consequence, no objective comparison between empirical and theoretical distributions has been possible.

In this respect, the work of a number of researchers whose work has only recently and, in fact, only partially been appreciated adequately, is of utmost importance. These scholars have proposed particular frequency distribution models, on the one hand, and they have developed methods to test the goodness of the results obtained. Initially, most scholars have (implicitly or explicitly) shared the assumption that there might be one overall model which is able to represent a general theory of word length; more recently, ideas have been developed assuming that there might rather be some kind of general organizational principle, on the basis of which various specific models may be derived.

The present treatment concentrates on the rise and development of such models. It goes without saying that without empirical data, such a discussion would be as useless as the development of theoretical models. Consequently, the following presentation, in addition to discussing relevant theoretical models, will also try to present the results of empirical research. Studies of merely empirical orientation, without any attempt to arrive at some generalization, will not be mentioned, however – this deliberate concentration on theory may be an important explanation as to why some quite important studies of empirical orientation will be absent from the following discussion.

The first models were discussed as early as in the late 1940s. Research then concentrated on two models: the Poisson distribution, and the geometric distribution, on the other. Later, from the mid-1950s onwards, in particular the Poisson distribution was submitted to a number of modifications and generalizations, and this shall be discussed in detail below. The first model to be discussed at some length, here, is the geometric distribution which was suggested to be an adequate model by Elderton in 1949.

## 2. The Geometric Distribution (Elderton 1949)

In his article "A Few Statistics on the Length of English Words" (1949), English statistician Sir William P. Elderton (1877–1962), who had published a book on *Frequency-Curves and Correlation* some decades before (London 1906), studied the frequency of word lengths in passages from English writers, among them Gray, Macaulay, Shakespeare, and others.

As opposed to Mendenhall, Elderton measured word length in the number of syllables, not letters, per word. Furthermore, in addition to merely counting the frequencies of the individual word length classes, and representing them in