Quantitative Linguistics, an Invitation

Karl-Heinz Best

Otto Rottmann

2017

RAM-Verlag

Studies in Quantitative Linguistics

Editors

Fengxiang Fan	(fanfengxiang@yahoo.com)
Emmerich Kelih	(emmerich.kelih@univie.ac.at)
Reinhard Köhler	(koehler@uni-trier.de)
Ján Mačutek	(jmacutek@yahoo.com)
Eric S. Wheeler	(wheeler@ericwheeler.ca)

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen*. 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro***.** 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in quantitative linguistics 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3.* 2013, IV + 403 pp.
14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics Vol. 4.* 2014, VIII+148 pp.
15. Best, K.-H., Kelih, E. (eds.), *Entlehnungen und Fremdwörter: Quantitative Aspekte.* 2014. VI + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, G. *Unified Modeling of Length in Language.* 2014, VIII + 123 pp.
17. G. Altmann, R. Čech, J. Mačutek, L. Uhlířová (eds.), *Empirical Approaches to Text and Language Analysis dedicated to Luděk Hřebíček on the occasion of his 80th birthday*. 2014. VI + 231 pp.
18. M. Kubát, V. Matlach, R. Čech, *QUITA Quantitative Index Text Analyzer*. 2014, VII + 106 pp.

19. K.-H. Best, *Studien zur Geschichte der Quantitativen Linguistik.* 2015. III + 158 pp.
20. P. Zörnig, K. Stachowski, I.-I. Popescu, T. Mosavi Miangah, P. Mohanty, E. Kelih, R. Chen, G. Altmann, *Descriptiveness, Activity and Nominality in Formalized Text Sequences.* 2015. IV+120 pp.
21. G. Altmann, *Problems in Quantitative Linguistics Vol. 5*. 2015. III+146 pp.
22. P. Zörnig, K. Stachowski, I.-I. Popescu, T. Mosavi Miangah, R. Chen, G. Altmann, *Positional Occurrences in Texts: Weighted Consensus Strings.* 2015. II+178 pp.
23. E. Kelih, R. Knight, J. Mačutek, A.Wilson (eds.), *Issues in Quantitative Linguistics Vol. 4.* 2016. III + 231 pp.
24. J. Léon, S. Loiseau (eds.), *History of quantitative linguistics in France*. 2016. II + 232 pp.

ISBN: 978-3-942303-51-4

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
Germany
RAM-Verlag@t-online.de
http://ram-verlag.eu

Preface

The present book is a strongly extended translation of the German original „Quantitative Linguistik. Eine Annäherung", 3rd edition. (Göttingen: Peust & Gutschmidt Verlag 2006). It is written for beginners in a rather "non-mathematical" language and contains many computations showing the procedures which should be followed by a researcher who begins to work with a language. It aims at showing some law-candidates concerning length, diversification, evolution, borrowings, ranking, mutual relations, first steps in synergetics, etc. A hypothesis can become law only if it is derived from a theory and corroborated on as many texts and languages as possible. An introduction to the history of quantitative linguistics shows that there is a very old tradition beginning about 2500 years ago.

Writing an introduction one never knows whether the contents and the way of presentation will attract or discourage the young readers. On the other hand, there are a small number of introductions which can show the reader some backgrounds of quantitative linguistics. Usually, the authors fulfill the books by mathematics, deriving of models, statistical tests, computing some probabilities etc. and the beginner does not know what to do. The present book shows how to define, how to count, how to fit a function or a distribution, how to evaluate it and how to present it in a publication. The rich bibliography attached to the book helps the reader to find ready evaluations, descriptions of the problem, the way of attaching the results to an existing theory, etc. The authors refer to many works written by their students and published in form of articles or dissertations.

The main motive of the book is: Begin and do not cease! Mathematics is no misfortune, on the contrary, it is a means to exactly express our findings and attach them to other ones. Even simple quantitative data collections are better than speaking about something in a non-formal language without any trials for corroboration.

The editorial board of this series recommends both to beginners and to advanced scholars to imitate the results using as many languages as possible in order to corroborate the existing results.

Gabriel Altmann

Contents

Preliminary Remarks

The following explications are addressed to those readers who are often "missed" by authors and researchers whose specialty is Quantitative Linguistics (QL), but represent a large potential circle of readers and staff members: philologists (linguists, but also literary scholars and historians), who have never thought of seriously dealing with phenomena of linguostatistics, partly, since they did not allocate any relevance to such subjects, partly, since they assume that the methods required can only be learned elaborately and with uncertain results. In contrast to requirements in the tuition of physicians, psychologists, sociologists and economists the subjects of the philosophical faculties mostly do without the recommendation of acquiring knowledge in statistics or even the integration of the subject into the relevant courses of studies. Thus, a hurdle is developed which is an extensive obstacle for most representatives as to access to the possibilities of QL. The "success" of that situation is that thematically versatile international research with an orientation to language statistics exists, which, however, is almost exclusively taken note of by specialists, not by the majority of philologists, though they could benefit from it. On the other hand, there exist evident needs with many linguists and literary scholars concerning precise statements related to their observations which find expression in statistics for most different objects, e.g. if attempts are made to describe stylistic particularities of a text, text type / genre or author.

Now, it has to be shown that and how it is possible to acquire part of the relevant research and even do own research work without long-time studies of statistics. That is made possible by the large improvement of the tools of the disciplines mentioned, especially if an interested scholar is ready to participate in cooperations. This means that it has to be attempted to remove some of the hermetic character from the quantitative studies in linguistics and literature. To achieve that the following subjects will be dealt with:

- development of quantitative linguistics (and studies in literature);
- subjects which can be processed quantitatively thus furnishing scientific progress;
- statistical observations concerning German vocabulary;
- detection of linguistic laws;
- theory;
- perspectives.

The above subjects are mainly described from the perspective of quantitative linguistics (QL) in the German speaking countries; despite that restriction the object can only be dealt with in excerpts and by means of examples. The main subject is the validation of law-like hypotheses. "Quantitative Literaturwissenschaft"[1] – a

[1] Quantitative studies in literature

term used by the German physicist Wilhelm Fucks (1968: 77, 88) – will play a marginal role only; mentioning it has a more programmatic character. However, it has to be stated that many questions dealt with in quantitative linguistics can be applied to literary texts and also to other genres of art without problems. "Quantitative literary science can teach us to understand better what actually takes action in the author or is done by him when writing his works by the mapping of issues in texts onto mathematical models. Generally, however, he is hardly aware of what he does formally."[2] (Fucks 1968: 88). This corresponds to all authors, not only the literary ones. A lot of information on what quantitative studies in literature can do is found in Altmann (1988a) and Altmann & Altmann (2005), Popescu et al. (2015).

Though explications are mainly addressed to non-specialized readers, and this is done in the hope that some interest can be aroused in them, specialists should profit from reading this book as well; they may confidently skip the trivial descriptions of elementary information. Dealing with linguistic laws concentrates on those whose testing the authors of this book were or still are somehow involved in the course of the projects performed in Germany, Austria and China. Many results appear there for the first time; with respect to subjects already published improved files have been elaborated or new calculations have been performed in some cases already. To obtain a depiction that is as demonstrative as possible almost all files are presented in the form of tables or graphical illustrations. It is not originality that is the primary target of these explications, but an overview that is as compact as possible; understandability is mainly intended. As contributions to quantitative linguistics can hardly be found in more widely circulating professional publications in linguistics to allow the further induction into subjects mentioned here, their theoretical backgrounds and the mathematical aspects in a lot of references are given to be used, if required.

Due to the restrictions concerning subjects it is not claimed that this book can be an introduction into quantitative linguistics being sufficient for all purposes. With respect to that important subjects like linguistic typology, stylistics and readability research come off badly. But all beginners can use it directly and extend the store of languages scrutinized up to now. Any analysis performed with the presented means would either corroborate a hypothesis or force us to revise it.

We'd like to recommend that the readers use the book *Quantitative Linguistics. An International Handbook*, edited by Köhler, Altmann and Piotrowski (2005), containing many domains of linguistics.

The present book is a strongly revised version of the third edition of K.-H. Best, *Quantitative Linguistik. Eine Annäherung*. Göttingen: Peust & Gutschmidt Verlag, 2006.

[2] „Eine quantitative Literaturwissenschaft kann uns durch die Abbildung von Sachverhalten in Texten auf mathematische Modelle besser verstehen lehren, was eigentlich im Autor vorgeht oder von ihm getan wird, wenn er seine Werke verfaßt. Dabei wird der Autor sich dessen, was er formal tut, im allgemeinen kaum bewußt sein."

1. Development of Quantitative Linguistics (QL)

Oksaar (1972: 630, footnote 2) states as follows: "The beginnings of QL go back to the antiquity (lists of the Hapax legomena)"[3]. Colometry and stichometry are supposed to begin in the 3rd century BC (Pawłowski 2008); combinatorial reflections on the formation of linguistic units begin with Xenokrates in the 4th century BC (Biggs 1979: 113f., Best 2005b: 79f.). In the 7th century Masoretes in Palestine compiled "an exact **frequency statistic** concerning words" (Merten 1983: 35) whose use was intended for the codification of the Hebrew Old Testament. About 1466 Leon Battista Alberti (1404-1472) issued his paper on encryption and decryption *De componendis cifris* (Gadol 1969: 207), which comprises linguistic analyses based on statistics concerning the ratio of vowels and consonants, the relative frequency of vowels and consonants among themselves, etc. (Gadol 1969: 210). Bauer (1995: 213) therefore assigns knowledge of "statistical linguistic, especially …frequencies"[4] to Alberti.

The beginnings of shorthand, which contributed "to the compilation of frequency dictionaries in a fertilizing way", are dated to the 16th century by Wolff (1969: 211) who expressly refers to Bright (1588) who "compiles a list of the most frequent English words to which he allocates symbols; the shortest shorthand symbols are allocated to the most frequent words."

In his *Dissertatio de arte combinatoria* (1666; 1962: 61; Eco 1997: 278ff.) published in the 17th century Leibniz reflects about the number of terms which can be formed if the available alphabet comprises 24 letters and the terms formed do not have more than 24 letters. The result is a number of more than 620 sextillion terms. For comparison: The German alphabet comprises 30 letters incl. <ß> with the umlauts with upper and lower cases not being considered yet. So, the number of combinable terms is increased accordingly. Leibniz's special attainment is appreciated by Schmidt (1966: 52) as the "disjointed transition from the qualitative to the quantitative combination." (As to the linguistic understanding in the Enlightenment see also Ricken 1981: 548; for the development of combinatorics till the ideas of Leibniz see: Eco 1997; Gardt 1994: 206-226.)

Finally, in the 18th century disputes concerning the orthodoxy of the Pietists were decided on the basis of a quantitative analysis of contents by counting the religious key terms (Merten 1983: 35).

It was in the middle of the 19th century at the latest that the continuous development of quantitative linguistics began (for this, see the time table in Meier [2]1967: 349-351). For example, Lord (1958: 282) quotes from a letter by De Morgan dated 18-8-1851 in which he proposes to solve the problem of the authorship of the Paul's Epistles by means of the criterion of word length.

[3] "Die Anfänge der quantitativen Linguistik reichen bis ins Altertum zurück (Listen der Hapaxlegomena)"

[4] "statistische Gesetzmäßigkeiten der Sprache, insbesondere ... Häufigkeiten"

L.A. Sherman (1888) dealt with observations on record length in English prose. A quantitative analysis of subjects in New York papers was also done in the 19th century (Merten 1983: 36). According to Kempgen (1995: 13) Russian linguistics began with sound and grapheme statistics at the end of the 19th century; Grzybek (2003:103f.), however, refers to programmatic reflections on possibilities of language statistics published by Bunjakovskij in 1847 already.

In German linguistics reflections on combinations by Harsdörffer (Best 2005a) and Leibniz (Best 2005b) are at the beginning of the development, before Jean Paul (Best 2006d) performed statistical analyses of linguistic phenomena. Kaeding (1897: 38) refers to Gabelsberger's efforts concerning shorthand from 1822; amongst others, Förstemann's analyses on sound statistics in the German linguistic development phases (i.e. Gothic, Old High German, Middle High German, Modern High German) and in the classical languages (Förstemann 1846, 1852; Best 2006b) as well as Drobisch's hexameter studies (1866ff) in the same languages follow. Statistical efforts concerning the periodization of Plato's works (Cherubim & Hilgendorf 1998: 235; Dittenberger 1881: 326) must be mentioned (for further information see the Guiraud's thematically and chronologically sorted bibliography 1954.)

Pott – asked for an overview of results in linguistics for the 1st volume of Techmer's magazine – takes up Förstemann's work as well as Whitney's relevant lecture (Pott 1884: 24f.) and emphasizes: "The **statistical** conduct of classes of sounds and individual sounds is of particular relevance as to the overall effect on the ear and otherwise...". [5]

Further to that, Pott's reflections (1884: 19; Best 2006) are based on Leibniz's combinatorial considerations concerning the formation of phrases and integrated knowledge of word lengths in some languages (Greek, Old High German, Algonkian and others). However, he points out that – due to the "incompatibility of several sounds side by side and also the numerous repetition of the same sound" (Pott 1884: 19) - the actual vocabulary of languages is much smaller than the one that can be determined by way of calculation (for this, see: Beutelsbacher 1997: 26f.).

The masterminds in quantitative linguistics in Germany also include Gabelentz, who – in a posthumously published paper (Gabelentz 1894) – outlines the program of a quantitative language typology appearing to be very modern and says that: "a dozen of known properties of a language should allow to be surely indicative of hundreds of further properties; the typical tendencies are obvious" [6] (Gabelentz 1894: 7). Best known, however, should be Kaeding's (1897/98) Frequency Dictionary of the German Language, which was continued by Meier (²1967) later.

Thumb who moved for the use of experiments and statistics in philological disciplines in his paper on methodology and demonstrated the usefulness of

[5] „Von besonderer Wichtigkeit betreffs der Gesamtwirkung eines Sprachidioms auf das Ohr und sonst ist aber das **statistische** Verhalten der Lautklassen und Einzellaute..."

[6] "aus einem Dutzend bekannter Eigenschaften einer Sprache müsste man mit Sicherheit auf hundert andere Züge schliessen können; die typischen Züge, die herrschenden Tendenzen lägen klar vor Augen"

statistics by taking the example of the position of verbs in the Greek language (Thumb 1911: 2; Best & Kotrasch 2005; Kotrasch 2004/05) was almost forgotten. Even before, he had performed association experiments to study analogy in cooperation with Marbe (Thumb & Marbe 1901). Marbe in turn dealt with quantitative analyses of the rhythm of German prose (Marbe 1904; for this, see Best 2001e, 2005c, 2006f and in this volume (p. 56). In a very differentiated way he comments on contemporary assumptions concerning the causes of linguistic changes, especially the thought of economy; he warns against the reduction of linguistic changes to just a few conditions (Marbe 1916: 63ff.).

In the 1930s Zwirner & Zwirner developed a school of phonological research based on the paragon of biometry that they called "Phonometry" accordingly. Their findings included the law that later became known as "Frumkina's law" (Zwirner & Zwirner 1935; Best 2005g). Further to that, they used the example of the distribution of the duration of short vowels to prove their thesis "that the implementation of linguistic standards bound to activities by organs must show legalities of distribution which can be captured by means of the binomial formula or specific derivations of the binomial law" (Zwirner & Zwirner 1936: 192). Their paper "Remarks on the history of phonetics ", which is more a history of linguistics in wide sections, can in parts be understood as an archaeology of the statistic background of some linguistic concepts like the affinity of languages and sound laws (Zwirner & Zwirner 1966: 17ff.).

Apart from Meier (1967) five authors must be mentioned most of all with respect to the further development of QL in Germany up to 1980: Thumb's disciple, the later Bonn phonetician Menzerath, whose studies aimed at the objective of a "Comparative Linguistic Typology" (Menzerath 1954, Preface) and detected the connection between construct and constituent which was later called the "Menzerath-Altmann Law" (Best 2006l); the Aachen physicist Fucks (1968) who analyzed statistical structures in arts, literature and music and founded the laws of distribution of differently long units (Aichele 2005); two scientists in Saarbrücken: the specialist in German studies Eggers (project "Syntax of Contemporary German"; Eggers 1962, 1973) and the anglicist Finkenstaedt (project "Computerized Thesaurus of English"; Finkenstaedt & Wolff 1973; Wolff 1969) as well as the Kiel linguist Winter, who was in charge of a project on "Quantitative Stylistics" (Pieper 1979). That enumeration could be extended.

Internationally, the situation of the QL during that period can be characterized by the fact that in some countries individual scientists or small groups of scientists were dealing with specific subjects. Important authors to be named include: Zipf and Greenberg (USA), Herdan (UK), Guiraud, Benzecri, Petitot, Thom and Muller (France), Brainerd (Canada), Arapov, Cherc, Frumkina and Piotrowski (Russia), Tuldava (Estonia), Orlov (Georgia), Mańczak and Sambor (Poland), Mistrík (Slovakia), Hřebíček and Uhlířová (Czech Republic), Levickij and Perebyjnos (Ukraine). This list includes some of the better known scientists only; for references see Köhler (1995). Several papers in Köhler, Altmann & Piotrowski [Eds.] (2005) are dedicated to the history of Quantitative Linguistics in individual countries.

Among the authors mentioned above it was especially Zipf who attracted a lot of attention in linguistics. The so-called Zipf Laws are common knowledge in linguistics: If words of a language are sorted according to their frequency, the result is a ranking of those words. If, then, the frequency rank *r* of the words is multiplied by the relevant frequency *f*, the result is an almost constant value *C*: $r \cdot f = C$ (Zipf 1949: 24, with data from *Ulysses*). Crystal (1993: 87) demonstrates the applicability and limit of that idea by means of examples from an English corpus of spoken language. A further correlation: the longer the words, the more infrequent their use (Crystal 1993: 87; for this, cf. p. 23 ff.). Zipf's basic idea was that language is mainly shaped by striving after economic behavior; hereby, the needs of speaker and hearer as against their language must be understood as antithetical. Therefore, language requires the best possible balance between them (Zipf 1949: 19ff.), which must always be reset in case of linguistic alterations ("steady state"). Beyond the area of linguistics Zipf's laws have been adopted in minimum ten scientific disciplines and are more recently quoted in chaos research as well.

Greenberg's idea of a quantitative linguistic typology by defining linguistic characteristics in the form of indices met a big response as well (1960). One of them is the synthetism index:

(1.1) $$S = \frac{Number\ of\ \ words}{Number\ of\ \ morphemes}$$

This index is used to calculate the morphological average length of words (Greenberg 1960: 187, in the form modified by Krupa 1965: 33f.). All in all, ten such indices were developed. Altmann & Lehfeldt (1973) worked out the correlations among those indices and, based on that, a "Graph of the correlations of properties" (Altmann & Lehfeldt 1973: 42-54). This line of research could be understood as a continuation of the ideas of Gabelentz (1894); but it has remained almost unknown for a long time. At the same time, it is a part of synergetic linguistics introduced by Köhler (1986).

A new development has been introduced since the seventies: In 1978 the publication of the book series *Quantitative Linguistics* began. On the one hand it contains monographs and articles as original papers (in the anthologies under the title of *Glottometrika*), on the other the series is a contribution to international communication by translations. In 1991, the *First International Conference on Quantitative Linguistics (QUALICO)* took place in Trier. In 1994, the *International Quantitative Linguistics Association (IQLA)* was founded on the occasion of the 2nd *QUALICO* (Moscow); in the same year, the publication of the *Journal of Quantitative Linguistics* began and soon it was considered the voice of the IQLA. Further congresses took place in Moscow 1994, Helsinki in 1997, in Prague in 2000, in Athens (USA) in 2003, in Trier in 2007, in Graz in 2009, in Belgrade in 2012, in Olomouc in 2014. Today, they are organized regularly. Since 1998 *Göttinger Beiträge zur Sprachwissenschaft* have been published with regular contributions to the QL; in 2001 *Glottometrics* and *ETC (Empirical Text and Culture Research/ Empirische Text- und Kulturforschung)* followed. Today,

there are further journals publishing quantitative linguistics works, e.g. *Mathematical Linguistics* in Japan and Iran, *Glottotheory* in Germany.

An essential contribution to the consolidation of the QL is Köhler's *Bibliography of Quantitative Linguistics* which was published in 1995, that bibliography that does not only consider the relevant literature, but allows its manifold analysis by means of several indices. A further stage is marked by the international handbook *Quantitative Linguistik – Quantitative Linguistics* (ed. by Köhler, Altmann & Piotrowski) published in 2005 and permitting topical and historical overviews. Further, biographical articles are dedicated to the *History of Quantitative Linguistics* in *Glottometrics* beginning with edition no. *6/2003 ff*. It has also proved helpful that Trier University established a professorship of linguistic data processing with one of the focal points being *Quantitative Linguistik* with its own right to award doctorates.

The present situation is characterized by increasing cooperation and organization. Internet communication helps bridge vast distances that still exist at least partly. *IQLA* was the first organization addressing relevantly interested scientists all over the world. In addition to the research groups above further groups were formed in Finland, Greece, India, Japan, China, Canada, France, the Netherlands, Austria, Iran, Italy and the Czech Republic. From the point of view of German linguistics the scientists to be named most of all are G. Altmann (Bochum), L. Hoffmann (Leipzig) and R. Köhler (Trier); Altmann and Köhler have a large share in the organizational and scientific progress in that discipline; Hoffmann was the source of decisive impulses in the area of languages for special purposes (Hoffmann & Piotrowski 1979; Hoffmann ^{2}1985).

Finally, we'd like to point towards some long-term projects in the German speaking area: From 1978 a focal point of Altmann's works is the idea of language being a self-regulating system (Altmann 1978) as well as the derivation of law-based hypotheses (Altmann 1980); an encyclopaedia of linguistic laws is being prepared (http://lql.uni-trier.de). For almost thirty years Goebl (2004, 2005) has continued to develop *Dialectometry* further, concentrated on France and Northern Italy. Since the middle of the 1980s R. Köhler has worked out the *Linguistic Synergetics* (Köhler 1986; Köhler & Altmann 1986; Köhler 2005). The self-organization of language is also a central concern of Fenk-Oczlon (1997, 2001). Following some analyses in the seventies and eighties of the last century the *Göttingen Project of Quantitative Linguistics* (K.-H. Best; http://www.user.gwdg.de/~kbest) has continuously been working and strives towards an improved empirical foundation of linguistic laws. The latest development which should be mentioned is the *Graz Project of Quantitative Text Analysis (QuanTA)* (http://www-gewi.uni-graz.at/quanta/projekt_descr.htm) founded by Grzybek and Stadlober in 2002, which - amongst others in the Russian, Slovenian and Croatian languages – systematically researches the factors influencing frequency distributions in texts (cf. Grzybek, ed., 2006).

In general, after the publication of *Quantitative Linguistics. An International Handbook*, edited by Köhler, Altmann and Piotrowski (2005) one can observe a trend towards theory formation. Even if there is a lot of testing some hypotheses using data from texts in many languages in order to corroborate them,

the researchers strive for the last step, namely modeling the phenomena, unifying the approaches and searching for links between properties in the control circuit. One uses differential and difference equations, stochastic processes, tries to interpret the parameters of functions linguistically.

There are two specialized book series, namely *Quantitative Linguistics* published by the de Gruyter Mouton publishing house and *Studies in Quantitative Linguistics* published by the RAM-Verlag. As part of the *Studies* started a book series *History of Quantitative Linguistics* (Best 2015, Léon & Loiseau 2016). The number of international QL projects elaborated by teams of researchers increases yearly.

Last but not least, a thorough bibliography of newer publications is elaborated continuously at Trier-University in Germany.

2. QL Topics

The development of QL shows that it dealt with a number of topics that according to our opinion can roughly be divided into five areas.

1. QL has always concentrated on the solution of specific tasks. Crystal (1993: 86) e.g. refers to Samuel Morse (1791-1872), the inventor of the Morse code, who used a frequency hierarchy of letters, as he had found them as types in a printing shop, when he developed his code. It is no coincidence that the grapheme <e> is represented by just one dot \·\[7], <t> by a dash \-\, however, <y> by a complex combination of dash-dot-dash-dash. Statistical linguistic instruments also supported the development of shorthand and the keyboard of a typewriter (Meier ²1967: 329ff.). Bauer (1995: 213ff.) and Beutelsbacher (1997: 25ff.) point out to the significance of statistical analyses for encryption and decryption.
2. Studies used for the solution of problems in neighboring disciplines are a further topic; it includes research concerning legibility and the comprehensibility of text for research in psycholinguistics (Groeben 1982: 148ff.; Ballod 2001; Best 2006g) and speech-language clinic (Bamberger & Vanecek 1984; Mikk 2000) or the problem of the identification of anonymous authors for literary studies (Fucks 1968: 95ff.; Wickmann 1981).
3. A third field of study can be seen in the fact that linguostatistical elicitations help with the discrimination of linguistic findings or hypotheses. So, Hentschel (1992) confronts the general hypothesis of high-frequent words bucking linguistic changes with observations made in Slavic languages which add up to the hypothesis that high-frequent words can even support innovation in the early phase of a linguistic change, retarding effects can be observed in the late phase. "The quantitative data of the last five phenomena definitely point out that frequent words can also be pioneers for innovations" (Hentschel 1992: 55). So, the new nominative plural with the ending -{a} instead of -{i} is prevalent especially with more frequent nouns (Hentschel 1992: 53). At this point, Labov's variables rule (1976) should be reminded of.
4. A further field of study is stylistics. Comprehensive quantitative studies were mainly performed in the following fields: research in technical languages (Hoffmann 1985) and literature (Fucks 1968; Kreuzer & Gunzenhäuser 41971), on the distinction of oral and written language (Höhne-Leska 1975), the study of poetic language (Popescu, Lupea, Tatar, Altmann 2015) and the research in language used in the press (Lindell & Piirainen 1980; Becker 1995) and in text types (Pieper 1979). This field of study also comprises the problem of characterization and differentiation of style in various text groups (Fucks 1955a; Oksaar 1972; Dshurjuk & Levickij 2003).

A new approach concerning the discussion referring to the "nominal style" in the German language was presented by Ziegler, Best & Altmann (2002); that presentation concerns an objective method of the determination of "nominal style" and shows that the stylistic characteristics mentioned are not new at all, but can also be evidenced in Early Modern High German text types.

[7] In this publication "\...\" is used to mark Morse characters.

However, the four fields of study mentioned so far do not have linguistic laws in the center of interest; they are – if taken into consideration at all – marginal or background phenomena. Those linguistic laws are dealt with in the fifth field of QL studies. The only research scientist coming from QL, but generally known in the overall linguistics is involved in this field: George Kingsley Zipf (1902-1950); he has become public property due to the "laws" named after him (Crystal 1993: 87; Prün 2005). Those laws also include the finding that words being used more frequently are shorter than those used more rarely (Zipf 1932: 59ff.). As a matter of fact, he was the precursor of linguistic synergetics.

Quite a number of further linguistic laws exist. We have also referred to early proposals by Zwirner and Zwirner in the thirties of the past century; many other linguistic laws were found by other researchers. You can find a number of examples in this publication. The systematization of linguistic laws offers the chance of developing a linguistic theory. Prün (1999) showed that Zipf's hypotheses could be grouped to form a consistent systematic linguistic conception that can be understood to be an early form of linguistic synergetics (Köhler 1986; Hoffmann & Krott 2002). Only recently suggestions have come up that linguistic laws that have been understood very differently so far can obviously be ascribed to one and the same basic mechanism (Wimmer & Altmann 2005, 2006).

3. Statistical Observations Concerning Vocabulary

3.1 How many words does the German language comprise?

Care for the survivability of one's own language or even one's curiosity may cause us to ask how many words our mother tongue has, probably with the tongue in cheek that its shape is the better the more comprehensive the vocabulary is.

The tenor of the subsequent remarks can be found with Eisenberg (1998: 201): "The voluminous inventory of words that is available in a language like German is not known exactly by an individual speaker. Nobody knows how many words the German language comprises at a certain time and nobody knows exactly which meanings a single word has." However, this does not prevent Eisenberg (1998: 33) from offering figures related to several aspects of the vocabulary. Of course, there are some possibilities for reasonable questions and searching for preliminary answers.

So, how many words does the German language include?

Combinatorial reflections concerning the formation of terms by means of an alphabet comprising 24 letters, which Harsdörffer, Leibniz and - later - Pott undertook, were already referred to. Even Pott characterized the calculated dimension to be unrealistic (cf. p. 4).

To obtain at least an approximate impression of the actual vocabulary of a language you can consult the respectively most comprehensive dictionaries or other information. The German language mainly offers two works: on the one hand the *Deutsches Wörterbuch* by Jacob and Wilhelm Grimm (1852 - 1960), which comprises the German vocabulary from the middle of the 15th century up to the middle of 20th century (Gardt 1999: 261); Schlaefer (1999: 10) indicates that its volume is "approx. 350.000 keywords" and remarks that larger parts of the vocabulary like foreign words, legal terminology and dialect words are missing, which resulted in the development of special dictionaries. They also include *Das Deutsche Rechtswörterbuch*, which according to its own statement (vol. 9, 1992-96: III) is set out to approximately 130.000 words.

To test an own method that can be pitted against Schlaefer's data, Ina Kühner (personal information) performed a further estimation. The basic thought of her test was that the dictionary had been compiled by several generations of researchers, which means that different processing methods can be expected. Therefore, a separate estimation was done for each of the 32 individual volumes. Evaluations included each 100th column; in most cases the columns were numbers 2, 102, 202, ... till the end. As the volumes comprised a different number of words, between 14 and 30 columns per volume were counted. This counting gave evidence of considerable differences concerning the number of keywords per volume resulting from different amounts of words per volume, but also individual keywords. The keyword "gewinnen" e.g. covers more than 71 pages (with two columns each), "Gewinn" further 30 pages. The estimation results were lowest in volume 18 with 862.8 keywords, and highest in volume 13 with

22,856.25 keywords. The addition of the projected values resulted in the estimation value of 267,691.67 keywords for the complete works. Attention is attracted by the fact that the number of estimated keywords in the older volumes (vol. 1-5, 10-16) is substantially higher than in the later ones. Maybe, this is an explanation for the different indications by different authors concerning the amount of keywords in Grimms' dictionary (Best 2000a: 36). The difference between Schlaefer's indications and Kühner's estimation can be explained by the fact that the individual word articles comprise further keywords that were not evaluated by Kühner.

The most voluminous present dictionary is *Duden.* According to its own particulars "*Das Große Wörterbuch der deutschen Sprache* (^{3}1999) comprises "more than 200,000 keywords with more than 300,000 explanations of meanings" (preface) from the period of the middle of the 18th century to the present in the third ten-volume edition. This dictionary comprises derivations and compounds; it does without the individually used vocabulary and ad-hoc formations, but includes words needed to understand the classical German literature as from Lessing.

Both works also include obsolescent vocabulary; today, it must be deducted in thought when answering the question of the volume of the German vocabulary without knowing in detail how many keywords are respectively concerned. On the other hand it is clear that even the most voluminous dictionary cannot include the complete vocabulary, since new words are permanently developed, either by loan or by word formation. *Meyers Enzyklopädisches Lexikon* in 25 vol. (1985) with approx. 250,000 keywords has almost the same volume like the *Duden*, however, it has to be taken into consideration that proper names play an especially major part in encyclopediae (for further details concerning the volumes of German dictionaries: Haß-Zumkehr 2001: 183; however, details concerning Grimms' dictionary are excessive).

But there are also words that are even missing in those most comprehensive encyclopediae and dictionaries, although they do not come under the restrictions mentioned and are widespread. In the time after WWII a four-wheel vehicle for children was available which was moved forward and backward by a lever similarly to a manually actuated draisine and which was known under the designation of "Holländer" (a children's vehicle actuated by muscle force); that designation is still known today by many people. This word cannot be found in this meaning in *Meyers Enzyklopädisches Lexikon* for example. The "Ballonroller" (= another little children's vehicle: scooter with so-called "balloon tyres") cannot be found in any of the consulted encyclopediae and dictionaries, not even in the *Brockhaus Enzyklopädie*. In an answer to a letter to the editor of *Der Sprachdienst* (44/ ed. 3-4, 2000, 125f.) reference is made to the missing of the word "Pornokratie" in recent dictionaries. In addition, regularly formed derivations and compounds are missing (Bär 2001: 170). It could be supposed that they are no individual cases, which means that portions of vocabulary must be assumed to be widely known, but they are not even included in the most voluminous encyclopediae. Even if interjections or words in comics or slang are not thought of, you can agree to the well-known author of horror fiction, Stephen

King (2000: 131), who said: "And then there are words which cannot be found in any dictionary, but are part of the vocabulary nevertheless."

If you want to have an idea of the vocabulary in a language, it is also important to know that the generally used vocabulary, the "central vocabulary", comprises approx. 70,000 words (*Duden. Deutsches Universalwörterbuch* ^{4}2001: 13). As to that volume Bergenholtz (1989: 773) agrees by stating "more than 60,000 lemmata"; this is deemed to be the vocabulary "the user [of a dictionary, author] will most likely meet again and again" (ibidem). The core area of the so-called basic vocabulary comprises the 1,000 to 1,200 most frequent words that – depending on the type of text – reach text coverage of 70 to more than 90 % (Krohn 1992: 166ff.). A comprehensive vocabulary has to be expected in technical and special languages; only a minor portion of that vocabulary can be expected in general dictionaries, which means only if a word is not only used in the communication area of the technical language. *Duden. Das Große Wörterbuch der deutschen Sprache* (Bd. 1/ ²1993: 7) mentions: electrical engineering: "approx. 60,000 technical terms, medicine: "approx. 250,000 technical terms and constructions", organic chemistry: "more than approx. 3.5 million denotations". Acc. to Winter (1986: 155) a "minimum of 20 million mass nouns" have to be expected in chemistry. The relatively new vocabulary in computer technology is already estimated to amount to 25,000 to 26,000 words (Wichter 1998: 1177; On the development of computer vocabulary: Busch 2000: 218; Best 2006j). As to linguistic terminology the widespread *Lexikon der Sprachwissenschaft* (Bußmann 1990: 7; approx. 3500 entries) and *Metzler Lexikon Sprache* (Glück [ed.] 1993; approx. 5000 articles) can be deemed relatively representative, even if completeness could not be achieved.

Let's summarize: a number of more than 200,000 words can be considered the lower limit of general vocabulary; when adding vocabulary of technical languages, it is evident that the number of entries would easily sum up to several millions of entries. The relevant literature mentions 300,000 to 500,000 entries for the German language (Menzerath 1954: 5; Miller 1993: 160). König (^{15}2005: 115) concludes: "The German language comprises approx. 4,000 basic morphemes. Those elements form the vocabulary. The orthography of the ›Duden‹ has approx. 110,000 entries, the most voluminous dictionaries include approx. 2 - 300,000 ones having up to half a million meanings. If you include technical languages and names as well, millions of words are taken into consideration. However, those estimations differ strongly."

Just for comparison some information on other languages: English is estimated to have 600,000 to 800,000 words, the estimation for the French language is 100,000 (*Meyers grosses Taschenlexikon*, 1987, Bd. 24: 200); Wolff (1969: 48): "More recent estimations indicate the number of 500,000 to 600,000 words for the English vocabulary; the German vocabulary is just below that number and the French vocabulary comprises about 300,000 entries."

Apart from those differences: Is French therefore an especially poor language? This distinctly shows a problem linked to such indications: If you consider that in French "Kartoffel" (= potato) is called "pomme de terre" and "Kartoffelbrei" (= mashed potatoes) is "purée de pommes de terre", it can be seen that

the designation is just governed by different structural methods of designation than it is the case in German. “Speech recognition and (la) reconnaissance de la parole“ have the same meaning. In other cases, there are differences which can be noticed when the volume of a dictionary is determined, only orthographic conventions are different: “Saint-Esprit“ vs. “Holy Ghost”, “triple saut“ vs. “triple jump”, “deux cents“ vs. “two hundred”. One and the same target of communication can be realized in different ways. A very extensive research has been performed for French (cf. Léon, Loiseau 2016).

In this context the following question is of interest as well: how many words must a language have as a minimum? Of course, the answer depends on communicative necessities. If they are rather restricted, this means that a vocabulary of approx. 500 lexemes may seem to be sufficient: Pidgin-English in New Guinea is said to have had that approximate amount in the 19th century, the same amount as the formerly planned “Colonial German“ (Stiberc 1999: 123, 128). There are – controversial – indications (Menzerath 1954: 5 f.) that the active vocabulary of some groups of population has almost the same extent.

There are several reasons for the uncertainties occurring when vocabulary is estimated: 1. The vocabulary of a language continuously alters by the loss of words on the one hand as well as loans and new formations on the other. Loans in the German language are statistically surveyed by Best (2001b), Körner (2004) and Ternes (2011) by the complete evaluation of dictionaries; as to the course of loans and growth of vocabulary also see the section “Incompletely implemented linguistic changes“ (p. 118 ff.). 2. It must be determined what has to be considered a “word“. Just to hint at some problems: it is always a problem to determine whether two units are considered a single word or two different ones. In case of homonyms like the German “Bank“ (bench or bank) it is easy to agree upon two words. In case of words usually seen as a polyseme – like “Pferd“ (animal or gymnastic apparatus) such a decision is not that easy. As a lot of words are plurivalent in that sense, the estimation of vocabulary strongly depends on how the decision is come to. How similar must the meanings of a word be to allow the interpretation of one and the same word? If just a few differences in meaning make you speak of different words, the volume of the vocabulary is increased (for further aspects of the problem see Störig 1997: 206f.). If it is not lexemes, i.e. entries of keywords in the vocabulary that are taken as referential elements but forms of words, the vocabulary will increase even more. Abbreviations, proper names, technical terms will also make it increase. Further to that it should be reflected whether or not certain syntagmata (functional verbal structures, in German: “Streckformen“) could be entered as words. Not even the problem of compounds is solved: what kind of compounds can be considered a word? Those written together, separated by a hyphen, written separately, joined by a preposition of conjunction, etc. And for some languages like Chinese or Japanese not even this is enough. This enumeration of possible differences shows that definitions are conventions which may be adapted to the aim of the topical examination.

3.2. The individual's vocabulary

3.2.1. Active and passive vocabulary

The uncertainty occurring when determining the vocabulary of a language continues when it is attempted to indicate individual persons' vocabularies. The question concerning active vocabulary is complex, but it is almost approximately solvable. If the active vocabulary is defined to be the word inventory implemented and documented at least once, the relevant information can be prepared for authors whose works were edited. The entire works by Pushkin are said to comprise 544,777 word occurrences (number of individual words in text = tokens) with 21,197 different words (types) (Hoffmann 1985: 148). Acc. to Zipf (1949: 23) *Ulysses* comprises 260,430 tokens and 29,899 types. In 36 dramas by Shakespeare there are 885,535 tokens und 21,150 types (Montemurro & Zanette 2001: 2). A multitude of details concerning the vocabulary in works at world literature level can be found with Orlov (1982b: 146ff.). Acc. to him *Finnegans Wake* by Joyce comprises 218,077 tokens with 63,924 types, Tolstoj's *Krieg und Frieden* contains approx. 472,000 tokens with just 2,146 types. Similar details concerning German texts and text mixes can be found with Billmeier (1969: 35): Kant's *Kritik der reinen Vernunft* has a length of text of 184,098 tokens based on 9,335 types. Th. Storm's complete works are said to comprise 22,500 different words (Braun 1998: 159); Benn's lyrics includes 50,000 references of "more than 10,000 individual words" (Lyon & Inglis 1971: preface); according to the estimation by Ina Kühner (personal message) the *Große Konkordanz zur Luther-Bibel* comprises approx. 19000 entries.

Goethe's vocabulary is deemed especially voluminous (Störig 1997: 197). Its extent is determined in the "Information for the user" in vol. III the Goethe dictionary (1998) as follows: "Goethe's passed-on vocabulary comprises approx. 90,000 words. The documentary evidence proliferated considerably by the excerption of newly edited texts ... ."

To check such statements you can evaluate an author's work index. Taking the *Wörterbuch zu HEINRICH VON KLEIST* (1989) as edited by Helmut Schanze, you have a lexicon of 482 pages in which the vocabulary used by Kleist in his narrations, anecdotes and short writings is listed in the form of lemmata (entries in the lexicon) and – allocated to each lemma – the different forms of words. As the author does not give any information on the scope of the dictionary, an estimation procedure offers itself: the lemmata and forms of words are counted, beginning e.g with page 20, then for each 48th page; following that, the values are added and the sum is converted to 482 pages. The estimation results in 10,507.6 lemmata and 18,364.2 forms of words. That is the number of words Kleist had – as a minimum – in his active vocabulary. Undoubtedly, more lemmata and forms of words have to be expected from his works (dramas, poems, …) not taken into consideration. As Kleist left behind an oeuvre that – comparably – is not very comprehensive, this estimation supports the information which can be found for other authors concerning the amount of lexical

entries and for which the vocabulary of 20,000 to 30,000 words used is suggested.

This, however, does not include the active vocabulary; so far, just the documented lemmata are concerned. However, once in his life any author will have used words, which have not been recorded anywhere and must therefore be missing in the complete concordance.

Beyond that, the estimation of the so-called passive vocabulary is really difficult: How many words in a language are really mastered by someone? It can be assumed that everybody knows words he had never had or wanted to use, which, however, he could use at any time, e.g. well-known taboo words, several technical terms, a number of obsolete words etc. (for this, also see the above examples "Ballonroller" and "Holländer").

Miller (1993: 159) describes an estimation process to deduce a person's vocabulary: "For that purpose we use the following calculation. Let's assume the dictionary we start with comprises 500,000 words. If 500 of them are chosen by chance to estimate the extent of a friend's mental lexicon, the sample factor is 1,000. That means that each word the friend recognizes results in a credited knowledge of 1,000 words that could have been selected, which, however, was not done. If the friend recognizes 100 of the 500 words, the estimated volume of the vocabulary is 100 x 1000, i.e. 100,000 words. If the analysis was, however, started with a dictionary of just 100,000 words, the friend should have recognized each test word to achieve the same estimated extent. So, the basic rule is: the more comprehensive the dictionary is on which the test is based, the higher the estimations being considered probable." The test process contains at least one further problem: the probability that a word is known in the sample of a test person changes with the extent of the underlying dictionary.

So, how many words does an individual master, actively and also passively? Information on this varies, and it does not always become clear what is meant exactly. It also has a distinct social component. The number of words of some classes of population in the country in England is estimated to be below 300 or about 400 words (Pott 1884: 21; Menzerath 1954: 5f.). *Meyer's Enzyklopädisches Lexikon* (vol. 25: article "Vocabulary") quantifies the active vocabulary of a German average speaker to be 12,000 to 16,000 words, including about 3,500 foreign words. Acc. to Miller (1993: 164) the vocabulary of a high-school graduate comprises approx. 60,000 different words if "proper names, numerals, foreign words, acronyms and many nondecomposable word compositions" are included in the counts.

Similar proportions should be found in German as well: "In his daily life a plain citizen needs some thousand words only. A more educated person, e.g. a scholar or an author, can use several ten thousand words (active vocabulary) and understand a lot more when they encounter them (passive vocabulary)" (Störig 1997: 207). Wellmann (1998: 408) quantifies the vocabulary used after attending the secondary grammar school to comprise 20,000 to 30,000 words. In a different publication the vocabulary of an averagely educated person is estimated to comprise 50,000 words, "the contrary extreme that a farm hand only needs some hundreds of words must certainly be corrected" (*Der grosse Brockhaus*, vol. 12:

598). Even higher estimations come "up to 250,000 words" (Rothweiler 2001: 21).

A different estimation method was developed by Wagner, Altmann & Köhler (1987: 132ff.) to determine the overall vocabulary of children. They made use of findings concerning the growth of vocabulary in texts. If e.g. a text or continuously spoken language is decomposed into equally long text blocks of e.g. 100 tokens each, it can be analyzed how many new words (types) occur in each text block. The occurrence of new words tends to decrease from the first to the final text block. Concerning this decrease various models have been developed, among them the formula which has been proven in linguistics again and again:

(3.1) $y = ax^b$

(Wagner, Altmann & Köhler 1987: 136). This formula has turned out to be a good model for the growth in vocabulary in a 430-minute long recording by the 12.2-year-old Christiane (Wagner, Altmann & Köhler 1987: 135; means from two corpora). Wagner, Altmann & Köhler (1987) analyzed the growth in new types in 57 consecutive text blocks of 100 token each; below, two of those text blocks were summed up to one (cf. Table 1.1).

Table 1.1
Increase in new types per 200 tokens with Christiane (age: 12.2)

Text block (per 200 tokens)	New types (observed)	New types (computed)	Text blocks (per 200 tokens)	New types (observed)	New types (computed)	Text blocks (per 200 tokens)	New types (observed)	New types (computed)
1	44.00	43.60	11	16.50	16.32	21	10.00	12.52
2	35.25	32.82	12	16.25	15.75	22	13.25	12.29
3	19.00	27.80	13	8.25	15.24	23	10.00	12.06
4	25.75	24.71	14	17.75	14.79	24	10.50	11.86
5	26.00	22.55	15	14.50	14.37	25	9.25	11.66
6	23.50	20.92	16	16.50	14.00	26	12.25	11.47
7	18.50	19.64	17	12.00	13.66	27	11.00	11.30
8	21.25	18.60	18	16.50	13.34	28	10.00	11.13
9	15.75	17.72	19	15.50	13.05	29	12.50	10.97
10	19.00	16.97	20	13.25	12.78			
a = 43.6000 b = -0.4098 D = 0.87								

a and *b* are the parameters; *D* is the coefficient of determination that - with *D* = 0.87 – signals a good conformity between the observations and the theoretical model that is also shown by the following diagram (Figure 1.1):

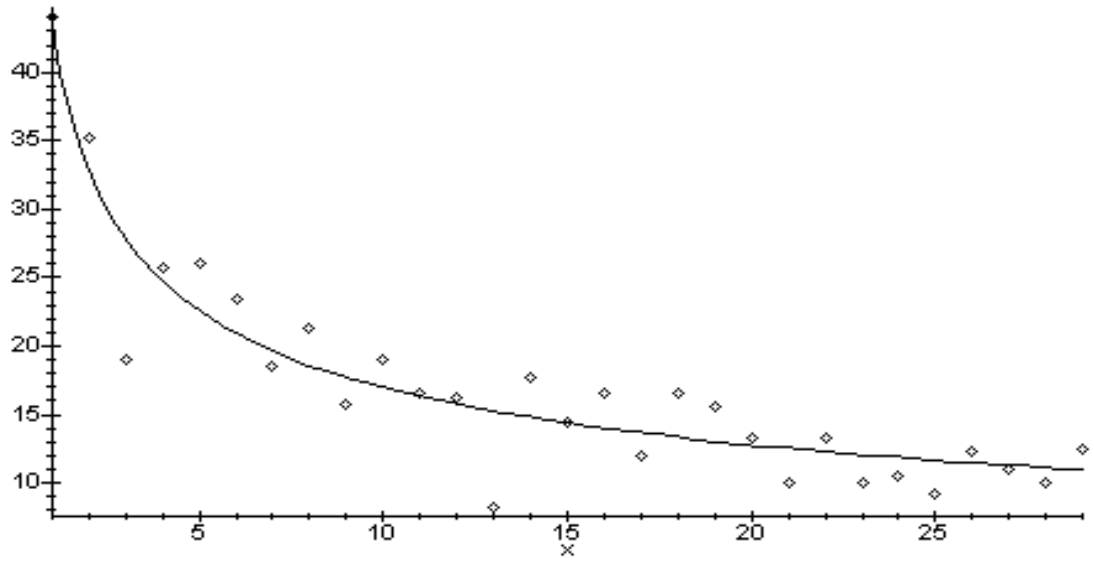


Figure 1.1. New types in text blocks
(*x*-axis: text blocks no. 1 to no. 29, from the beginning of text; *y*-axis: number of the new types = words in text section *x*).

The general trend that the increase in new words decreases when a text becomes longer is unambiguous. That trend becomes even clearer during observations on the spoken language by a group of 7-year-old Russian children (Tuldava 1995: 135), cf. Figure 1.2.

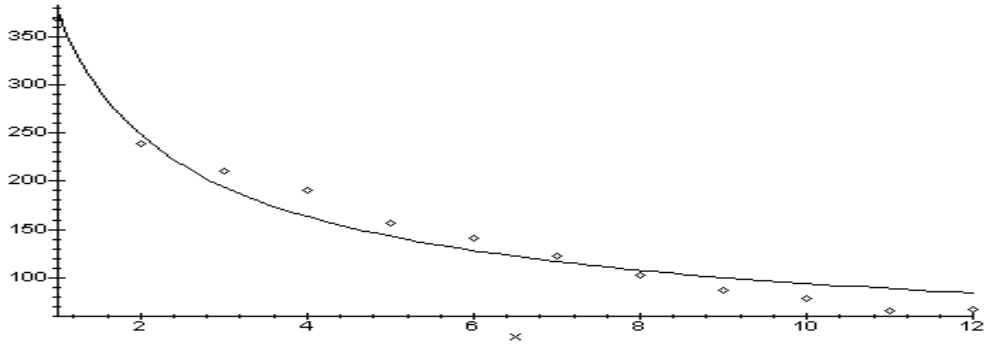


Figure 1.2. New words in spoken language
(*x*-axis: text blocks no. 1 to no. 12 with 1,000 tokens each when text length increases; *y*-axis: number of the new types = lexemes in text section *x*).

In that case the model applied explains 96.80% of the variance observed. It can be assumed that in the case of Christiane similarly good data could be achieved if more samples were available.

If the trend followed by the growth of vocabulary has been defined once, then it is possible to calculate beyond the linguistic section examined from which

text block less than a new word is introduced. On this basis it can be determined how many types (new words) a child would use till then. That value can be considered a temporary estimation of a child's present total vocabulary. Wagner, Altmann & Köhler (1987: 137) present the overview (simplified) displayed in Table 2.1.

Table 2.1
Age and vocabulary

Child	Age	Total vocabulary (estimated)
Katrin	1.5	3,717
Andreas	2.1	12,754
Gabi	5.4	3,800
Frederik	8.7	83,953
Regina	10.7	16,653
Christiane	12.2	47,321

(on different methods of estimation see Wimmer & Altmann 1999). The evaluation of an all-day recording (not estimated, but counted) of the 9.7 year-old girl Teresa resulted in 28,142 tokens and 3,825 types (Wagner 1974: 280). Wagner (1974: 301) estimates the passive vocabulary twice as much as the active one; Rothweiler (2001: 21) refers to assumptions concerning the language of adults that consider the difference to be lower: up to 90 % of the overall vocabulary is deemed to be actively usable.

As far as the numerical data are estimations, it is probably reasonable to understand them preliminarily such that they concern the approximate dimension of word estimations. But it seems to be still candid how reliable the different methods are. Differences among individuals should be large, even within educational or social classes. And: What does it mean when someone claims to know or even master a word? Is it really sure that an actively used word is always fully mastered? Do speakers or writers never use a word that they only partially know?

Here, a general remark must be inserted: What is the theoretical background of the study of an individual's vocabulary? It is evident with children but it is not evident with adults; it is evident with modern languages, but not evident with e.g. Sumerian; it is evident with individual texts or the complete creation of an author: it shows the vocabulary richness of individual texts; but it is not evident with individuals speaking the given language. It would be necessary to at least specify the parts of vocabulary, not simply compute them, knowing that the results are hard to interpret.

3.2.2. On the use of words

The use of language and words has already been thematized during the reflections of individual persons' active vocabulary. The following remarks deal with a special aspect of the use of words: which words – or even better – which classes

of words are preferred, which are used more rarely? This question can be approached under very different aspects, since words have different properties. This direction is more realistic and allows to approach theory. Two of the possibilities are to be addressed in short, before one of them is dealt with in greater detail.

The first aspect is the correlation between the length of words and their frequency. When looking into frequency dictionaries (Rosengren 1972, Wängler 1963: 59ff.) or at the frequency pyramids of the words (Braun 1998: 160f.; König [15]2005: 114; Meier 1967: 52) the following is found: short words are used most frequently; the more sounds or syllables a word comprises, the more rarely it is used. This fact can be seen most impressively in Zipf's overview concerning Kaeding's findings: among almost 11 million words the one-syllable ones comprise almost 50 %, whereas a 15-syllable word occurs only once (Zipf 1935/ 1968: 23). The word is used by the military and it is not documented, as Kaeding (1897 / 98) entered only such words that occurred at least four times.

It should be noted that measuring word length in terms of phonemes or letters does not bring relevant results – because the immediate constituents of the word are syllable and morpheme – and cannot be used e.g. for Chinese or Japanese.

A further connection is that increasing word length causes the decrease of the number of meanings (polysemy) of those words on average (Altmann & Schwibbe 1989: 66ff.). Word length is in the center of the complex of cooperating linguistic properties and units, as shown by Köhler (1986: 74) for the lexicon of the German language. That correlation can be presented as "Köhler's control circuit", cf. Figure 3.1 (Köhler & Altmann 1986: 261).

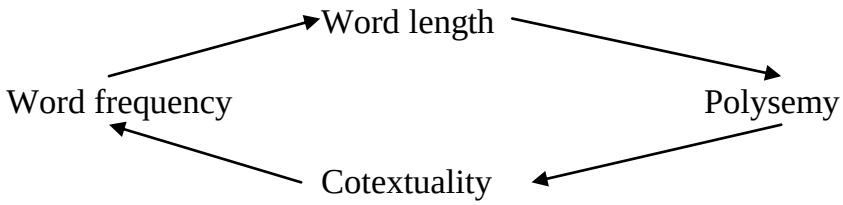


Figure 3.1. Part of control circuit

Cotextuality is the ability of appearing in different texts. The arrows can be read as "x has an effect on y". Those cooperating properties of words belong to the center of "linguistic synergetics" (Köhler 1986, 2005).

In the following paragraphs this aspect of word length is to be analyzed.

How long can German words be? According to the comprehensive analyses of word length distributions (Best [ed.] 1997) words comprising up to 20 syllables occur in German texts outside technical texts without having a parodistic meaning or exclusively intended to attract attention: "Rinderkennzeichnungs- und Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz RkReÜAÜG M V." Acc. to *DER SPIEGEL* (42/ 1999: 17) that is the "Title of a law discussed in the regional parliament of Mecklenburg-Vorpommern and committed to the board responsible for agriculture"; that title of a law attracted so much attention that it was quoted one day long in broadcasts

by the *North German Broadcasting Service (NDR)* again and again and was also listed among the words of the year 1999 (Bär 2000: 3; 17f.). "Mini-Funk-Passiv-Infrarot-Bewegungsmelder" (in an ad for Scharpf alarm systems, *DER SPIEGEL* 9/ 1992: 73), "Metalloxidhalbleiter-Feldeffekttransistor" (ad by Technics, *DER SPIEGEL* 49/ 1992: 232), "Maschinengewehrmunitionsmagazinen" (G. Kunert, *Erwachsenenspiele*, München: dtv 1999, 284), "Verkehrswegeplanungsbeschleunigungsgesetz" (*Naturschutz heute*, vol. 25/ no. 2/ 1993: 23), "Hochleistungsultrakurzwellengeradeausempfänger" (Moser 1971: 100) and the name of the institution "Sparkassen-Innovations-Beteiligungsfinanzierungsgesellschaft m.b.H." (*DER SPIEGEL* 35/ 1998: 75) are further extreme cases. What would Mark Twain have said when getting to know such expressions, the man who already sneered at the word "Generalstaatsverordnetenversammlungen", which only has twelve syllables, and even shorter compounds (Twain 1961: 449)? The example mentioned first distinctly exceeds the longest word found by Kaeding. In the technical language used in chemistry even more different dimensions of word formation are reached. In his sketch "At the pharmacy" Karl Valentin (1996: 42) sneered at this technical language by the wrongly reproduced "Isopropilprophe[n]ilbarbitursauresphenildimethildim[e]thylaminophirazolon" [8] (Boettcher et al. 1983: 130, corrected), a term whose "just" 30 syllables have by no means reached the peak of word complexity. In the handbook *Rote Liste* used by many physicians you can find the chemical shortcut "Ceftriaxon" for a Cefalosporin, an antibiotic, whose full denomination is "(6R,7R)-7-[2-(2-Amino-4-thiazolyl)-glyoxylamido]-3-(2,5-dihydro-6-hydroxy-2-methyl-5-oxo-1,2,4-triazin-3-yl-thiomethyl)-8-oxo-5-thia-1-azabicyclo[4.2.0]oct-2-en-2-carbonsäure-7²-(Z)-(O-methyloxim)" (*Rote Liste* 1989; List of chemical shortcuts of medical agents, 19).

It need not be enhanced especially that such words are an exception outside technical texts. It can also be contended whether such words belong to the German vocabulary, as they are restricted to a specific community of users. That argument cannot be applied to the IT vocabulary whose words can also be considerably complex (Grote & Schütte 2000: 100f.).

Is it really worth to deal with the length of words, an aspect of language that could spontaneously and easily be considered especially trivial and irrelevant? Best (1997a) contains several indications concerning the relevance of word length. Here are some additional ones:

In the linguistic system as well as in language use word length is a central structuring property, a fact which can be deduced – amongst others – from Köhler's control circuit (1986: 74). It depends on the size of the phoneme inventory of a language, the frequency of words in texts (Strauss, Grzybek & Altmann 2006), the length of the clauses to which words belong, and the volume of the lexicon of a language. It influences the length of constituents of a word (morphs, syllables), its polysemy and the duration of its sounds (Best 2006g). Longer words are rather subjected to folk-etymological conversions than shorter ones

[8] Thanks a lot to Dante Bernabei, Darmstadt, for subject-specific suggestions.

(Paul [4]1909: 221f.). In addition, Tuldava (1995: 21) pointed out that there is a distinct correlation between the age of words on the one hand and their length, frequency and polysemy (see page 138) on the other.

Further to that, word lengths are a property of linguistic styles. They evolve in the course of language acquisition as a speaker gets older (Best 2006c) and change in the course of language history (Best 2006i; Kromer 2006).

However, there are also extralinguistic effects influencing word length or being pertained by them: in case of proper names e.g. it plays a differentiating role. Seibicke (1982: 106) e.g. points out that in Germany there is a distinct tendency that girls' first names are longer than those of boys; in Italian press articles Hollberg (1997: 134ff.) observed that on the average Italian names are longer than non-Italian ones. Politeness and legibility are further aspects being related to word length (Best 2006g).

These and formerly given details allow the conclusion that word length – and the same applies to sentence length – is anything but a merely trivial linguistic property.

The text examples given in the following paragraphs are meant to demonstrate which word lengths must be expected as a minimum and by which linguistic laws they are controlled. This requires some simple theoretical reflections.

4. Discovery of Linguistic Laws I: The Distribution of Word Lengths

4.1. Preliminary reflections

Conceptual experiments can e.g. be made to get onto the track of linguistic laws. This is to be shown by three examples.

- Imagine you had to do with two languages, one has a phoneme inventory of 20 phonemes, the other has one of 40. What does this mean with respect to the average word length in those languages, if all the other circumstances (extent of vocabulary, allowed phoneme combinations in a word etc.) are identical? Which of the two languages will comprise the longer words?
- Imagine the words "bild-(en)", "Bildung", "Bildungspolitik": which of the three will occur most frequently, which will be the second most frequent word in a large corpus (acc. to your assumption)?
- Imagine you are writing two letters with each of them containing 100 words. The first letter comprises 50 one-syllable words, the second 70. What does this mean for the occurrence of two-syllable words, what for the words being even longer?

It should be clear that such reflections lead to expectations or hypotheses concerning the potential links between the properties mentioned. However, it has to be checked whether those reflections really prove successful. For that purpose, the respective hypotheses must be set up in an appropriate form, texts must be evaluated and the respective hypotheses must be tested for the material accumulated. In connection with the last conceptual experiment those steps of procedure should be presented below. It must be emphasized that discovering that regularity does not automatically mean the discovery of a law. The way to a law is very long: it begins with definitions, continues with measurements, testing the existence of some trend, derivation of the trend from a background theory, showing the interrelations of the regularity with other regularities, generalizations, further testing, etc. Going this long way, one usually cares for finding a well fitting function or distribution to individual cases that contain some boundary conditions, even if one does not know preliminarily how they influence the fitting. Step by step one generalizes, abandons the level of correlations and begins to theorize. But a beginner should stay at the lower levels and merely show the existence of some regularity. This can be done e.g. by showing the existence of a correlation.

The example chosen for the testing of the correlation between length of word and its frequency is the beginning of Pestalozzi's fable *Hühner, Adler und Mäuse* beginning as follows:

"Die Hühner rühmten ihr Gesicht und sagten selber zum Adler: „Auch das kleinste Korn liegt heiter vor unsern Augen." – „Arme Hühner!" erwiderte dieser, „das erste Kennzeichen eines guten Gesichts ist dieses: von allem dem nichts

zu sehen, was euch in die Augen fällt“ (Johann Heinrich Pestalozzi, *Fabeln*. Zürich: Manesse 1992, S. 48) [9]

If you now want to examine the number of syllables in the individual words you must define “word“ and “syllable“. “Word“ stands here for the “orthographic word“, which is an uninterrupted chain of graphemes – in alphabetic languages - bordered by blanks or punctuation marks. Hyphens and separation marks are not deemed to be punctuation marks; the same applies to apostrophes and dashes. The number of syllables per word is defined by the number of vowels, diphthongs/triphthongs or syllabic consonants (e.g. [r] and [l] in Slovak in words [hrst’], [vlk]) in a phonetic word. The number of syllables of the words in the overall text (without heading) is presented as follows:

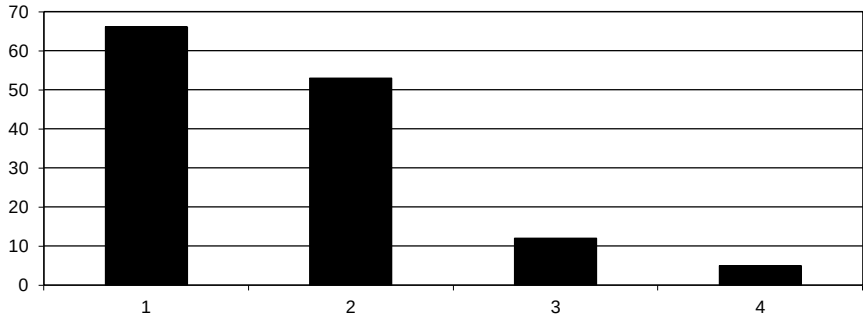
1-2-2-1-2-1-2-2-1-2-1-1-2-1-1-2-1-2-2-2-2-4-2-1-2-3-2-2-2-1-2-1-2-1-1-1-2-1-1-1-1-2-1-2-2-1-1-3-1-3-2-1-1-1-2-1-1-1-1-1-2-1-2-1-2-2-2-2-2-2-1-2-2-2-2-1-2-3-3-1-1-1-2-1-2-1-3-1-1-2-1-3-2-1-1-1-1-1-3-1-3-2-1-2-1-1-1-1-1-1-1-2-1-1-2-1-4-4-1-2-2-2-4-2-2-3-2-3-1-1-1-1-2-3-1-4.

The result is a sequence of numbers without any evident regularity; it can only be recognized that this text does not contain words of more than four syllables. Three- and four-syllable words are relatively rare. If numbers – and therefore words of different length – are ordered according to frequency, the overall text results in the following table (cf. Table 4.1):

Table 4.1
Frequency of different word lengths (number of syllables per word) in text 1: Pestalozzi, Hühner, Adler und Mäuse (Pestalozzi, *Fabeln*, p. 48)

x	n_x
1	66
2	53
3	12
4	4

[9] “The chickens praised their faces and even said to the eagle: “We can even see the smallest grain clearly before our eyes.“ – “Poor chickens!” the eagle replied. “The first feature of a good face is this: never see anything which might drop into your eyes.“ (Johann Heinrich Pestalozzi, *Fabeln*. Zurich: Manesse 1992, p. 48)



.

Figure 4.1. Graphical results of Table 4.1
Key: *x*: syllables per word (diagram: *x*-axis);
n_x: number of words with x syllables (diagram: *y*-axis)[10].

The table and the graphical diagram show that the seeming disorder is nevertheless based on a regular distribution. The following four examples (cf. Table 4.2 to 4.5) show that this is not an individual, accidental result.

Table 4.2
Frequency of different word lengths (number of syllables per word) in text 2: Pestalozzi, Das Menschenvertilgen (Pestalozzi, *Fabeln*, p. 42f.); text 3: Letter by H. Böll to E.-A. Kunz, 11.11.52 (Strobel 1996: 72); text 4: Kletternder Kastenklau, *ET*[11], 3.5.93, p. "Duderstadt" (Best 1997c: 9); text 5: Das Ende des Propheten, *PC Professionell* 9/ 1996, p. 60 (Nitsch 1997: 67)

	Text 2	Text 3	Text 4	Text 5
x	n_x	n_x	n_x	n_x
1	130	226	78	315
2	93	125	37	195
3	21	57	20	79
4	10	13	10	36
5	1	4	7	11
6		1	1	1
7			0	3
8			1	

[10] For reasons of space we do without the indication of x and n_x in the diagrams; a table is given for each diagram, and therefore it is easy to obtain the correct allocation.
[11] *ET: Eichsfelder Tageblatt*, local edition of the *Göttinger Tageblatt.*

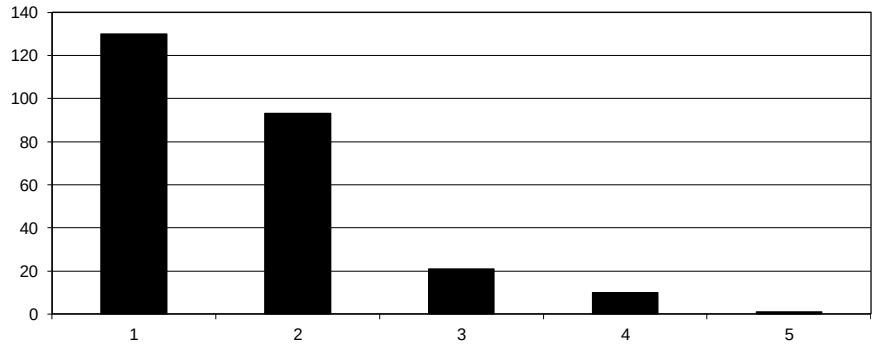


Figure 4.2a. Diagram related to text 2: Pestalozzi, *Das Menschenvertilgen*

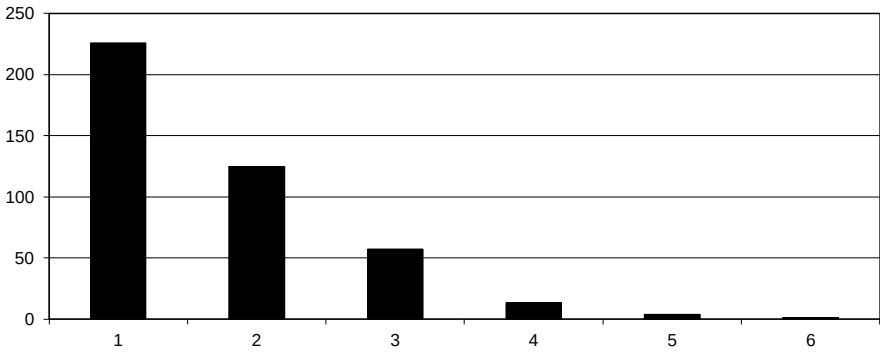


Figure 4.2b. Diagram related to text 3: *Letter by H. Böll to E.-A. Kunz, 11.11.52*

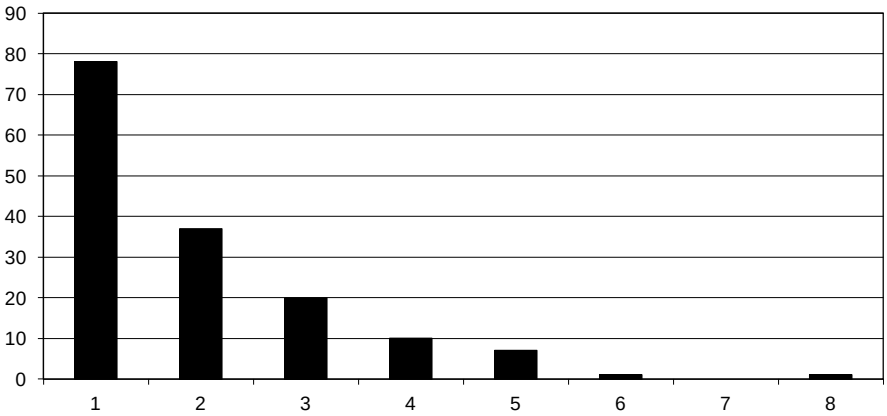


Figure 4.2c. Diagram related to text 4: *Kletternder Kastenklau, ET,* 3.5.93, p. ”Duderstadt“

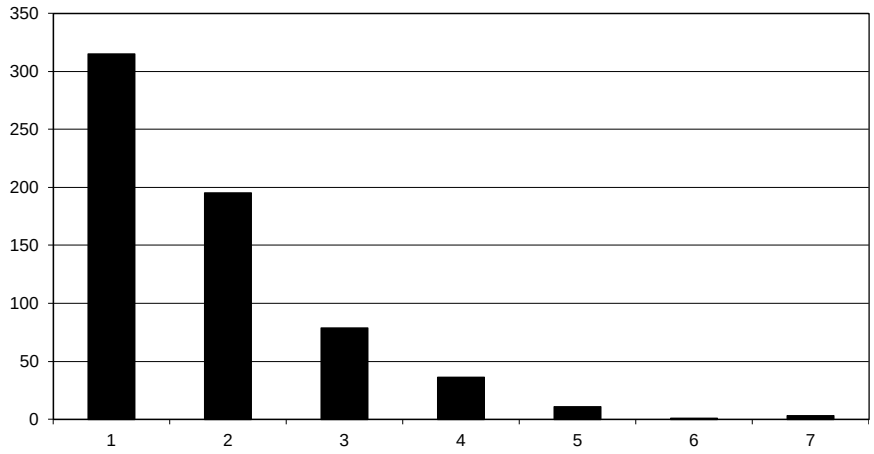


Figure 4.2d. Diagram related to text 5: Das Ende des Propheten, *PC Professionell* 9/ 1996, p. 60

The two texts by Pestalozzi show an especially strong concentration of one- and two-syllable words; in the other three the longer word lengths play a major role. It is obvious that differences originate from the fact that text sorts are very different (this represents a boundary condition, i.e. a special character of the given text). However, in all examples the shorter words are the more frequent ones, if the only rarely occurring longest word lengths are abstained from. All texts seem to be governed by a similar ordering principle.

The question is how to get ahead? Besides, there are languages using no alphabetic script and other ones in which the non-syllabic clitics are written separately or together with the word (cf. Slavic languages). There are no objective criteria; one must perform the analysis applying all possible definitions. The "correct" solution is that which optimally agrees with the given hypothesis. However, several positive results are possible. In order to accept one of them, the next more abstract criterion is necessary, namely the relationship of the given property measured in the given way to another property. Here one accepts that solution which has the strongest and most links to other properties. This way is, so to say, infinite and is practiced in all exact sciences.

4.2. A law concerning the distribution of word lengths

Remember the third and last of the thought experiments proposed above whose solution can be commented on as follows:

If e.g. one- to four-syllable words are used, it is most probable that the letter with the 70 one-syllable words contains fewer two-syllable ones than the one with just 50 one-syllable words. For the two- to four-syllable words just 30 possibilities remain, whereas there are 50 in the other letter. Of course, this is not quite certain, because 28 of the remaining 30 words could have two syllables; e.g. one three-syllable word and one four-syllable word could come in addition. But in the letter with 50 one-syllable words 20 two-syllable words and 15 three- and four-syllable words each are thinkable. This solution, however, is less proba-

ble, as a view of the tables and graphic charts of the five evaluated texts shows, though, of course, they cannot be deemed really representative.

The same observation, which is applicable to the ratio of the two- and one-syllable words in a text, can then be applied to all word lengths in a text: the number of the three-syllable words in a text depends on the number of the two-syllable ones etc. To put it briefly: the frequency of words of a specific length occurring in a text depends on the frequency of the shorter words. Specific proportions exist among the differently long words. This can be expressed by the following formula:

(4.1) $P_x \sim P_{x-1.}$

(P_x: probability of the occurrence of a word of the length x; P_{x-1} for words of the length x-1 accordingly).

As long as just two adjacent word lengths are considered, that proportion can be deemed constant: $P_2 = a\ P_1$. However, the graphic chart allocated to the fable by Pestalozzi shows that the ratio between the different word lengths is very different. That fact is taken into account by inserting a function $g(x)$ for that proportion:

(4.2) $P_x = g(x)\ P_{x-1.}$

The meaning of $g(x)$ can be characterized as follows: "The function $g(x)$ is an order parameter caring that a text serves its purpose, i.e. the communicative functionality of the linguistic system survives, the speaker does not fluctuate excessively etc." (Altmann & Best 1996: 166-167). At another point another author says: "The function g(x) represents a self-organizational entity, for example an attractor, which imposes an order on word length in terms of frequency in spite of the fact that the temporal sequence of words is chaotic with regard to length" (Wimmer et al. 1994: 102).

This, however, is no testable model yet. The question is how a more exact determination of $g(x)$ can be obtained. This requires further reflections that according to Altmann (1988a: 56) are linked to reflections by Zipf (1949: 21) on the two counteracting forces "Force of Unification" and "Force of Diversification" that can be transferred into the discussed phenomenon of word lengths rather easily. The effect of "speaker's economy" is simplifying and unifying, whereas "auditor's economy" is diversifying. This means that it is the speaker's aim to reduce his energy input when speaking. This can also result in a tendency to shorten words. The listener's efforts are contrary: the less linguistic information offered by the speaker, the bigger his problems concerning the understanding of the speaker. Maybe, he must ask the speaker to repeat his statement, to speak more distinctly etc. In communication it is important that a balance is established between the partners to avoid an unnecessarily high effort by the speaker in connection with his statements and an undue effort by the listener trying to understand what he hears. On the other hand, the diversifying force can also come from the speaker, e.g. when he/she varies his/her wording for reasons

of style that also results in varying word lengths and in a rejection by the listener due to difficult comprehension. Further requirements of speaker and hearer are defined in Köhler (2005).

If a parameter *a* is defined for the diversifying force and a parameter *b* for the unifying force, the function

(4.3) $g(x) = a/(b+x)$

can be utilized. If function (4.3) is inserted into (4.2), some transformations performed result in the hyper-Poisson distribution, which is defined as:

$$(4.4) \qquad P_x = \frac{a^{x-1}}{b^{(x-1)}\,{}_1F_1(1;b;a)}, \quad x = 1,2,3,...$$

The formula is stated here in its 1-displaced form, as in most cases zero-syllable words are not assumed. In the Slavic languages, zero-syllable words, e.g. simple prepositions, may be considered proclitics. Here, ${}_1F_1(1;b;a)$ is the hypergeometric function, or simply the sum of the individual values; it is necessary for normalization. And $b^{(x-1)} = b(b+1)(b+2)\ldots(b+x-2)$. One can, of course, use also a simple function instead of a normalized distribution. The above formula (4.4) represents a distribution that seems to have a basic significance for word lengths in texts:

So far, no distribution in approx. 4,000 texts in approx. 50 languages (state May 2002) has proved to be an appropriate model nowhere near as frequently as this one.

This distribution plays an outstanding role especially in ancient languages (Old Greek, Ancient Hebrew, Old High German, Old Icelandic, Old Church Slavonic/Old Bulgarian, Latin); however, it did not prove successful in some Old Russian texts.

Wimmer et al. (1994) and Wimmer & Altmann (1996) show that the hyper-Poisson distribution belongs to a family of distributions that can all be generated from approach (4.2) in case of slightly modified assumptions for *g(x)*. Further distributions can be developed from that by different methods (randomization, mixing, convolution, generalization acc. to Feller et al.). For a detailed and critical overview of those and other attempts concerning the modeling of word length distributions see Grzybek (2006); for an overview of the state of research see Best (2005). A unifying approach can be found in Popescu, Best, Altmann (2014).

The 1-displaced hyper-Poisson distribution is a hypothesis that can now be tested:

Hypothesis: Word length frequencies in German texts are distributed according to the hyper-Poisson distribution.

The test method applied is the X^2 test (Chi-square test), which is used in all sciences applying test methods. Appropriate software exists for this test: the

Altmann-Fitter (1997). That software allows the fitting to and testing of almost 200 different distributions in connection with arbitrary files, among them the hyper-Poisson distribution. The result of the fitting is the information on two test values: *P* is the probability that the fitting of the distribution conforms to the empirically found data at least as well as or even better than the value indicated. For linguistic purposes it is sufficient when $P \geq 0.05$. Unfortunately, there are cases in which this test value fails: the longer a file is the worse *P* is despite otherwise identical ratios. *P* cannot be defined either when a file comprises too few classes (here: too few words of different lengths). In both cases a different test value, the discrepancy coefficient C signaling satisfactory fitting with $C \leq 0.01$, must be taken into consideration then (cases with $0.01 \leq P < 0.05$ and $0.02 \geq C > 0.01$ are not deemed satisfactory, but, however, they are not that bad that the fitting must be considered a failure).

The *Altmann-Fitter* (1997) should now be taken to test whether the hypothesis applies when word lengths in German texts occur in compliance with the hyper-Poisson distribution. For that purpose, the fitting of the 1-displaced hyper-Poisson distribution to the first of the Pestalozzi texts is demonstrated and explained by using the *Altmann-Fitter* (1997).

Excursus I: Working with the *Altmann-Fitter* (1997): the fitting of distributions to files

To be able to fit a distribution to a text file this file must be available in a specific form. After having loaded the data according to the Handbook, load the start menu and proceed as follows: Start ➔ all programs ➔ accessory equipment ➔ C: prompt; then load an arbitrary data carrier x (disc, hard disk, USB stick) and enter: x: > edit (enter command "edit"). A text window is displayed. Then enter your file flush left in the following form:

1	66
2	53
3	12
4	5

This is the file for *Pestalozzi, Hühner, Adler und Mäuse*.
Apart from those two columns of figures the files must not contain any other information – not even additional blanks being invisible on the screen! Each file is provided with a file name on its own on the separate data carrier, which in this case is "WLSI15.dat". The text before the dot is arbitrary, however, the name of the file must end with the suffix "dat" after the dot. After the files have been entered, leave the editor "edit" by means of the command "exit".

Then, the Fitter can be loaded. As the smallest class in the file is $x = 1$ (i.e.: one-syllable words with which usually zero-syllable words are not accepted), the Fitter fits the distribution chosen in the 1-displaced form.

The result:

-- ALTMANN-FITTER 2.0 -- Result of fitting

Input data: WLSi15.dat
Distribution: hyper-Poisson (a,b)
Sample size: 136

Moments:
M1 = 1,6765
M2 = 0,6159
M3 = 0,5287
M4 = 1,4482

Best method is method 2 of 5

Parameters:
a = 0,542685237113938
b = 0,710498128391065
DF = 1
X^2 = 1,5640 $P(X^2)$ = 0,2111 C = 0,0115

X[i]	F[i]	NP[i]
1	66	65,9343
2	53	50,3613
3	12	15,9780
4	5	3,7264

This is the test result obtained when the hypothesis is tested with respect to the 1-displaced hyper-Poisson distribution being a good model for the text by Pestalozzi.

The new version of the *Altmann-Fitter* yields many other functions applied in text analysis.

Explanations

The fitting of the 1-displaced hyper-Poisson distribution (with the parameters *a*, *b*) to the text file supplies the following information:

Input data: This is the name given to the file.

Best method: The 1-displaced hyper-Poisson distribution comprises five different iterative methods to be applied to the fitting to a file. The best fit is chosen automatically; in this case it is method 2.

Moments: "M1" – "M4" indicate the so-called "Moments" of the distribution (Altmann 1988a: 46-51; Altmann ²1995: 133ff.). "M1" is the 1st moment, which is the arithmetical mean (mean value) of the distribution, "M2", the 2nd moment, expresses the spread/variance (Swoboda 1974: 37ff.). "M3" is the 3rd moment of the distribution, which helps calculate skewness[12] / asymmetry b_1 (Swoboda 1974: 72f.) as

$$b_1 = \frac{M_3}{M_2^{\frac{3}{2}}}; \tag{4.4}$$

a distribution is symmetrical, if $b_1 = 0$, it is asymmetrical to the left with $b_1 < 0$ and asymmetrical to the right with $b_1 > 0$.

The excess (steepness) b_2 of the distribution (Swoboda 1974: 73) is calculated by means of the 4th moment "M4" as

$$b_2 = \frac{M_4}{M_2^2} - 3. \tag{4.5}$$

A distribution with $b_2 = 0$ is called normal, with $b_2 < 0$ being flat and $b_2 > 0$ being steep.

Parameters: Those parameters are always parameters concerning the distribution fitted to the input file. It is often difficult to interpret those parameters, a problem the already mentioned *Grazer Projekt zur Quantitativen Text-Analyse* (2002) will deal with (see p. 7).

Further details: (Chi² =) X^2 is the chi-square; *DF* is the number of the degrees of freedom (DF). *P* is the probability that X^2 is as good as the result of the calculation or even better, and *C* is the discrepancy coefficient $C = X^2/N$; it is used when the probability *P* cannot be determined because of $FG = 0$ (*N*: sample size). Even when a very large file has to be processed, *C* should be used as the test criterion. The parameters of this distribution are indicated as *a*, *b*. *X[i]* stands for the words with 1, 2 ... syllables, *F[i]* is the observed number of words with 1, 2 ... syllables in the respective text and *NP[i]* expresses the computed number of words with 1, 2, ... syllables that were calculated on the basis of the 1-displaced hyper-Poisson distribution.

Evaluation of the test results: see p. 30 for the criteria for the evaluation of the test values *P* and *C*. It can be seen that the Altmann-Fitter indicates $P = 0.21$ as the test value; that means that the 1-displaced hyper-Poisson distribution is a good model for the file tested, which comprises 136 words only and is therefore

[12] Skewness and excess are defined to be a deviation from the normal distribution: Altmann 1995: 144f.; Sachs [5]1978: 81.

rather small; the criterion of $P \geq 0.05$ is met. The discrepancy coefficient C is neglected here, because the file is small and the degrees of freedom can be determined.

Degrees of freedom: "Let's assume we found out that a small sample of 4 cases results in a mean value of 25. In that case the total must be 4 x 25 = 100. Irrespective of the actual values of the individual cases we are *free* to ascribe any arbitrary value on condition that the total of the values is always 100." (Kennedy 1985: 195).

So, the result could have been 20 – 30 – 25 – 25 or 25 – 29 – 22 – 24, but the total is always 100. In such cases you are free to determine the first three values; with respect to the fourth value, however, you are not free to determine it: "Should we ascribe the values of 25, 29 and 22 to three of the four cases and sum them up, we get the total of 76. If we ascribe those values to those cases, we *restrict* the option for the value of the fourth case. The value of the fourth case is simply 100 – 76 = 24." (Kennedy 1985: 195).

So, in case of x classes you have $x - 1$ degrees of freedom. The number of degrees of freedom is further restricted by the number of parameters of the distribution to be adapted. The Pestalozzi text results in four classes and two parameters of the hyper-Poisson distribution: 4 – 2 – 1 = 1, so there is one degree of freedom. That is exactly what the Altmann-Fitter determined.

4.3. Test results concerning word lengths (measured by the number of syllables per word) in the five texts already mentioned

a) to e) include the test results concerning the word length distributions in the five texts already indicated. For reasons of clarity the tables are always accompanied by graphical diagrams in which the values observed and those calculated can always be compared directly; the left black column includes the values observed, the right white column includes the calculated ones. Test results were as follows:

Table 4.3

a) Text 1 – Fitting the 1-displaced hyper-Poisson distribution to the word lengths (syllables per word) in: Pestalozzi, *Hühner, Adler und Mäuse* (Pestalozzi, *Fabeln*, p. 48)

x	n_x	NP_x
1	66	65.93
2	53	50.36
3	12	15.98
4	5	3.73
$a = 0.5427$ $b = 0.7105$		$X_1^2 = 1.564$ $P = 0.21$

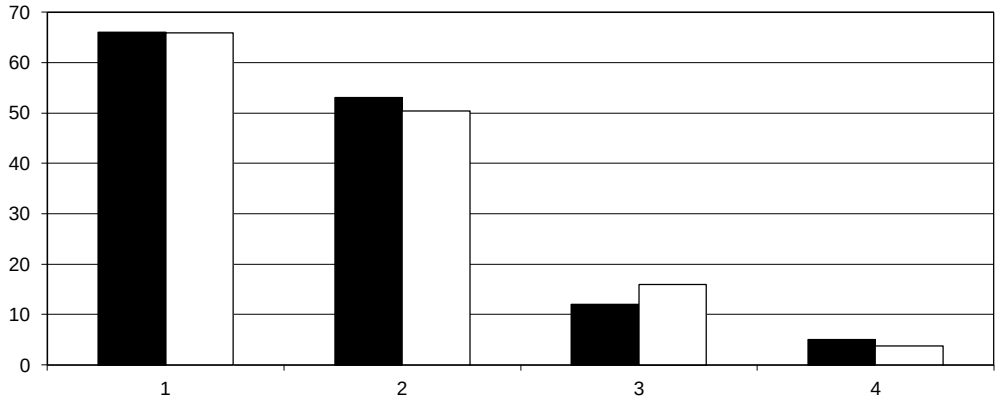


Figure 4.3. Concerning the data in Table 4.3.

Key for the tables:

x: number of syllables per word;
n_x: observed number of words with x syllables;
NP_x: number of words with x syllables calculated on the basis of the hyper-Poisson distribution;
a, b: parameters of the hyper-Poisson distribution;
X_k^2: value of the chi-square with k degrees of freedom;
P: probability of exceeding the given chi-square;
C: discrepancy coefficient.

The test result signals – as already mentioned - by $P = 0.21$ (i.e.: $P \geq 0.05$) that the 1-displaced hyper-Poisson distribution for the word lengths in this fable by Pestalozzi indicates an appropriate model. The relevant graphical diagram already shows that just minor differences exist between the columns for the observed (black) values and the calculated (white) ones.

Compared to this the tests of the other four texts show the following results (cf. Tables 4.4 to 4.7):

Table 4.4

b) As to Text 2: fitting of the 1-displaced hyper-Poisson distribution to the word lengths (syllables per word) in: Pestalozzi, *Das Menschenvertilgen* (Pestalozzi, *Fabeln*, p. 42f.)

x	n_x	NP_x
1	130	128.99
2	93	87.97
3	21	29.93
4	10	6.79
5	1	1.33
$a = 0.6791$		$X_2^2 = 4.566$
$b = 0.9958$		$P = 0.10$

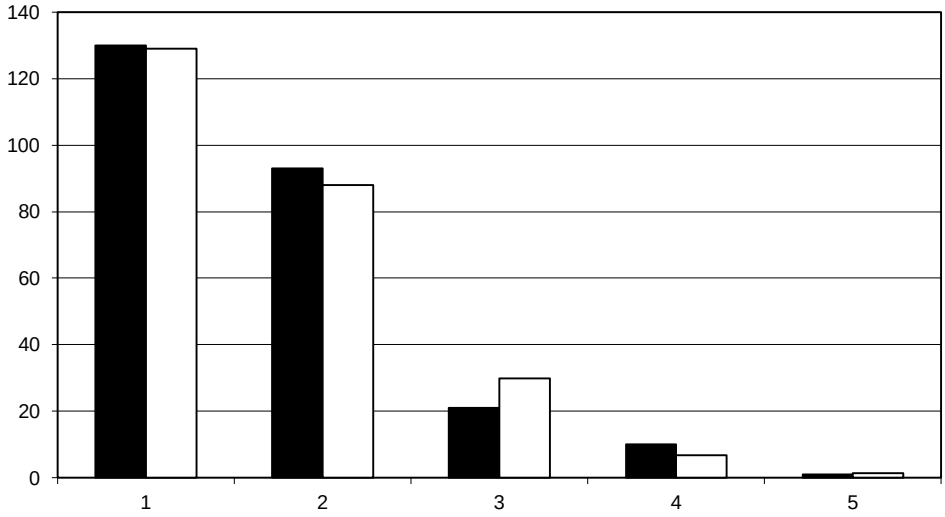


Figure 4.4. Concerning the data in Table 4.4

In this case the test criterion is met again ($P = 0.10$).

As to the following text the fitting of the hyper-Poisson distribution results in $P = 0.75$ and is distinctly better than in the cases before (cf. Table 4.5):

Table 4.5
As to Text 3: fitting the 1-displaced hyper-Poisson distribution to word length (syllables per word) in: Böll, Brief an E.-A. Kunz, 11.11.52

x	n_x	NP_x
1	226	226.22
2	125	128.09
3	57	51.05
4	13	15.70
5	4	3.93
6	1	1.01
a = 1.3462		X_3^2 = 1.233
b = 2.3775		P = 0.75

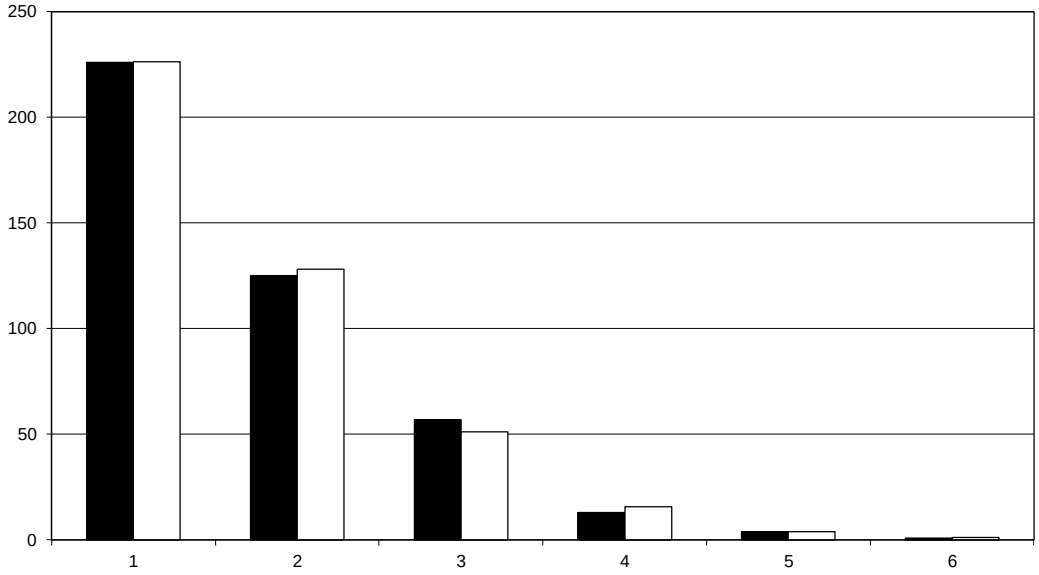


Figure 4.5. Concerning the data in Table 4.5

Table 4.6
As to Text 4: fitting the 1-displaced distribution to word lengths (syllables per word) in: Kletternder Kastenklau, *ET*, 3.5.93, p. "Duderstadt"

x	n_x	NP_x
1	78	74.90
2	37	40.16
3	20	20.54
4	10	10.04
5	7	4.70
6	1	2.11
7	0	0.91
8	1	0.64
a = 11.0726		X_4^2 = 2.287
b = 20.6493		P = 0.68

The vertical lines with the word lengths $x = 7$ and $x = 8$ indicate the pooling of the respective length classes here as well as in the other ones. The size of the expected class must be greater than 1.

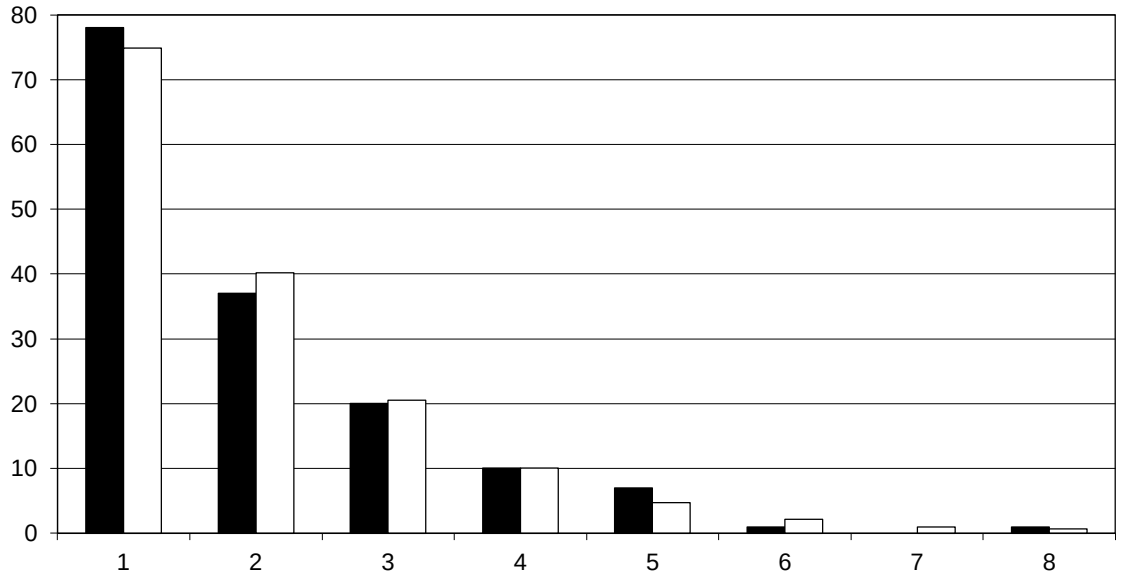


Figure 4.6. Concerning the data in Table 4.6.

Table 4.7

As to Text 5: fitting the 1-displaced hyper-Poisson distribution to the word lengths (syllable per words) in: *Das Ende des Propheten*, *PC Professionell* 9/ 1996: 60

	n_x	NP_x
1	315	317.08
2	195	187.69
3	79	87.13
4	36	33.27
5	11	10.79
6	1	3.04
7	3	1.00
$a = 2.1523$		$X_3^2 = 1.285$
$b = 3.6359$		$P = 0.73$

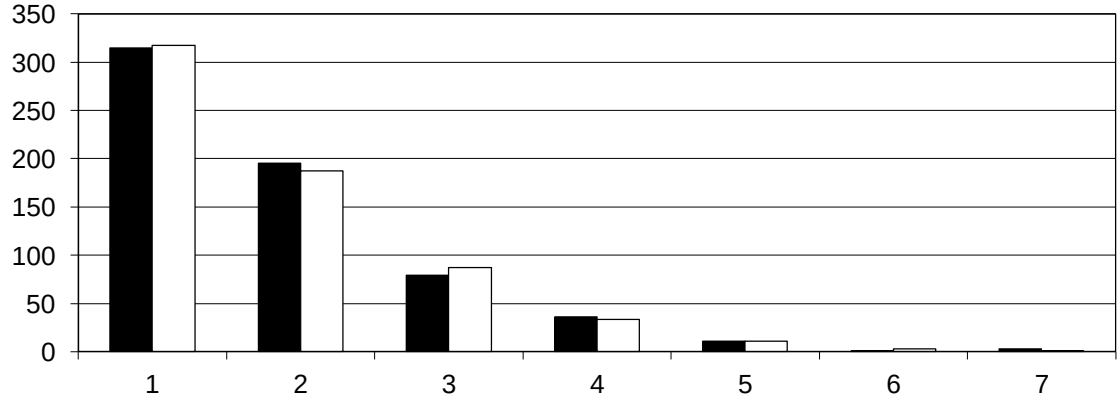


Figure 4.7. Concerning the data in Table 4.7

As a result referring to all five texts it can be stated that the 1-displaced hyper-Poisson distribution is an appropriate model for them; it is in full compliance with our hypothesis. In all cases the test criterion $P \geq 0.05$ is met.

4.3.1 Summary

Together with Zipf's reflections on unification and diversification forces the third thought experiment (p. 29 f.) obviously resulted in a theoretical approach that turned out to be corroborated with respect to testing word lengths in German texts. So far, approximately 1,400 texts from the Old High German period to the present have been tested and the hyper-Poisson distribution failed five times only as the appropriate model. In those cases, however, distributions that can be derived with a slightly altered approach for $g(x)$ could partly be fitted. Thus, after some modifications the approach $g(x) = a/x$ leads to the Poisson distribution, which is the model which was first considered the law for the distribution of word lengths by the Russian army physician Sergej Grigor'evič Čebanov (1947) and a little later – but obviously without knowing Čebanov's findings – by Wilhelm Fucks (1955):

$$P_x = \frac{e^{-a} a^{x-1}}{(x-1)!}, \qquad x = 1,2,... \tag{4.7}$$

Altmann & Altmann (2005: 105f.) used this model for the word length distribution in Goethe's ballad *Erlkönig*.

For further relations see Wimmer & Altmann (1996: 114). Details will not be developed here any further; however, it is important to know that distributions used in this paper allow substantiations which are all based on modifications of $g(x)$ in the hypothesis (4.2).

Why are such modifications needed? Three answers exist:
Observations show that word lengths do not follow the hyper-Poisson distribution in all languages. In Finnish e.g. different distributions must be applied (Vettermann & Best 1997).

It must generally be expected that in one and the same language the distribution of different linguistic entities[13] is different:

Different entities can conform to different distributions; e.g. syllable or sentence lengths need not conform to the hyper-Poisson distribution, even if word lengths do so.

A linguistic entity like word length can follow different models in a language, if several authors, styles or text types are analyzed or texts of different development phases in a language are processed. This is e.g. the case with Finnish poems and prose texts (Vettermann & Best 1997).

[13] Entity: unit (e.g.: sentence, word) or class of units (e.g.: words of a certain length; word class).

When modeling a phenomenon, one always strives for unification. If various models have been derived, one seeks a general solution that is able to yield all models. But this is always a dream of future. A unified model for all type of lengths in language has been found by Popescu, Best, Altmann (2014) but since it has three parameters, sometimes the number of classes is too small. In that case one can add one-two classes containing zero observations.

4. Usually, one uses a discrete probability distribution as a model of length but one can use also a continuous function/distribution. Mathematical models have nothing common with truth, they merely express exactly our hypotheses, they are testable and can be manipulated formally.

4.4. Corpus problems

To perform tests like the ones shown with word lengths the selection of texts to be processed requires some reflections concerning the following questions:

1. Are there texts that are especially appropriate or especially inappropriate for such testing?
2. Must texts have a specific minimum length?
3. Are there maximum limits for text lengths?
4. Should short texts to be processed be united in a small corpus and be analyzed together?
5. Should texts be evaluated completely?
6. What is the result of observations concerning the homogeneity of short texts?
7. What has to be taken into consideration when processing dictionaries?

As to 1. Again a thought experiment may be used for introduction: Imagine you are sitting at your desk to write a text. It is a difference if you write a letter to your brother or sister or the text is to become a tale or a script for a lecture. An essential aspect is that the decision to write a letter to a family member results in a volume that is usually much smaller than in the other two cases. In quantitative linguistics, however, it is well known that the length planned for the text results in a different flow of information, e.g. the growth rate of new words (Orlov 1982b: 125).

So, even the aspect of text length is relevant. In addition, the longer texts let grow the risk that writing is interrupted; in turn, this can result in forgetting the original plan of text composition and a new conception is implemented from the point of interruption. Example: an originally planned short letter becomes distinctly longer by new ideas. However, this means that the originally homogeneous planning of the text has a breach; in short: as to style the letter is no longer homogeneous. Hammerl (1990) gave a reason for the problems occurring when processing word classes in texts by saying that the length of the few problematic texts is decisive. So, stylistic homogeneity is an important aspect; texts will often be more homogeneous if they are shorter, less revised and written more spontaneously. Any restriction of those conditions need not result in faults and therefore in problems during the adaptation of distributions; nevertheless, it is possi-

ble. Therefore, the ideal text is a spontaneously written private letter, which after being written is not corrected or only a little and mailed then. This does not mean that other texts should not be dealt with; it only means that faults concerning the homogeneity of texts may cause problems that probably would not occur otherwise (Orlov 1982b dealt with problems of the text volume planned and – related to it – text homogeneity; also see Altmann 1992).

As to 2/3. When dealing with the other two questions concerning text volumes it has to be considered that they can only be answered with respect to the subject processed. So far, the analysis of word lengths was based on how many syllables the words comprised. But now it has to be considered that in German texts only rarely words of more than ten syllables can be observed. In such cases texts of distinctly less than 100 words are sufficient (Best 1997b: 11-13). Hammerl (1990: 151) tentatively specified 2000 words as the upper limit that did not turn out to be empirically problematic. Those specifications can only be understood as a rough guide. No statistician has ever said what is the ideal size of a sample.

An important point is the definition of the units processed. Word lengths can also be defined on the basis of letters, sounds or phonemes per word. If one of those three possibilities is used, you will in any case get definitely more length classes than in case of the syllable being the measuring unit. If, however, the number of different word lengths is higher, it is necessary to process longer texts to achieve that rarer lengths are occupied to some degree as well. Nevertheless, it can be recommended to measure the length of a unit always in terms of its immediate constituents.

Those reflections have to be transferred to other units accordingly. If units smaller than words are analyzed, e.g. syllables, somehow shorter texts will do, whereas longer texts are needed in case of longer units (clauses, sentences). The choice of texts is also influenced by its functional style (Sowinski 1991: 33f.): in scientific texts longer words can be expected than it is the case in private letters.

As to 4. This question refers to the problem of text. Is it not better to unite very short texts to become a small corpus instead of processing them individually? At this point we should refer to what has been said about "Homogeneity". Every text must be considered a stylistic unit; several texts more or less implement different styles. So, a text mix violates the principle of homogeneity. Altmann (1988a: 68f.) demonstrated that fitting results could become worse when uniting two texts to a new mixed text. A corpus as a whole can be used as a source of e.g. grammatical rules, but not for the evaluation of text properties.

As to 5. For the same reason the formation of arbitrary text sections should be avoided. The best solution is an individual, continuous text being processed completely. If text sections are to be considered, natural sections like e.g. chapters of appropriate length should be chosen.

As to 6. On the other hand, those reflections on homogeneity should not be exaggerated. In the course of processing some observations did not meet expectations: In a German – Latin poem (Best 1996: 149, Text 20) there did not occur any problems with the fitting of the appropriate distribution; the same applies to texts that partly show a very long literary tradition till the latest edition (fables), do not originate from native-language authors (French or Latin letters from German speaking authors) or the like. Such development conditions could be a reason for insufficient text homogeneity, which, however, does not always prove to be correct. A computer simulation by Schmidt-Samoa (2002), who analyzed German texts of different text types with respect to the distribution of word lengths, is also noteworthy; as always, results were quite convincing. In further operations he decomposed those texts into paragraphs and used them to artificially compose into new texts in different splittings. The astonishing result was that fitting with those artificial texts was not significantly worse. This requires clarification: Does this only apply to very short texts like the ones analyzed? Does it always apply to longer ones? Does it apply to texts of all kinds? And so on. Meanwhile, further analyses should be based on the homogeneity assumption; it permits the development of ideas when problems with the fittings of distributions could possibly be expected.

As to 7. The different types of dictionaries are based on different conceptions that will most probably be reflected in word length distribution. Further to that, dictionaries are a completely different type of texts compared to letters, novels etc. So, distribution models that can be expected are more different than it is the case in connection with "natural" texts.

For further details concerning corpus problems see Grotjahn & Altmann (1993: 143ff.); on the homogeneity of files see Altmann (1992).

4.5. Test results concerning word lengths in dictionaries

In the meantime, several experiences have been obtained on the distributions governing word lengths in more than 50 languages; especially for the German language we could learn how other entities behave, especially lengths of rhythmic units as well as lengths of morphs, syllables and sentences, but also word classes. Just in individual cases the distributions of those and other entities are known for further languages, e.g. those for word classes in French, the length of syllables in English, the length of sentences in English and Russian, the lengths of rhythmic units in Classical Greek as well as those for the ideographs in Chinese. However, there remains a lot to be done to obtain a somewhat useful overview of such phenomena. So far, there is no hint that the distributions of any entity in any language behave differently with respect to the statement of the law hypothesis (4.2).

Now, it has to be shown which distributions are required unless only those of syllables per word in texts are analyzed. Attention has to be paid that the individual examples do not express that the same phenomenon in German must al-

ways be distributed in the way presented here; at the present state of analyses it is too early for generalization.

If word length distribution is not analyzed in individual texts, but in a comprehensive corpus as it is presented in frequency dictionaries, you have to do with a mixed text that actually is not really wanted (see previous section). However, this is a very voluminous mixed text that may compensate for the disadvantages mentioned thanks to its large volume. Concerning Kaeding (1897/98), however, the hyper-Poisson distribution turned out to be successful:

a) Word length distribution in frequency dictionaries

Table 4.8

Fitting of the 1-displaced hyper-Poisson distribution to the word lengths (syllables per word) in Kaeding's dictionary (acc. to Zipf 1935/ 1968: 23; 10,906,235 words)[14]

x	n_x	NP_x
1	5426326	5432151.34
2	3156448	3124163.92
3	1410494	1472860.87
4	646971	588308.17
5	187738	203852.30
6	54436	62371.57
7	16993	17084.58
8	5038	4236.04
9	1225	959.34
10	461	199.95
11	59	38.60
12	35	6.94
13	8	1.17
14	2	0.18
15	1	0.03
a = 2.6151	b = 4.5470	X_{10}^2 = 11871.54 C = 0.0011

[14] "Subsequently corrected by Kaeding to 10,910,777." (Zipf 1935/ 1968: 23)

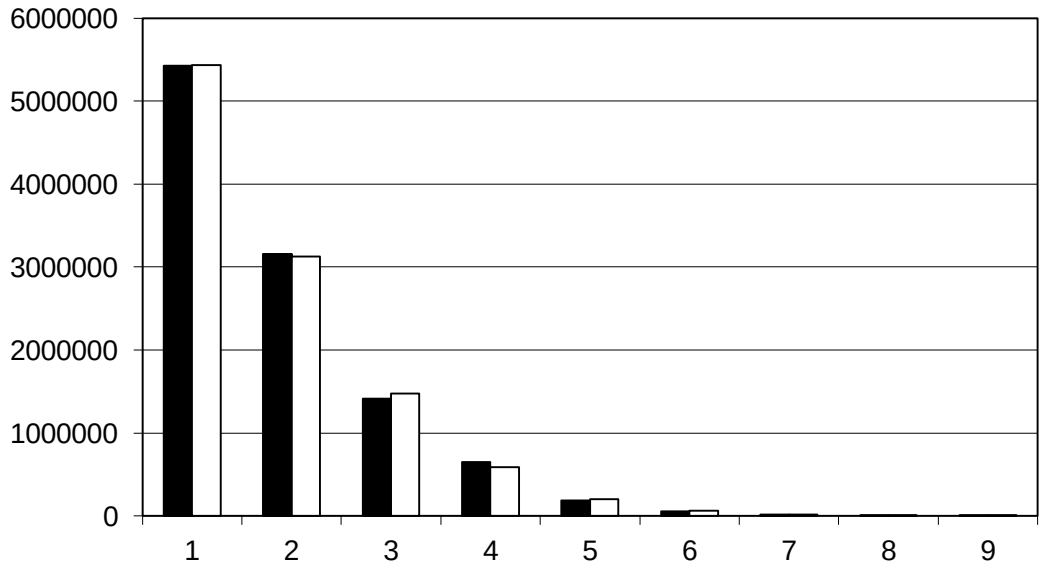


Figure 4.8. Concerning the data in Table 4.8

In contrast to individual texts dictionaries form very comprehensive files; therefore, the discrepancy coefficient *C* must be used as the test criterion instead of probability *P*; this applies to this case as well. *C* with *0.0011 < 0.01* indicates satisfactory fitting. The graphical diagram (figure 4.11) elucidates that, though it only represents the one- to seven-syllable words noticeably. Since *C* decreases with sample size, one can fit simple (not-normalized) functions to data as well and judge the results on the basis of the determination coefficient. Using this way, we do not falsify reality because mathematical models do not represent truth but merely express our human way of treating reality. We construct models for the sake of easier understanding, for more exact description, and, if possible, for explaining the behavior of the observed entities. If all respective entities can be captured by the same model and the model can be derived deductively, we may begin to speak about laws. In the present case one could apply e.g. the pulse function y = a*exp(-(x-b)/c)*(1-exp(-(x-b)/c)), yielding an excellent determination oefficient (D = 0.9998) which is a combination of two exponential functions but has merely three parameters.

b) Word length distribution in the alphabetically ordered dictionary

Table 4.9
Fitting the 1-displaced Conway-Maxwell-Poisson distribution to the word lengths (syllables per word) in Viëtor's *Deutsches Aussprachewörterbuch* (Menzerath 1954: 8, 98; 20,453 words)

x	n_x	NP_x
1	2245	1806.09
2	6396	6543.79
3	6979	6995.99

4	3640	3662.67
5	920	1155.44
6	214	246.06
7	42	38.01
8	11	4.47
9	6	0.48
a = 3.6232 b = 1.7609 X_5^2 = 192.341 C = 0.0094		

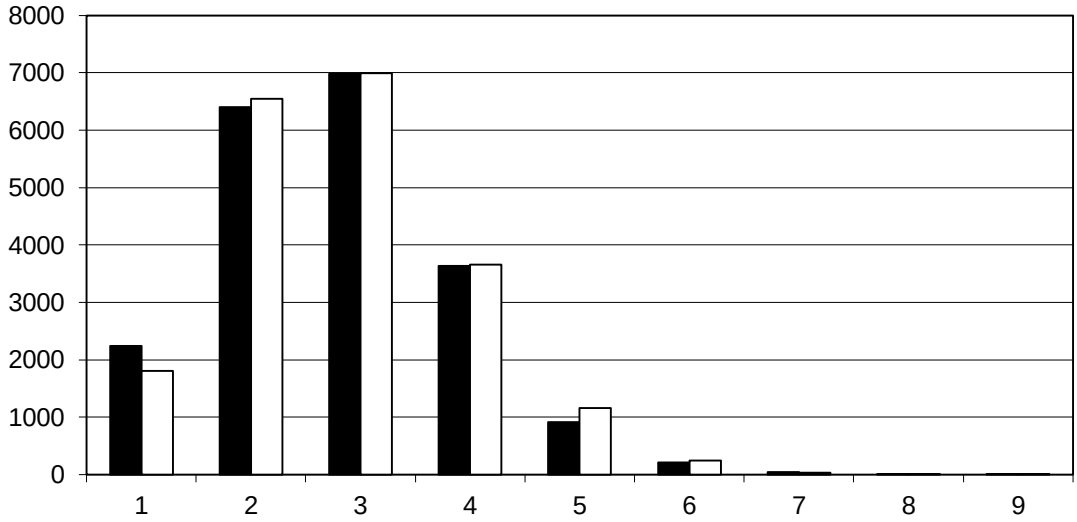


Figure 4.9. Concerning the data in Table 4.9

In such an alphabetically ordered dictionary the conditions are different, which is shown by the fact that each lexeme is entered only once. Those dictionaries do not reflect the frequency at which words are used. Therefore, it is not astonishing that in that case a completely different distribution must be chosen. This fitting, however, meets the criterion $C \leq 0.01$. The Conway-Maxwell-Poisson distribution is derived from $g(x) = a/x^b$ (Wimmer & Altmann 1996).

In dictionaries one may apply different sampling techniques: (a) systematic sampling, e.g. taking each first word on the page, or (b) random sampling, e.g. opening the dictionary randomly and taking some of the words, (c) authoritative sampling, e.g. deciding a priori what kind of words should be sampled. In any case, we have here different boundary conditions, which should be taken into account with modeling.

In general, the results for frequency dictionaries are very misleading. It is known that X^2 increases linearly with sample size, hence it cannot be used successfully in all cases. On the other hand, the coefficient C decreases with the increase of sample size and signalizes good results especially in cases where the sample size is large.

4.6. Word lengths measured by means of different units

So far, word length has been determined by the number of syllables in a word. However, word length can also be determined in a different way. If length is

measured by the number of morphs, phonemes or letters per word, results are as follows:

a) Word length distribution (number of morphs per word)

Table 4.10
Fitting the 1-displaced hyper-Poisson distribution to the word lengths in text 1: Pestalozzi, Hühner, Adler und Mäuse (Pestalozzi, *Fabeln*, S. 48) in terms of x = number of morphs per word.

x	n_x	NP_x
1	44	43.60
2	67	68.15
3	22	20.42
4	3	3.83
a = 0.3708		X_1^2 = 0.316
b = 0.2372		P = 0.57

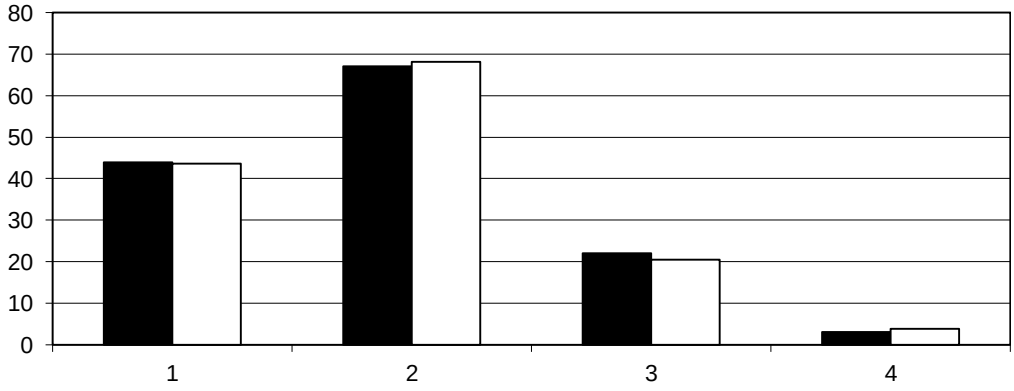


Figure 4.10. Concerning the data in Table 4.10

For further results concerning the number of morphs in words see Best 2001e.

b) Word length distribution (number of phonemes per word)

If one skips a level, one must reckon with boundary conditions that can be captured by additional parameters or whole functions. Unfortunately, the parameters or functions found in one language need not hold in another. Thus generalizing this approach may lead to a very extensive family of functions which may increase with every new language or text studied.

Table 4.11
Fitting the displaced hyper-Pascal distribution to word lengths in text 1: Pestalozzi, Hühner, Adler und Mäuse (Pestalozzi, *Fabeln*, 48) in terms of x = number of phonemes per word.

x	n_x	NP_x
2	20	22.05
3	41	39.89
4	23	27.29
5	21	17.59
6	13	11.10
7	7	6.93
8	4	4.30
9	4	2.65
10	1	1.63
11	0	1.00
12	0	0.61
13	2	0.96
k = 0.2169	m = 0.0722	q = 0.6027
X_6^2 = 3.71	P = 0.72	

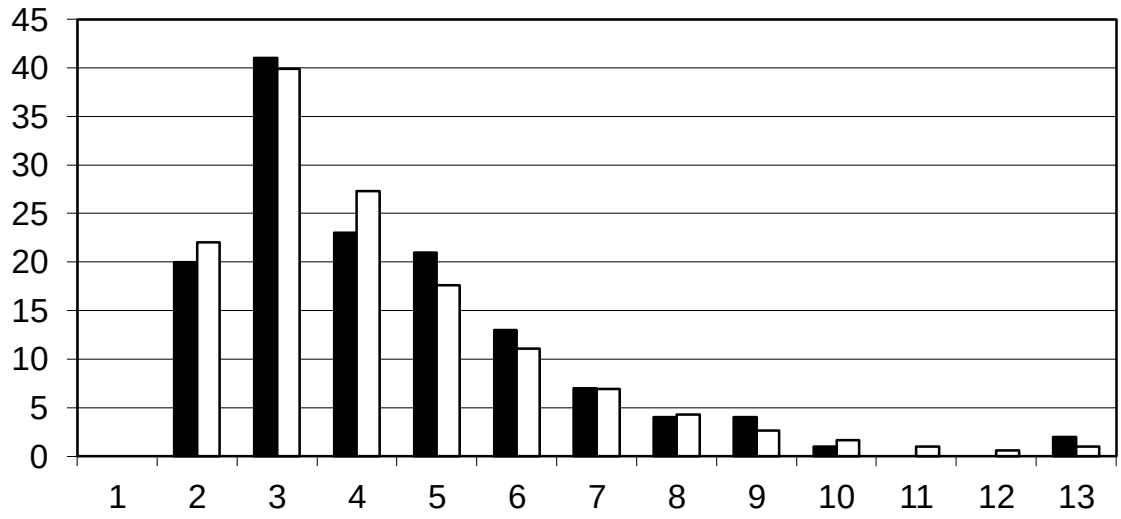


Figure 4.11. Concerning the data in Table 4.11
The hyper-Pascal distribution results from the approach $g(x) = (a+bx)/(c+dx)$. k, m and q are the transformed parameters of the hyper-Pascal distribution.

In the analyses of 60 press texts by Schneemann (2001), the hyper-Pascal distribution could be fitted except for one text. In these analyses, words with 1 and 2 phonemes, 3 and 4 phonemes etc. were pooled.

c) Word length distribution (number of letters per word)

Theoretically, this approach is not fruitful because letters are no immediate constituents of words, and in some languages this approach is not even appli-

cable (e.g. Chinese). Here, we do not neglect boundary conditions but link incommensurable units. A result is shown in Table 4.12 and Figure 4.12.

Table 4.12
Fitting the displaced hyper-Pascal distribution to word lengths in text 1: Pestalozzi, Hühner, Adler und Mäuse (Pestalozzi, *Fabeln*, 48) in terms of x = number of letters per word

x	n_x	NP_x
2	6	6.17
3	47	36.17
4	12	30.61
5	27	22.01
6	17	14.86
7	10	9.71
8	5	6.21
9	5	3.92
10	2	2.44
11	2	1.51
12	1	0.93|
13	2	1.45|
k = 0.5212	m = 0.0520	q = 0.5853
X_7^2 = 16.955	P = 0.0177	C = 0.1247

This test result is not satisfactory; P is within the interval $0.01 \leq P < 0.05$ and therefore not satisfactory. C does not produce an acceptable result either. Even the fitting of other (200) distributions results in rejections. This is a sufficient reason to consider this kind of measurement as irrelevant.

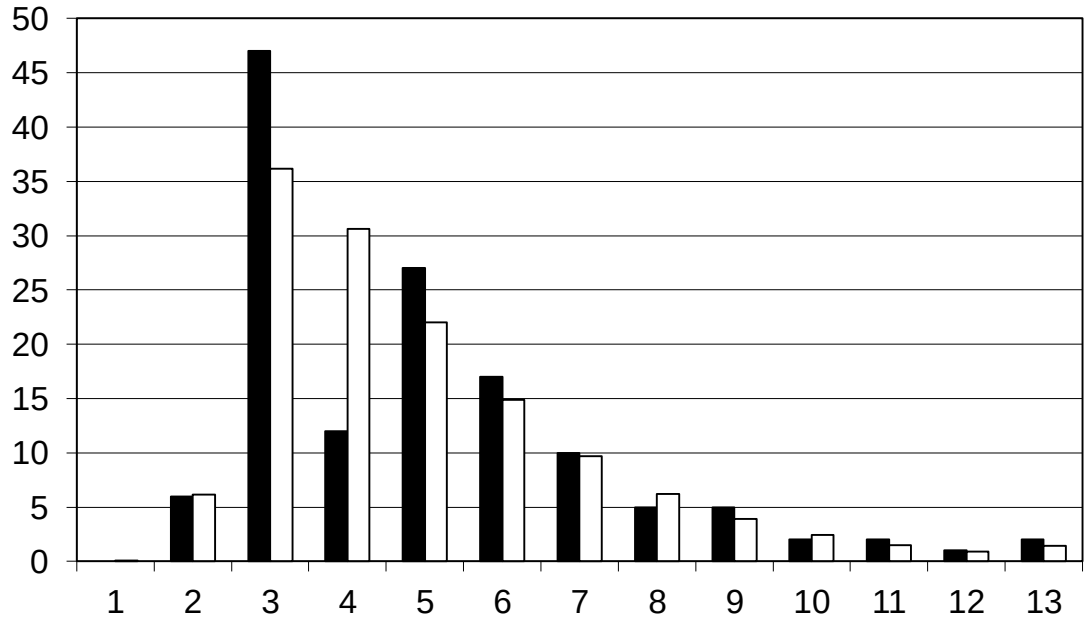


Figure 4.12. Concerning the data in Table 4.12

Table and graphical diagram show that the problem is caused by words with 3 or 4 letters. To compensate for such "local irregularities" as they occur here with words of 3 or 4 letters it is possible to "smoothen" the data generated by "pooling". When pooling words with 1 or 2 letters in class x = 1, those with 3 and 4 letters in class x = 2 etc. the result is as follows (cf. Table 4.13):

Table 4.13
Smoothed data from table 4.12

x	n_x	NP_x
1	6	7.07
2	59	60.85
3	44	37.39
4	15	17.86
5	7	7.69
6	3	3.13
7	2	2.01
$k = 0.8940$ $m = 0.0349$ $q = 0.3358$ $X_3^2 = 1.911$ $P = 0.59$		

The resulting fit turns out to be very satisfactory. Such a pooling of several word lengths, sentence lengths, word lengths etc. in one class is a usual process in QL. Kelih & Grzybek (2004) deal with this problem systematically by concentrating on sentence lengths. The kind of pooling can have an effect on the suitability of a model. However, the compensation for local irregularities can also be done in different ways (Wimmer, Witkovský, & Altmann 1999).

Herdan (1966: 23) presented a further and more comprehensive file concerning word lengths in German, determined according to the number of letters per word.

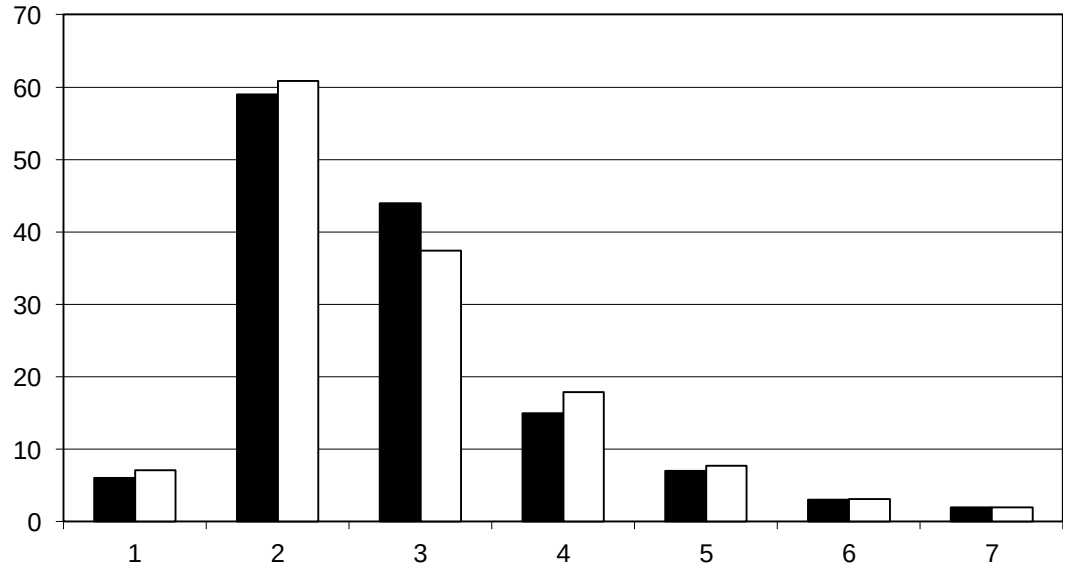


Figure 4.13. Concerning the data in Table 4.13

However, we do not recommend to apply this kind of measurement for four reasons:

(1) Not all languages have an alphabetic script (Chinese).

(2) Some languages apply the Latin alphabet, but the use of letters is so redundant that one counts something that does not exist (English, French) or it differs in each dialect.

(3) Since a level has been skipped, the boundary conditions can differ from language to language and the more languages one analyzes, the more special cases will be found. Hence the derivation of a law will be rather more difficult.

(4) The inclusion of such a diversified result into the Köhlerian control circuit will be almost impossible.

4.7. On the length of compounds

Sowinski (1979: 110; 1998: 67) used ten advertisements each to find out how many compounds of two parts, three parts etc. occurred in the total of 20 texts. Summarizing those results led to the following table (cf. Table 4.14):

Table 4.14

Fitting the 1-displaced hyper-Poisson distribution to lengths of compounds
x = 1: two-part compounds (compounds consisting of two lexemes);
x = 2: three-part compounds; etc.

x	n_x	NP_x
1	192	192.41
2	63	60.66
3	10	12.66
4	1	1.97|
5	1	0.24|
6	1	0.06|
a = 0.6182		X_1^2 = 0.901
b = 1.9607		P = 0.34

Though this is a file originating from a text mix, the result nevertheless shows a good conformance between observation and theory (see the graphical diagram). This is, perhaps, caused by the fact that one does not take all words into account but only special ones (compounds).

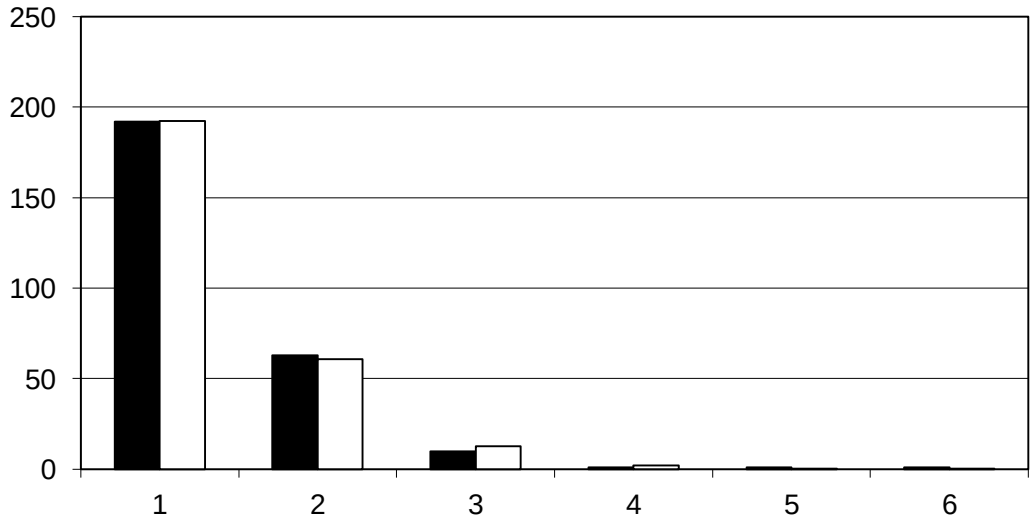


Figure 4.14. Concerning the data in Table 4.14

For further results concerning 21 German press texts abiding by Waring distribution see Poppe (2007) .

4.8. Summary

Word lengths were analyzed in texts as well as dictionaries and defined in different ways. In most cases the number of syllables was taken as the relevant criterion, once the number of morphs per word. That means that an immediate constituent (morph, syllable) was the measuring unit for the length of the construct (word). In those cases it was found in the text and in the lexicon that word lengths were distributed according to the linguistic law and conform to the hypothesis (4.8)

$$P_x = g(x)\, P_{x-1}$$

with different assumptions for $g(x)$. In the meantime approx. 4,000 texts from approx. 50 languages have been analyzed. Only in a small number of texts the fitting could not be implemented successfully (yet) – for this, see Best (2001a).

Word length distributions could also be modeled successfully with some German text corpora, which could not really be expected with mixed texts (Best 2006i).

Distributions turned out to be a little more complicated, if word length was not determined by the number of immediate constituents, but by that of the indirect ones (letters, sounds, phonemes). First attempts (mainly: Schneemann 2001) show that the fitting of distributions is not that easily successful and fails a little more frequently. It is sure that the omitted linguistic level of immediate constituents is a disturbing factor. However, it has to be considered that experience with such cases is rather restricted.

In case of languages like Chinese the number of ideographs or the number of strokes or components from which the ideographs are formed can be used as a

criterion for the measuring of word length (Best & Zhu 2001; Dictionaries: Bohn 1998). However, it cannot be said whether specific problems can occur, because so far just a few files have been processed that way.

The subject of the following overviews is to find evidence of the applicability of the above mentioned hypothesis for other entities in the German language as well. There is no reason to assume that other entities or properties of entities should not meet this hypothesis. To achieve that files found in literature are used, in several cases, however, new analyses were made. The definition of the relevant entities is only dealt with if it cannot be considered generally known or can be read about in the literature mentioned.

5. Application of the Theory to other Linguistic Entities

5.1. Morph length (number of phonemes per morph)

Since sounds, phonemes and letters are the immediate constituents of the morpheme, morph and syllable, we show some fitting in the sequel. The definition of morph is not simple. In general, the zero-morph is not taken into consideration (but it can) and letters should not be used because not all languages have an alphabetic script, hence the results concerning letters cannot be generalized. A German example is displayed in Table 5.1:

Table 5.1
Fitting the 1-displaced hyper-Poisson distribution to the morph lengths in text 1: *Hühner, Adler und Mäuse* (Pestalozzi, *Fabeln*, 48)

x	n_x	NP_x
1	70	69.19
2	91	89.95
3	62	59.55
4	20	26.44
5	10	8.83
6	1	2.36
7	3	0.68
a = 1.3488		X_3^2 = 2.168
b = 1.0375		P = 0.54

x: number of phonemes per morph

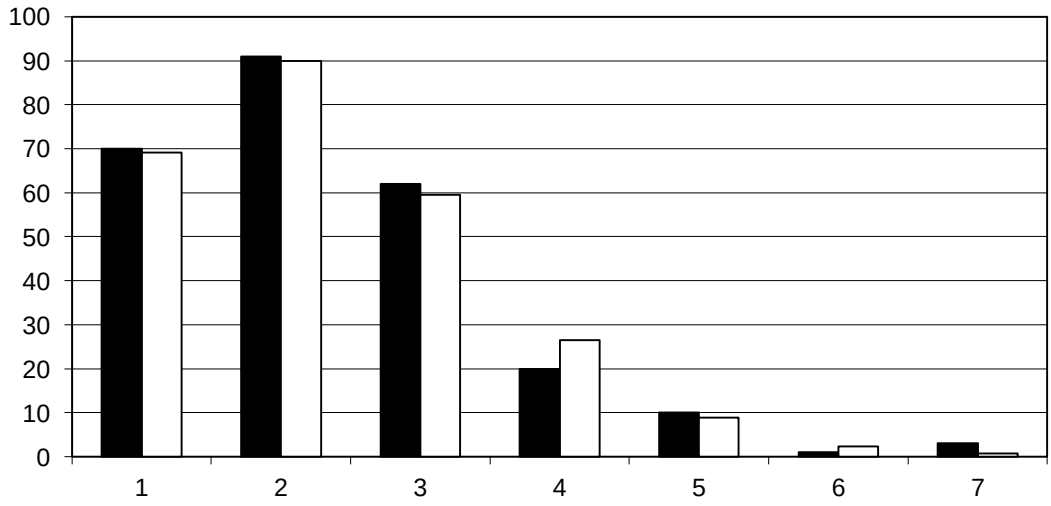


Figure 5.1. Concerning the data in Table 5.1.

So far, morph lengths have been analyzed only rarely. For further results concerning German see Best (2001d); for morph lengths in Lakota see Pustet & Altmann (2005) and in Spanish Saporta (1963: 69). An overview of research works can be found in Best (2005m).

5.2. Syllable length (number of phonemes per syllable)

Table 5.2
Fitting the 1-displaced Conway-Maxwell-Poisson distribution to the syllable lengths in text 1: Pestalozzi, Hühner, Adler und Mäuse (Pestalozzi, *Fabeln*, 48)

x	n_x	NP_x
1	10	9.42
2	95	101.94
3	104	94.25
4	13	20.67|
5	6	1.72|
a = 10.8116		X_1^2 = 2.024
b = 3.5477		P = 0.15

x: number of phonemes per syllable.

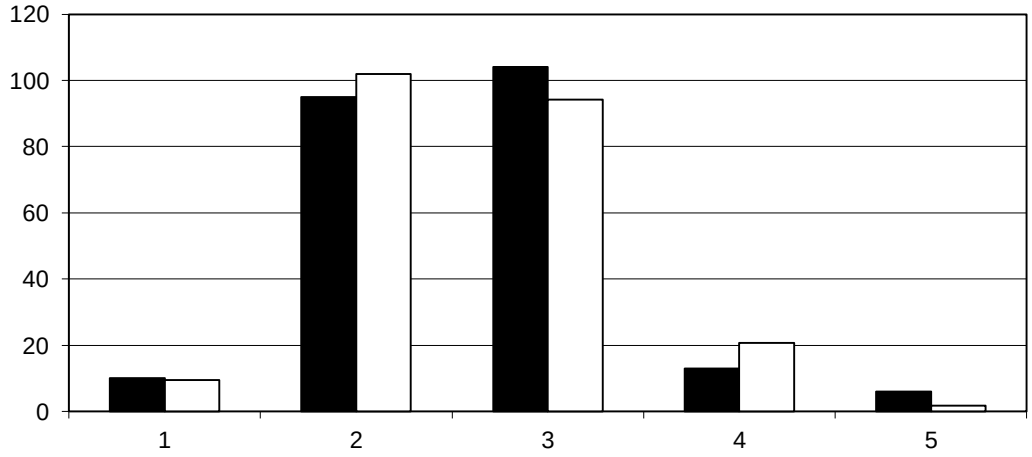


Figure 5.2. Concerning the data in Table 5.2

Astonishingly enough, the hyper-Poisson distribution has proved appropriate in connection with distributions of word and morph lengths, but it is not especially suitable for syllable lengths. Much better results are obtained by using the Conway-Maxwell-Poisson distribution (Best 2001g; Cassier 2001). When analyzing 60 texts in the German music press, positive results were obtained with 45 of them. In further twelve cases the results were only just acceptable; in three cases there was no acceptability (Schneemann 2001).

5.3. Length of constituents (number of words per sentence constituent)

Table 5.3

Fitting the 1-displaced hyper-Poisson distribution applied to lengths of constituents in Pestalozzi, *Hühner, Adler und Mäuse* (Pestalozzi, *Fabeln*, 48)

x	n_x	NP_x
1	37	35.36
2	11	16.12
3	7	7.35
4	4	3.35
5	3	1.53
6	2	0.70
7	1	0.58
a = 452252.4062		X_3^2 = 5.560
b = 991636.0098		P = 0.14

x: number of words per constituent.

The sentence constituents are determined in compliance with Bünting & Bergenholtz (1989).

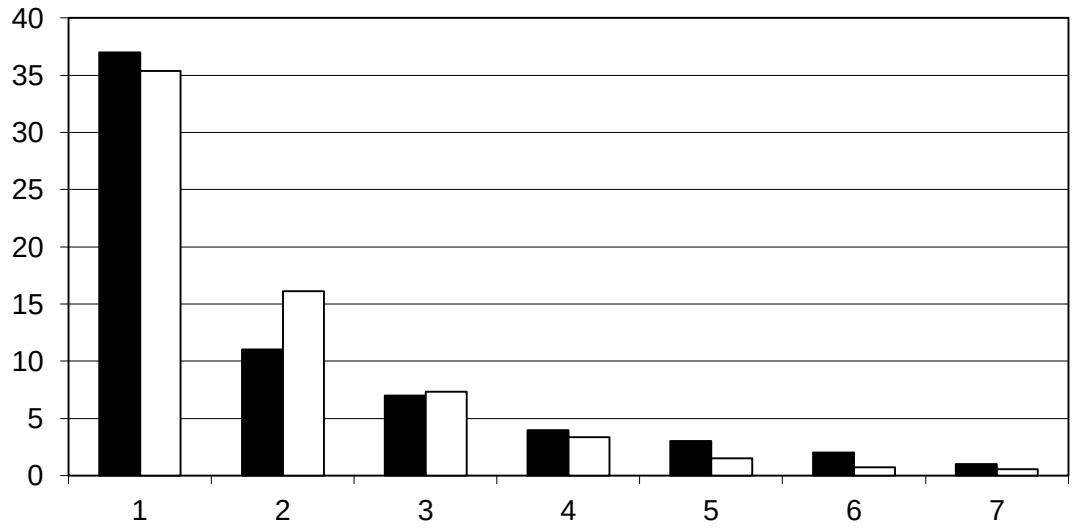


Figure 5.3. Concerning the data in Table 5.3.

With this fitting extremely high parameter values attract attention. In such a case the hyper-Poisson distribution converges to the geometric distribution (Wimmer & Altmann 1999a: 282). If then the geometric distribution is fitted to the file, the value P = 0.57 is a distinctly better result.

5.4. Constituent-related attributes

The "length of constituents" is a property that can be determined especially easily with respect to the different complexity of constituents, which, however, can also be measured in a different way. There is reason to believe that the distribution laws do not only apply to the different lengths of constituents and other entities. The constituent-related attributes are presented as an example. The criterion for that phenomenon here is the number of attributes in those constituents, and this without respect to their interrelationship.

Table 5.4
Fitting the 1-displaced hyper-Poisson distribution to Pestalozzi, *Hühner, Adler und Mäuse* (Pestalozzi, *Fabeln*, 48)

x	n_x	NP_x
1	46	46.00
2	14	14.17
3	4	3.75
4	1	1.09
a = 1.8707		X_1^2 = 0.0261
b = 6.0750		P = 0.87

x stands for the different degrees of complexity. x = 1 stands for constituents without attributes. x = 2 stands for constituents with one attribute, x = 3 are those with two attributes etc.

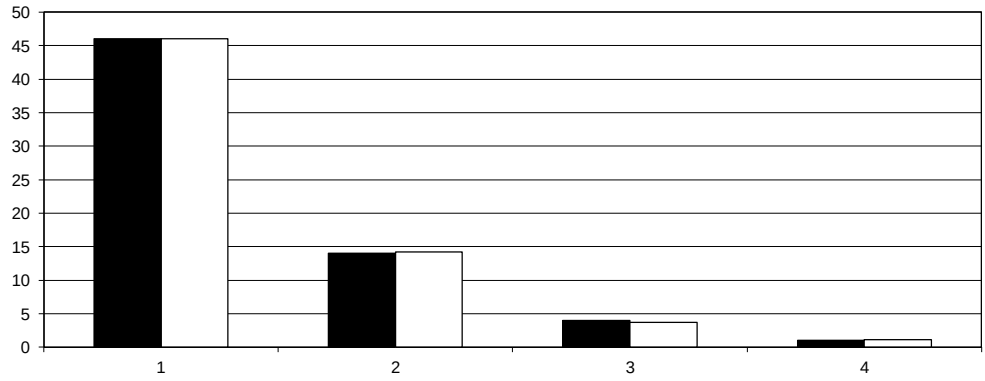


Figure 5.4. Concerning the data in Table 5.4.

5.5. Sentence lengths

Sentence lengths can be determined in different ways. If they are measured on the basis of the number of words, the analyst will find sentences of very different lengths (König [15]2005: 116; Meier 1967: 186). Meier (1967: 192) found especially long sentences in H. Broch's *Der Tod des Vergil* whose longest sentence com-

prises 1077 words. If one considers phonemes, one attains a fractal. Again, the measurement in terms of non-immediate constituents, e.g. words, yields many problems that can – perhaps – be solved by taking into account a number of boundary conditions concerning language, text type, style, etc. Here we show only one example.

A text by G. Wohmann is taken as an example of three different evaluations concerning sentence length.

a) Number of words per sentence; each sentence length is analyzed individually:

Table 5.5

Fitting the 1-displaced negative binomial distribution to sentence lengths taken from: Wohmann, *Die Bütows* (In: Best 2001i)

x	n_x	NP_x	x	n_x	NP_x	x	n_x	NP_x
1	0	0.43	12	4	7.38	23	2	0.98
2	1	1.47	13	6	6.59	24	1	0.77
3	3	2.96	14	6	5.78	25	0	0.60
4	5	4.63	15	3	4.98	26	0	0.47
5	9	6.22	16	1	4.22	27	2	0.36
6	6	7.50	17	3	3.54	28	0	0.28
7	8	8.39	18	2	2.93	29	0	0.21
8	14	8.84	19	3	2.39	30	0	0.16
9	10	8.89	20	3	1.94	31	0	0.12
10	9	8.60	21	0	1.56	32	0	0.09
11	9	8.07	22	2	1.24	33	1	0.41
k = 5.0156		p = 0.3305	X^2 = 14.377		FG = 21	P = 0.85		

k, p: parameters of the negative binomial distribution.

This text does not include single-word sentences. The longest sentence comprises 33 words.

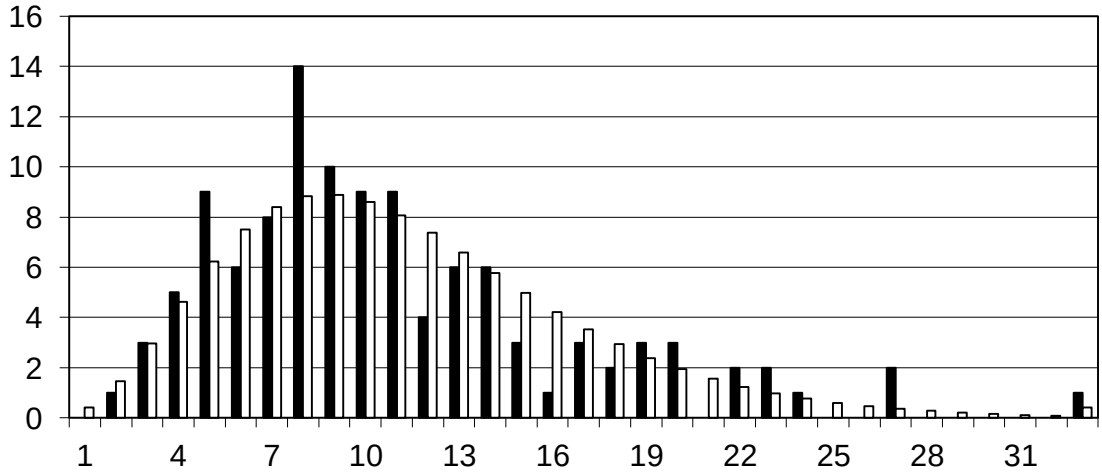


Figure 5.5. Concerning the data in Table 5.5

The negative binomial distribution results from the formula $g(x) = (a+bx)/x$.

The compliance between observation and model is very good. Discrepancies in the graphical chart partly result from the fact that the observations only comprise integer numbers.

b) The number of words per sentence: sentences comprising 1-5, 6-10 etc. words are grouped in class x = 1, 2 etc.:

Table 5.6

Fitting the 1-displaced negative binomial distribution to sentence lengths taken from: Wohmann, *Die Bütows* (In: Best 2001i)

x	n_x	NP_x
1	18	24.38
2	47	37.34
3	28	28.63
4	12	14.67
5	5	5.64
6	2	1.74|
7	1	0.60|
k = 531.8396		X_3^2 = 4.951
p = 0.9971		P = 0.18

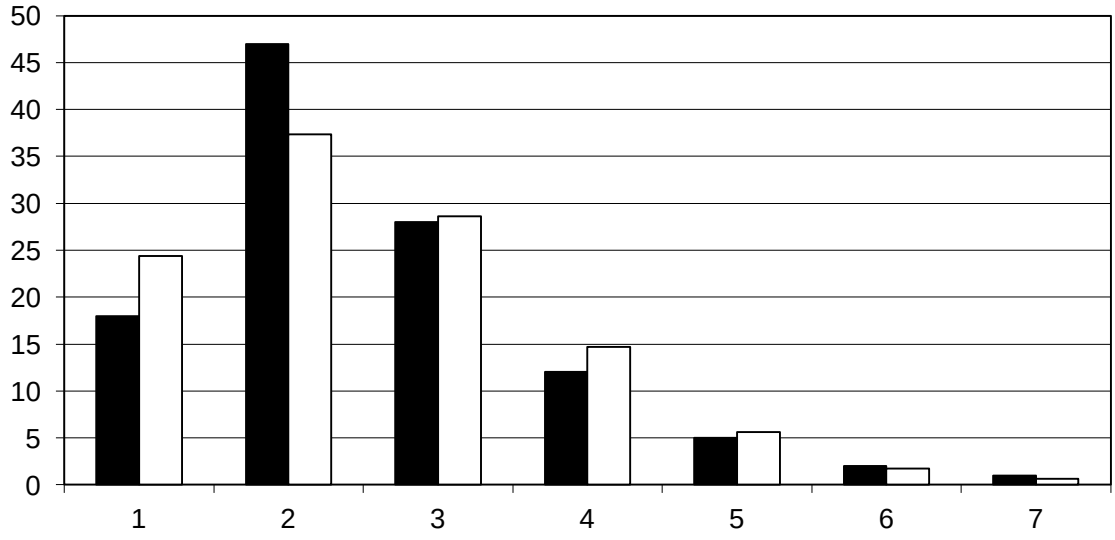


Figure 5.6. Concerning the data in Table 5.6

In a new analysis, Livesey (2001) fitted the hyper-Poisson distribution to 40 German press texts taken from *Spektrum* und *Die Zeit*, whereas the negative binomial distribution proved suitable for 40 English texts taken from the *Guardian* and *Time*. Most analyses dealing with sentence lengths and determining the latter by the number of words per sentence group the sentence lengths in different ways: Meier (1967: 186) forms the first length class by means of one to three

words per sentence, the second from four to six per sentence etc., but also points out that other groups affect the forms of different curves or circle diagrams in graphical depictions. Groups of one to five, six to ten etc. words are more usual (Altmann 1988a: 157; Kelih & Grzybek 2004).

Since we skipped a level - because the word is not the immediate constituent of sentence -, one may expect that exceptions form a non-smooth course of frequencies. Here, smoothing may be attained either by pooling several classes – as has been done above – or by considering only very long texts.

c) Number of clauses per sentence

Clause is the immediate constituent of the sentence, hence better results can be expected. We present an example in Table 5.7 and Figure 5.7. The fitting of the distribution to observed data proved excellent in all three cases. For further analyses of clause length in sentences see Altmann (1988); Niehaus (1997); Roukk (2001); Wittek (2001).

Table 5.7

Fitting the 1-displaced negative binomial distribution to sentence lengths taken from: Wohmann, *Die Bütows* (In: Best 2001i).

x	n_x	NP_x
1	69	69.13
2	29	28.02
3	9	10.27
4	4	3.63
5	1	1.26
6	1	0.69
k = 1.2355		X_2^2 = 0.233
p = 0.6719		P = 0.89

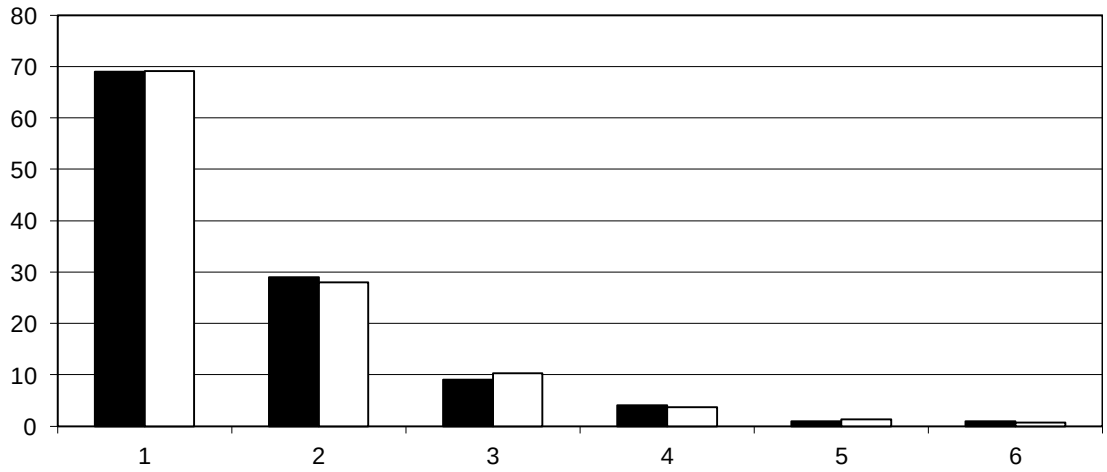


Figure 5.7. Concerning the data in Table 5.7

The number of clauses per sentence is determined by the number of finite verbs in a sentence (cf. discussion of criteria in Niehaus 1997: 222ff.; Wittek 2001: 220ff.)

Sentence lengths can also be determined in a different way. Fucks (1968: 87) demonstrates sentence length distributions obtained by the counting of syllables per sentence. Problems that might develop by the different measuring units are dealt with by Altmann (1988a: 149, 156) and Popescu, Best, Altmann (2014); for an overview of the modeling of sentence lengths see Best (2005n).

Jing (2001) analyzed Chinese texts on the basis of the number of ideographs per sentence and there are works measuring sentence length in terms of letters. It must be remarked that skipping some levels between construct and component is analogous to measuring of the size of earth in terms of the number of atoms it contains.

5.6. Length of rhythmic units

Prose rhythms were first dealt with by Marbe (1904); see (Best 2001h). Several of his disciples and his colleague Albert Thumb (Best 2006f) also dealt with that subject concerning different aspects. Rhythmic units are passages in which two stressed syllables follow one another directly ($x = 1$) or two stressed syllables are separated by one ($x = 2$), two ($x = 3$) or more unstressed ones ($x = 4, 5,...$). Values were compiled without taking sentence boundaries into consideration. In contrast to analyses by Marbe and his staff who mainly analyzed arbitrary text passages a complete text is dealt with here (for further examples including the evaluation of complete texts see Best 2001f, 2002; Kaßel 2002).

Table 5.8

Fitting the 1-displaced hyper-Poisson distribution to the lengths of rhythmic units in: Pestalozzi, Hühner, Adler und Mäuse (Pestalozzi, *Fabeln*, 48)

x	n_x	NP_x
1	1	3.40
2	28	21.52
3	19	24.16
4	14	14.88
5	7	6.32
6	4	2.73
$a = 1.3651$		$X_3^2 = 5.473$
$b = 0.2158$		$P = 0.14$

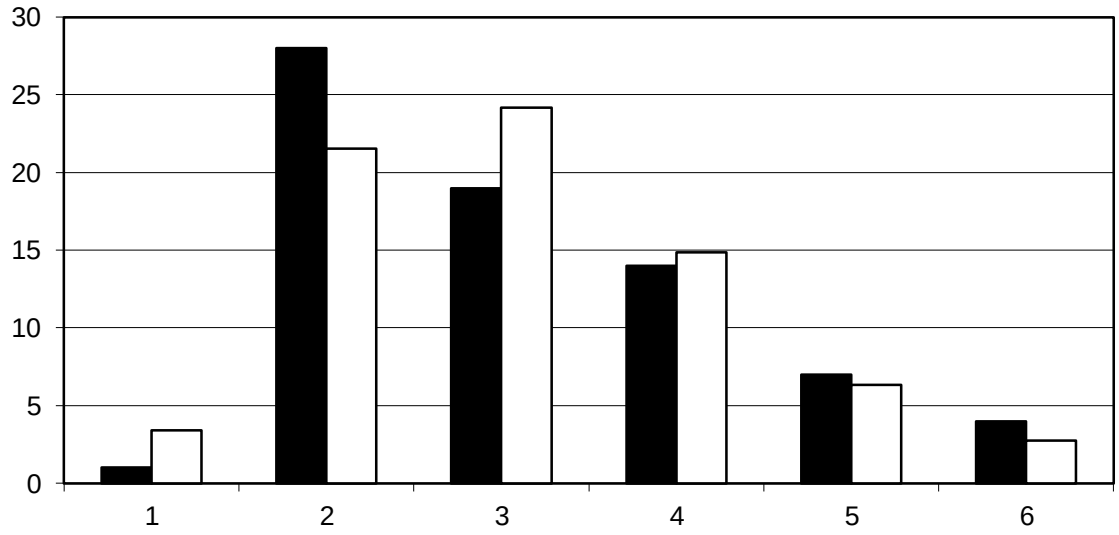


Figure 5.8. Concerning the data in Table 5.8.

The lengths of rhythmic groups (Knauer 1936) also permit the fitting of the hyper-Poisson distribution (Best 2006h). For an overview of research works concerning rhythmic units see Best (2005o). For Russian see Knaus (2008).

5.7. Complexity of ideographs

Languages with complex ideographs like Chinese volunteer to follow up the hypothesis that ideographs composed of a different number of components are used in texts with different frequency. Such an analysis was performed by Yu (2001). The complexity of the ideographs was determined by the number of strokes used to compose them. If then ideographs composed of 1 - 3, 4 – 6 etc. strokes are pooled, the analyst obtains distributions to which – in most cases – the 1-displaced binomial distribution can be fitted. An example is shown in Table 5.9 and Figure 5.9.

Table 5.9

Fitting the 1-displaced binomial distribution to the ideographs in: Binxin, Wangshi [II] (*Erinnerungen* [II]) (Yu 2001)

x	n_x	NP_x
1	36	31.83
2	79	87.75
3	105	96.75
4	47	53.33
5	17	14.70
6	2	1.64
n = 5.0000		X_3^2 = 3.322
p = 0.3654		P = 0.34

x = 1: ideographs composed of 1 to 3 strokes; x = 2: ideographs composed of 4 to 6 strokes etc.

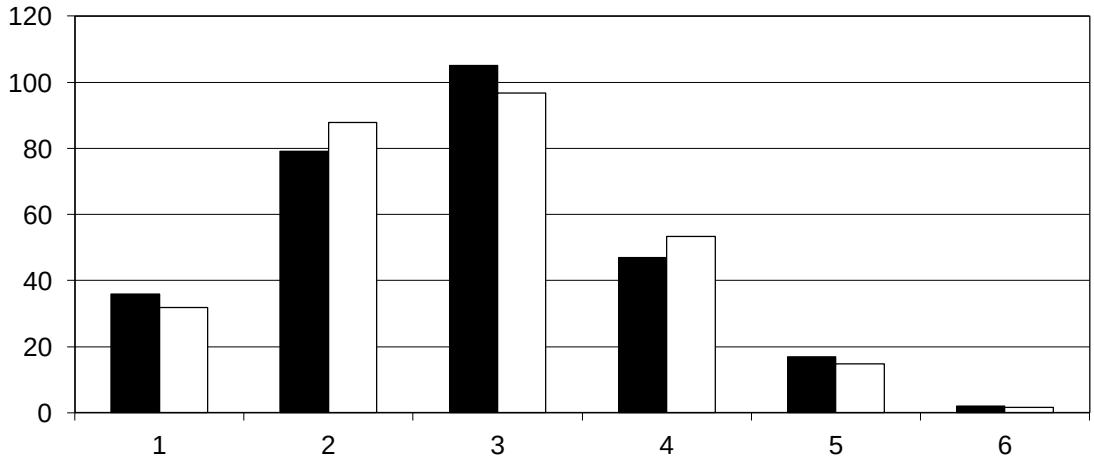


Figure 5.9. Concerning the data in Table 5.9

This case shows a good conformity of observation and theory. Chinese ideographs under the aspect of phoneticity are scrutinized by Bohn (1998: 78ff.). Proposals for the measuring of the complexity of writing in arbitrary systems can be found in Altmann 2004).

5.8. Length of illocution chains

Finally, two files are dealt with, which show that the same linguistic laws can obviously be observed with even larger linguistic entities than assumed so far: illocution chains in talks and dialogues. Therefore, Rothe, Altmann & Wagner (1992: 54) found out in a longer talk among four individuals how long illocution chains were; those chains were expressed by the participants (two adults – two children), before it is the subsequent participant's turn. They could demonstrate that the positive (0 truncated) negative binomial distribution is a good model for this; below, you can find the file for an adult woman (Nora) (cf. Table 5.10 and Fig. 5.10):

Table 5.10

Fitting the positive negative binomial distribution to illocution chains

x	n_x	NP_x
1	366	375.31
2	111	99.66
3	28	32.09
4	12	11.24
5	4	4.13
6	1	1.56|
7	2	0.60|
8	1	0.40|
$r = 0.2205$		$X_3^2 = 2.90$
$p = 0.5649$		$P = 0.41$

$x = 1$: turn containing one speech act only;
$x = 2$: turn containing two speech acts etc.

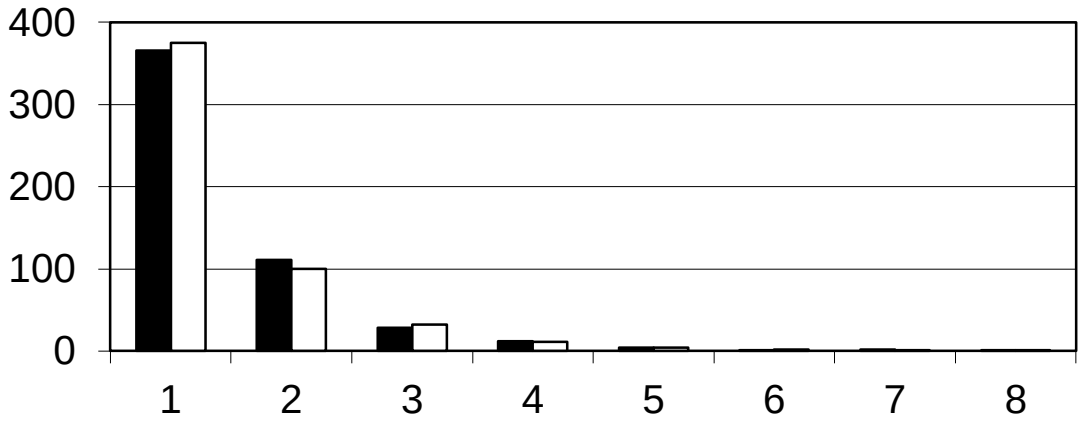


Figure 5.10. Concerning the data in Table 5.10

5.9. Length of SMS dialogues

In an analysis concerning SMS communication, Androutsopoulos & Schmidt (2002: 61) determine the length of news bound in a dialogue; in that case the geometric distribution is a good model. An example is shown in Table 5.11 and Figure 5.11.

Table 5.11
Fitting the displaced geometric distribution to dialogue lengths

x	n_x	NP_x
1	76	71.38
2	40	41.58
3	19	24.23
4	15	14.11
5	8	8.22
6+	13	11.47
p = 0.4174	X_4^2 = 1.751	P = 0.78

x = 1: the SMS dialogue finished after two dialogue turns;
x = 2: the dialogue finished after three turns etc.

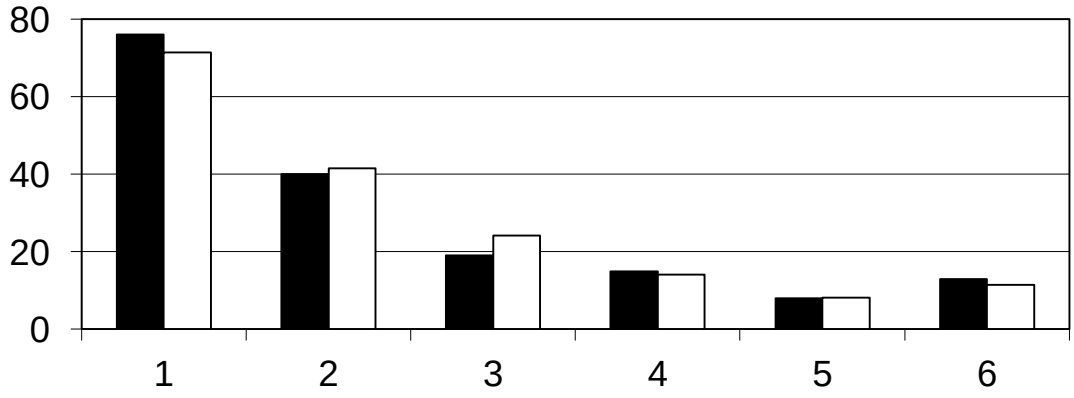


Figure 5.11. Concerning the data in Table 5.11

The empirical basis for the length of linguistic entities like illocution chains and dialogues being longer than a sentence is still very narrow; so, it can happen easily that further analyses require the testing of different models.

A unified model of length distribution of any unit in language has been described in Popescu, Best, Altmann (2014).

6. Application of the Theory on Categorially or Functional-Semantically Determined Classes of Entities

The subsequent three fittings do not deal with classes of entities of different complexity, but with classes of entities with different morphological and syntactical properties. Assumptions that they also obey to specific linguistic laws (Best 1997a, 1998) are allowed.

6.1. Distribution of word classes

Word classes may be defined in various ways. One usually adheres to the Latin parts of speech and tries to transfer this classification system also in other languages. Nevertheless, there are many other possibilities and the task of linguistics is the finding of a general system holding true for all languages. Here we shall restrict ourselves to German and Russian and search for a model for ranked data.

Table 6.1

Fitting the 1-displaced negative hypergeometric distribution to the ranking of word classes in text 1: Pestalozzi, *Hühner, Adler und Mäuse* (Pestalozzi, *Fabeln*, 48) **(***K*, *M* and *n* are the parameters used in this distribution.

Rank	Word class	n_x	NP_x
1	Noun	32	32.49
2	Pron.	24	23.35
3	Verb	19	19.38
4	Art.	17	16.66
5	Prep.	15	14.41
6	Adj.	12	12.32
7	Adv.	9	10.11
8	Conj.	8	7.29
$K = 2.0905$		$n = 7$	$FG = 4$
$M = 0.7532$		$X^2 = 0.262$	$P = 0.99$

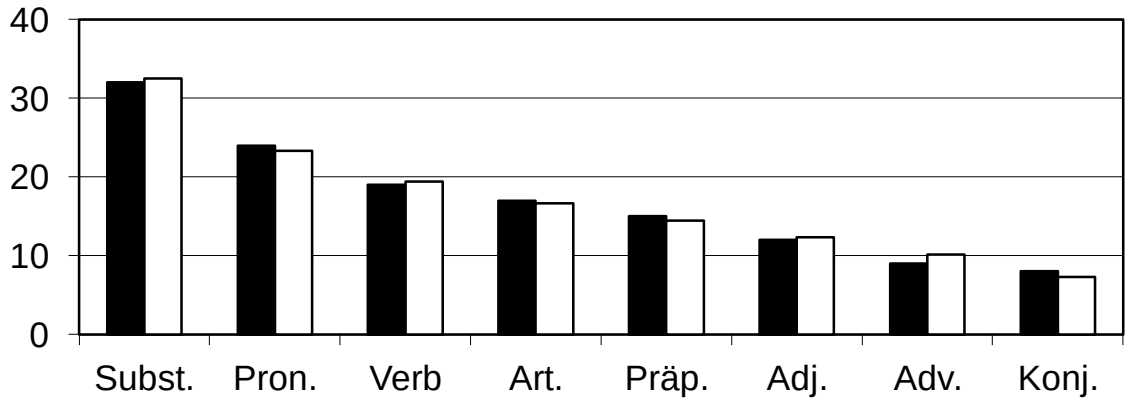


Figure 6.1. Concerning the data in Table 6.1.

In more recent analyses concerning 23 advertisements (Best 2000) and 30 press texts (ten local squibs and ten agency reports in the *Eichsfelder Tageblatt* as well as ten texts in the column "Panorama" in *DER SPIEGEL*) (Best 2001c) the negative hypergeometric distribution has always turned out to be a good model for the ranked distribution of word classes. This model also proved appropriate for this fable by Pestalozzi.

In the analysis of word classes by Hammerl (1990) that can be considered the most comprehensive one so far the right truncated Zipf-Alekseev distribution (also known as "Zipf-Alekseev") was fitted successfully to 50 German and 60 Russian texts. A modified form of that distribution (Wimmer & Altmann 1999: 574f.: "right truncated modified Zipf-Alexeev distribution") could also be used for the fable "Hühner, Adler und Mäuse" by Pestalozzi; the result was very good (P = 0.94). So, that distribution is a model that should always be taken into consideration in connection with word classes. Later on, Popescu, Best and Altmann (2015) showed that the Zipf-Alekseev function may be used to fitting of any type of length in language.

However, it was found that that distribution does not always allow for optimum results. It seems that in many cases the negative hypergeometric distribution is more appropriate. This, however, is a distribution that in contrast to the Zipf-Alexeev distribution can be derived from the same approach as the models already applied to word and sentence lengths. Unfortunately, the boundary conditions present in the data have not been scrutinized as yet.

Now it is completely clear that word lengths (and the lengths of other entities) are very different linguistic phenomena than the word classes. However, conformity can be found: in both cases classes of entities or properties are formed, which is followed by an analysis of the frequency at which those classes are represented in texts. Therefore, it could make sense to apply the general model

$$P_x = g(x)\, P_{x-1} \qquad (6.1)$$

to word classes as well, after the classes, i.e. word classes in this case, have been ranked according to their frequency. This would mean that possibly classes of linguistic entities of any kind are subjected to this one law.

As it was the case with word classes, it can of course not be expected that word classes in texts by arbitrary authors, languages, text types and times always follow the same distribution. However, it could be found that this model is an important form of the distribution law for word classes, as it is e.g. the hyper-Poisson distribution for word lengths and other entities that has proved appropriate again and again, but not always. In such cases it is revelatory to see when a certain distribution can be adapted and when it does not work. Such differences can be reflected when sufficient observations are available. Perhaps, those (boundary) conditions under which a distribution is appropriate or not can be determined generally in the future. However, it seems that there is a specific corre-

lation between the usable distributions and the assignment of the relevant language to a certain type.

For data concerning the occurrence of word classes in a comprehensive corpus of spoken language see König ([15]2005: 116).

The thought that the distribution law could generally be applicable to arbitrary classes of entities is to be illustrated below by means of three further examples:

6.2. Distribution of constituents determined according to their category

Table 6.2
Fitting the 1-displaced hyper-Poisson distribution to the categories of constituents in: Pestalozzi, Hühner, Adler und Mäuse (Pestalozzi, *Fabeln*, 48)

Rank	Constituents (categorial)	n_x	NP_x
1	nominal	28	28.04
2	verbal	19	19.02
3	prepositional	11	10.40
4	adverbial	4	4.76
5	adjectival	3	2.79
$a = 2.8070$, $b = 4.1367$, $X_2^2 = 0.172$, $P = 0.92$			

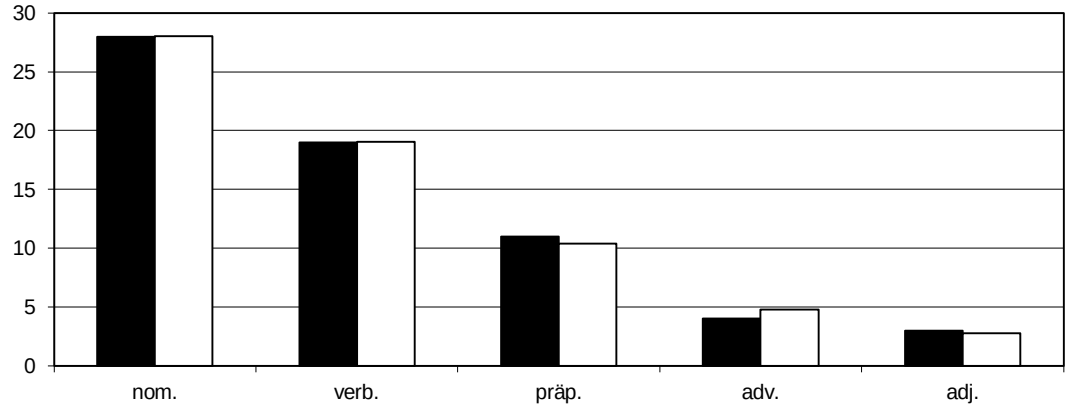


Figure 6.2. Concerning the data in Table 6.2.

The categories of constituents have been taken from Bünting & Bergenholtz (1989: 59ff.). Participial constituents did not occur in that text; if they occur in texts, they could be assigned to the verbal constituents. So can sentence-equivalent infinitives be assigned to the verbal constituents.

6.3. Distribution of constituents determined functionally

Table 6.3
Fitting the 1-displaced hyper-Poisson distribution to the functions of the constituents in: Pestalozzi, *Hühner, Adler und Mäuse* (Pestalozzi, *Fabeln,* 48)

Rank	Constituents (functional)	n_x	NP_x
1	Predicate	19	18.13
2	Adverbial phrase	15	17.31
3	Subject	14	13.02
4	Object	9	8.08
5	Predicative noun	4	4.27
6	Prepositional object	2	1.96
7	Vocative	1	1.23
$a = 3.5513$ $b = 3.7207$ $X_4^2 = 0.588$ $P = 0.96$			

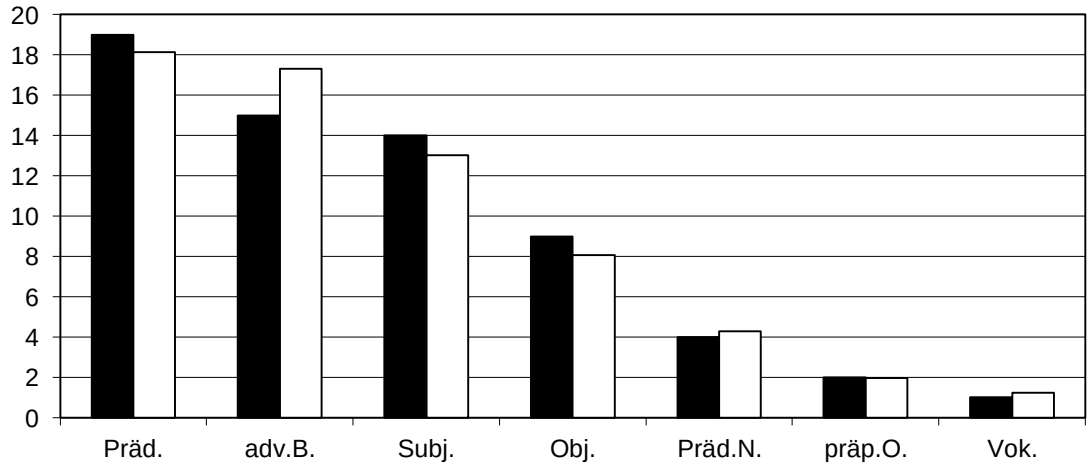


Figure 6.3. Concerning the data in Table 6.3.

The constituent functions have also been adopted from Bünting & Bergenholtz (1989: 67ff.). In addition, the constituent function "Vocative" was taken as a nominal constituent serving as a form of address.

6.4. Distribution of the cases of constituents and the constituent parts

Table 6.4
Fitting the 1-displaced hyper-Poisson distribution to the cases of constituents and constituent parts in:
Pestalozzi, *Hühner, Adler und Mäuse* (Pestalozzi, *Fabeln*, 48)

Rank	Case	n_x	NP_x
1	Nominative	17	16.96
2	Dative	13	13.78
3	Accusative	9	7.93
4	Genitive	5	5.34
a = 1.9718		X_1^2 = 0.211	
b = 2.4273		P = 0.65	

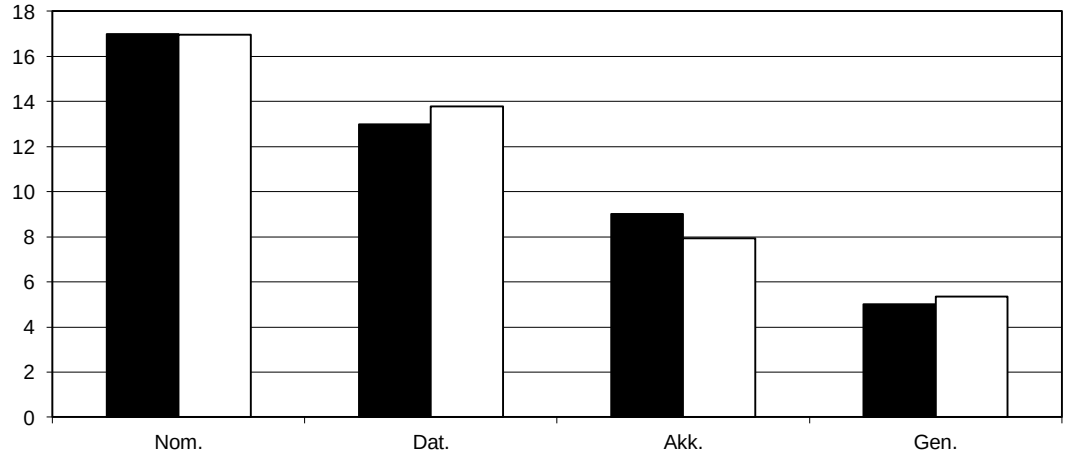


Figure 6.4. Graphical results to Table 6.4.

Such analyses can also be performed in a much more differentiated way. Rothe (1991) discussed a syntactical-semantical differentiation of the cases and tested her hypothesis using the example of the genitive for which she developed more than 50 categories.

König ([15]2005: 118) and Meier (1967: 199ff., 261) supply comprehensive data concerning the use of the cases in German. It is very difficult to generalize this aspect because marked cases are present only in some languages; in strongly analytic languages they are not present. Further, they need not be equal in the languages that have them. Thus it is not possible to compare e.g. German and Hungarian, not even German and English because the latter passes slowly to strong analyticity concerning nouns. Before continuing in this direction, one must define the relevant boundary conditions and, taking them into account, one must modify the difference or differential equations leading to a model.

Excursion II:Working with the Altmann-Fitter (1997): Troubleshooting

After a lot of distributions of classes of entities have been dealt with, some problems are to be dealt with shortly; those problems occur occasionally and can usually be subjected to troubleshooting easily. After 20 texts related to a certain subject have been processed, it can happen that fittings to individual texts do not succeed at the first attempt. The following cases a) to d) may occur occasionally; they are files in which the distribution of word lengths being determined by the number of syllables per word is presented in texts.

a) A text has too few different classes:
When Old High German texts were analyzed (Best 1996) it turned out that the 1-displaced Poisson distribution is a good model in most cases. But in the following text (*Blutspruch*, 13. Jhd.) fitting was not satisfactory at first:

Table 6.5

x	n_x	NP_x
1	22	24.69
2	23	16.92
3	4	7.39
a = 0.6855	X_1^2 = 4.028	P = 0.04

If, however, an additionally empty class is added, a distinctly better and satisfactory fitting can be achieved:

Table 6.6

x	n_x	NP_x
1	22	24.70
2	23	16.91
3	4	5.79
4	0	1.60
a = 0.6849	X_2^2 = 4.621	P = 0.10

(Best 1996: 147). Such an attempt frequently pays, if only a very low number of different length classes were observed, especially if no degrees of freedom could be observed in connection with a specific fitting. One degree of freedom is gained by the addition of an empty class.

b) In contrast to other texts of the same corpus the distribution that is successful otherwise cannot be fitted at all or not satisfactorily.
Apart from the hyper-Poisson distribution the Poisson distribution is a good model for Old High German texts. However, the fitting of the 1-displaced Pois-

son distribution to the Old High German recipe produced an insufficient result only (Best 1996: 148, text 14):

Table 6.7

x	n_x	NP_x
1	35	47.45
2	72	49.96
3	19	26.30
4	7	9.23
5	3	3.06
$a = 1.0529$	$X_3^2 = 15.555$	$P = 0.0014$

The application of the discrepancy coefficient with $C = 0.11$ did not produce an acceptable result either. Instead, the fitting of the positive Cohen-Poisson distribution, a modified Poisson distribution, produced a good result:

Table 6.8

x	n_x	NP_x
1	35	34.72
2	72	67.93
3	19	24.77
4	7	6.78
5	3	1.80
$a = 1.0940$		$X_2^2 = 2.396$
$\alpha = 0.7205$		$P = 0.30$

(as to this distribution, cf. Wimmer & Altmann 1996: 119). It may be caused by the fact that this text is a recipe in contrast to the other texts (effect of the different text type).

c) Occasionally, really unusual texts like the following Finnish poem can be observed:

Table 6.9

x	n_x
1	0
2	45
3	0
4	12
5	1

(L.Onerva, *Tulit tuska* [You came, pain]; Vettermann & Best 1997: 260.)
In that case meanings of the independent variable are as follows: $x = 1$: one-syllable words; $x = 2$: two-syllable words etc.

It turned out that no distribution could be adapted to that file. However, the analyzer gets the impression that the generally observed tendency, i.e. that longer words are the rarer ones, applies to this poem as well. In that case – and that is similar to the analyses of sentence lengths and Chinese ideographs – length classes can be grouped. This results in the following file to which the 1-displaced Poisson distribution can be fitted very well.

The meanings of the independent variable are as follows: x = 1: words consisting of 1 or 2 syllables; x = 2: words consisting of 3 or 4 syllables etc.

Table 6.10

x	n_x	NP_x
1	45	45.38
2	12	11.14
3	1	1.49
a = 0.2455	X_1^2 = 0.229	P = 0.63

A similar phenomenon that lengths of words did not decrease consistently, but increase and decrease in turn, could be observed with Chinese texts; the problem was also solved by the grouping of adjacent word lengths (Best & Zhu 2001: 107).

d) Many irregularities can be observed with the higher classes comprising the more complex and therefore rarer entities; they can cause problems with the fitting of distributions (see press text in Best 1997b: 11). So, the application of the positive negative binomial distribution to this file does not turn out to be very successful:

Table 6.11

x	n_x	NP_x
1	27	23.21
2	15	16.15
3	9	10.45
4	4	6.51
5	9	3.96
6	0	2.37
7	1	1.40
8	1	1.95
k = 1.5329		X_5^2 = 11.236
p = 0.4505		P = 0.047

That fitting did not fail completely, but is not satisfactory either. The problem seems to be caused by the longer word lengths. Such variations as they are produced by the five-syllable words can be remedied by grouping the relevant length classes. For that reason the analyzer chooses in the menu of the Altmann-Fitter

(1997) "Options" ➔ "Optimization criteria" ➔ "Minimal class size" and there he/she changes from 1.0000 to a higher value, e.g. 5.0000; the effect achieved can be seen by the decreasing number of degrees of freedom. The result obtained is often definitely better, which is also the case here:

Table 6.12

x	n_x	NP_x
1	27	23.99
2	15	16.49
3	9	10.44
4	4	6.34
5	9	3.74
6	0	2.17
7	1	1.24
8	1	1.59
k = 1.6128		X_2^2 = 2.156
p = 0.4740		P = 0.34

7. Ord's Criterion

If texts of different text types are analyzed, one question can be of interest: are the texts to one another similar or different? To find out whether the relevant processed texts are similar to one another with respect to the distribution of the property analyzed, all known methods and a modification of Ord's criterion (Ord, 1972: 98f., 133ff.) can be used; this modification is based on the "moments" of the used distributions (Altmann, 1988a: 48ff.; Popescu, Lupea, Tatar, Altmann 2015: 122 ff.). In case of discrete distributions, Ord's criterion assigns them either points, lines or areas in the Euclidean space. The first is the arithmetic mean m_1

$$m_1 = \frac{1}{N}\sum x\, f_x \quad , \tag{7.1}$$

the others are the (central) moments

$$m_r = \frac{1}{N}\sum (x - m_1)^r\, f_x, \quad r \geq 2, \tag{7.2}$$

with m_2 being the variance and m_3 being the skewness or asymmetry of the distribution. They can be the basis for the calculation of two values, I and S, as $I = m_2/m_1$ and $S = m_3/m_2$. This is to be demonstrated by the example of Old Icelandic texts (Best 1996a). The bases for the analyses were 13 songs in the *Edda*, the 20 chapters of the *Hrafnkels saga freysgoða* as well as 2 chapters from the *Heimskringla*. The latter will not be taken into consideration for the moment. Now, two questions have to be answered: 1. can the songs of the *Edda* and the chapters taken from the *Hrafnkels saga freysgoða* - seen individually with respect to the word length distributions – be considered homogeneous text groups? 2. Do the two text types originating from one single language form a homogeneous text type? For this, see the following table (cf. Table 7.1):

Table 7.1
Ord's criterion in Old Icelandic

Text	m_1	m_2	m_3	I	S
1	1.5616	0.4131	0.2607	0.2645	0.6311
2	1.5627	0.3280	0.0924	0.2099	0.2817
3	1.5876	0.3625	0.1062	0.2283	0.2930
4	1.5993	0.4030	0.2262	0.2520	0.5613
5	1.6269	0.3725	0.1508	0.2290	0.4048
6	1.5401	0.4087	0.2773	0.2654	0.6785
7	1.5411	0.3689	0.1620	0.2394	0.4391
8	1.6634	0.4591	0.2138	0.2760	0.4657
9	1.5163	0.3506	0.1382	0.2312	0.3942

10	1.5093	0.3758	0.2632	0.2490	0.7004
11	1.6384	0.4692	0.3038	0.2864	0.6475
12	1.5946	0.5654	0.6732	0.3546	1.1907
13	1.5458	0.3969	0.2169	0.2568	0.5465
14	1.5590	0.6055	0.8459	0.3884	1.3970
15	1.6489	0.7810	1.1956	0.4736	1.5309
16	1.5152	0.6376	1.0293	0.4208	1.6143
17	1.5096	0.3887	0.2568	0.2575	0.6607
18	1.4823	0.4790	0.4865	0.3231	1.0157
19	1.5195	0.4449	0.4124	0.2928	0.9269
20	1.4804	0.3890	0.2930	0.2628	0.7532
21	1.5268	0.5439	0.6326	0.3562	1.1631
22	1.5205	0.4887	0.5511	0.3214	1.1277
23	1.4545	0.3915	0.3435	0.2692	0.8774
24	1.5163	0.4574	0.4001	0.3017	0.8747
25	1.5448	0.4719	0.3507	0.3055	0.7432
26	1.5073	0.4301	0.3580	0.2853	0.8324
27	1.6139	0.5539	0.4913	0.3432	0.8870
28	1.5332	0.4876	0.4929	0.3180	1.0109
29	1.5598	0.5541	0.5563	0.3552	1.0040
30	1.5866	0.6143	0.6898	0.3872	1.1229
31	1.5149	0.4546	0.4101	0.3001	0.9021
32	1.4706	0.3997	0.3244	0.2718	0.8116
33	1.5342	0.5222	0.6233	0.3404	1.1936
34	1.5532	0.5770	0.5753	0.3715	0.9971
35	1.5368	0.5083	0.5109	0.3308	1.0051

(Texts 1 to 13: songs from the *Edda*, texts 14 to 33: chapters from *Hrafnkels saga freysgoða*, texts 34 & 35: 2 chapters from the *Heimskringla*.)

The values *I* and *S* can then be entered in a coordinate system <*I*,*S*> and used for the visualization of the homogeneity of the text groups, as it is shown in the diagram below (cf. Figure 7.1):

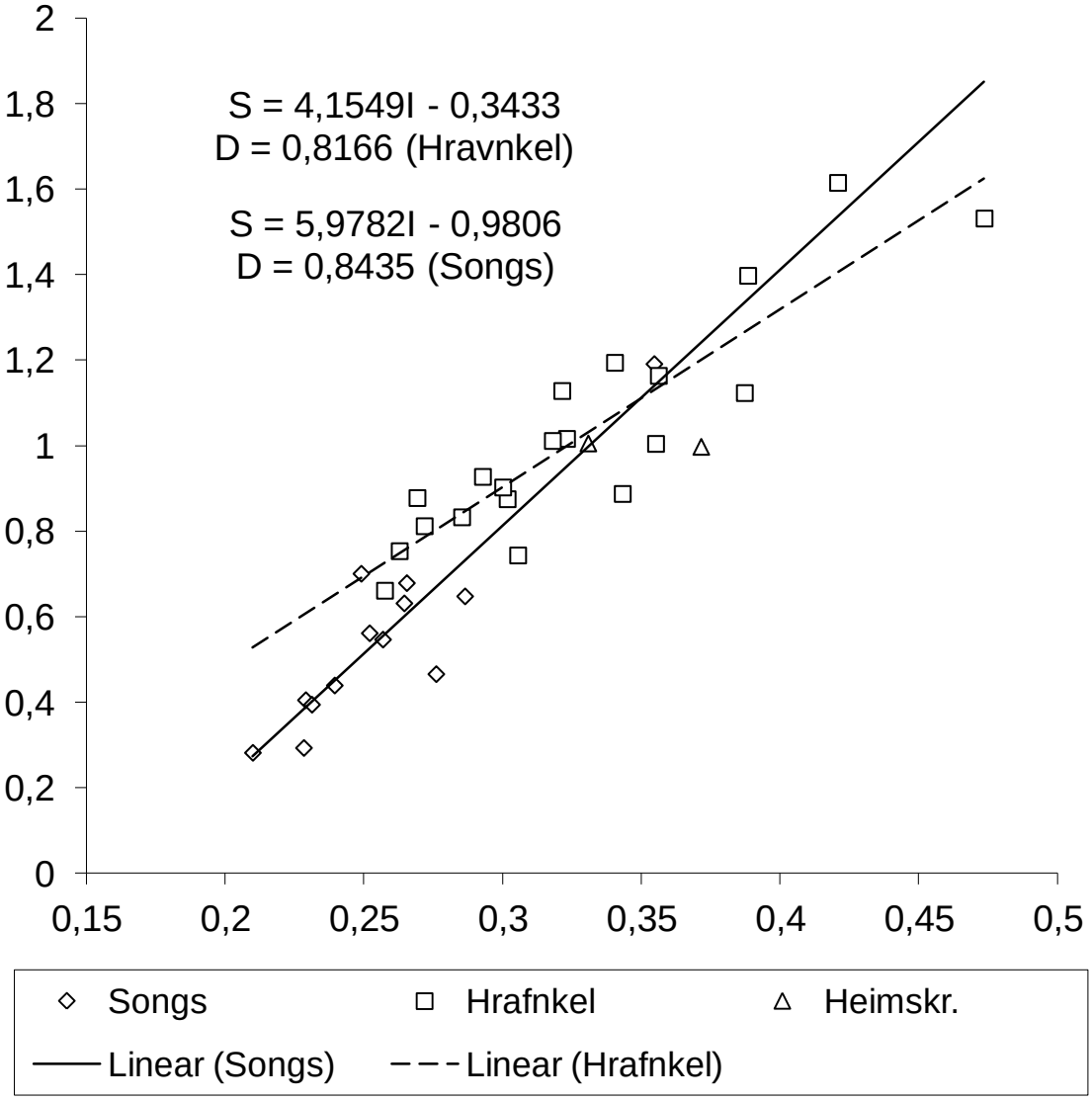


Figure 7.1. Ord's criterion in Old Icelandic, differentiated acc. to text types

The two chapters from the *Heimskringla* have not been taken into consideration for the calculations. Those calculations can be performed by means of the Office package issued by Microsoft, which can be expected to be used by many readers; special software is not required.

The optical impression of this diagram already shows that the texts from the *Edda* as well as those from the *Hrafnkels saga freysgoða* – seen individually with respect to their word length distributions – are somewhat homogeneous. This is confirmed by the linear regressions with their determination coefficients $D = 0.82$ or $D = 0.84$, which can be considered acceptable by $D \geq 0.80$.

If it is analyzed how strongly the two text groups differ from one another, it is the two very acute angles of the trendlines already that show that they are very similar. The regression for all analyzed Old Icelandic texts together including the two chapters from the *Heimskringla* turns out to be rather good:

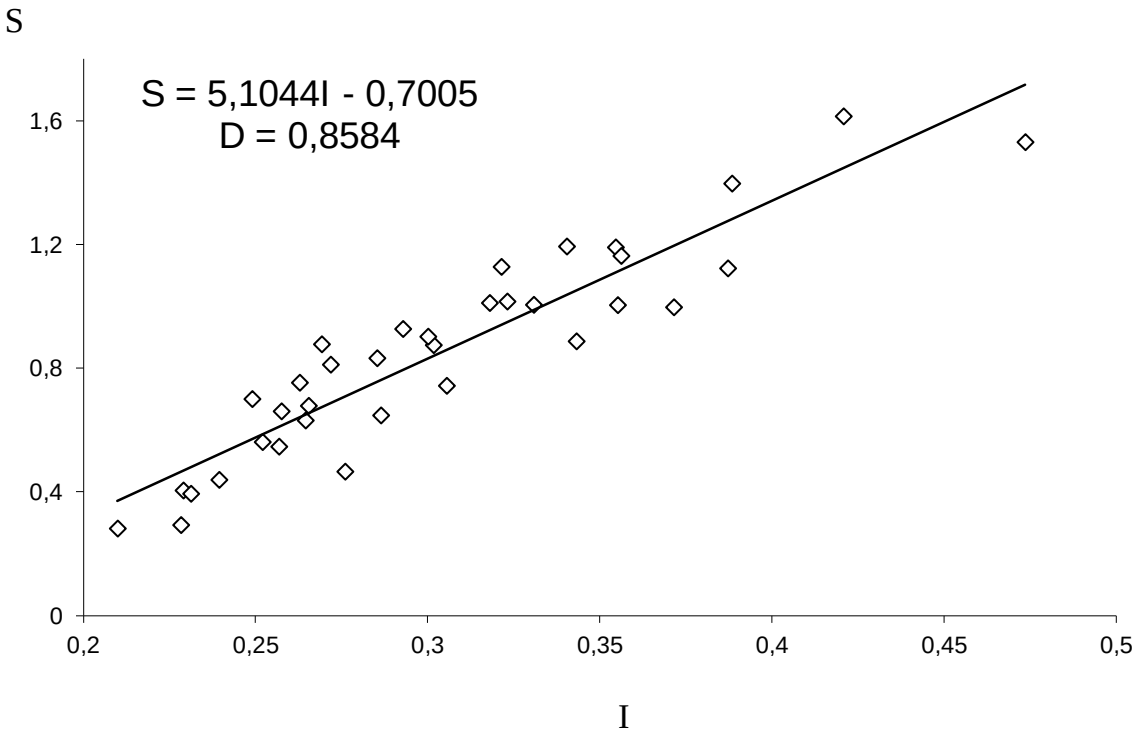


Figure 7.2. Ord's criterion in Old Icelandic (all texts)

One may automatically ask whether there are different areas for texts of various types in other languages. The same question may be asked concerning the lengths of different text units beginning with morphs and syllables up to sentences. Since we have to do with moments, the variances of I and S can easily be derived and at least a normal tests can be applied for goups of texts. This is, of course, a task for the future.

8. Relationships between the Parameters of Distributions

A further aspect of the analyses is the examination of the interrelationships among the parameters of the distributions. In the Old Icelandic texts analyzed in Best (1996a) the 1-displaced hyper-Poisson distribution could be fitted in all cases. There are narrow interrelations among the parameters of this distribution, as can be demonstrated as follows:

Table 8.1
The parameters of the hyper-Poisson distribution in Old Icelandic texts

Text	*a*	*b*	Text	*a*	*b*
1	0.1521	0.1802	19	0.2444	0.3724
2	0.0756	0.0737	20	0.1864	0.2961
3	0.1362	0.1371	21	0.6081	1.1909
4	0.1228	0.1211	22	0.2168	0.3370
5	0.1022	0.0858	23	0.1587	0.2744
6	0.2144	0.2911	24	0.2333	0.3672
7	0.1378	0.1658	25	0.3347	0.5249
8	0.2429	0.2382	26	0.1828	0.2764
9	0.1337	0.1718	27	0.4573	0.6363
10	0.0888	0.1129	28	0.2260	0.3406
11	0.1998	0.2032	29	0.6067	1.0719
12	0.3742	0.4932	30	0.6986	1.2526
13	0.1751	0.2220	31	0.2808	0.4552
14	0.2727	0.4286	32	0.1744	0.2910
15	0.4597	0.7011	33	0.3870	0.6504
16	0.3354	0.6515	34	0.7499	1.4472
17	0.1849	0.2579	35	0.3912	0.6563
18	0.4447	0.9118			

The narrow linear interrelationship among the parameters is expressed by $b = 1{,}9788a - 0{,}1109$, $D = 0.97$ and elucidated in the subsequent graphical diagram:

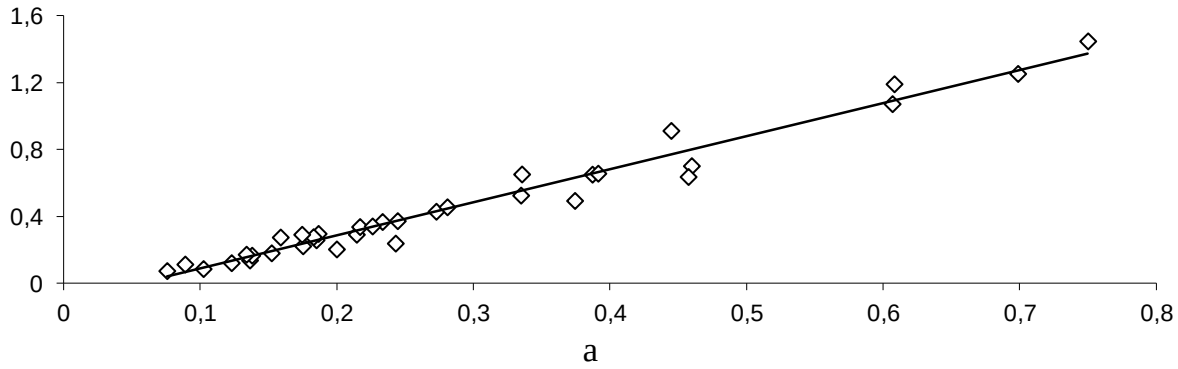


Figure 8.1. Concerning the data in Table 8.1: Relationship between the parameters *a* and *b* in all Old Icelandic texts

9. Discovery of Linguistic Laws II: Rank-Frequency Distributions

Preliminary remark – In those cases in which the distribution of entities of different lengths or complexity was dealt we always used the theoretical approach developed by Wimmer et al. (1994) and Wimmer & Altmann (1996) and it was also explained, if possible. The subsequent linguistic laws do without such detailed explanations of the theoretical bases as well as the distributions used; the relevant descriptions can be found in the references mentioned. As far as newly occurring distributions are concerned, reference is made to the manual related to the *Altmann-Fitter* (1997) as well as Wimmer & Altmann (1999a).

In this section quite different problems are dealt with: How are specific entities distributed in texts? So far, the subject has always been the distribution of classes of entities or properties of entities. Now the subject will be the individual entities themselves.

In literature the problem to be addressed here shortly is known as one of Zipf's laws in its simple form (Guiter & Arapov (ed.) 1982). In a frequency dictionary words are listed according to their frequency; the most frequent one gets rank 1, the second most frequent one rank 2 etc. The following law applies to those words: $r \cdot f = C$; i.e. the product of rank and frequency is a constant. It is known that this law models medium ranks very well; however, there are deviations with the very frequent words as well as the very rare ones. So, we will analyze here whether the Zipf-Mandelbrot distribution

$$P_x = \frac{(b+x)^{-a}}{F_n}, \qquad x = 1, 2, \ldots, n$$

$$F(n) = \sum_{i=1}^{n} (b+i)^{-a}$$

is a good model for such cases instead.

Applications of the Zipf-Mandelbrot law do not only include rank-frequency relations in a frequency dictionary, but also those in texts. Several examples are meant to illustrate them. At first, the distribution of letters in a text is to be dealt with. The example used is a text from G.Chr. Lichtenberg, *Sudelbücher*.

9. 1 Rank-frequency distribution of letters

Table 9.1
Fitting the Zipf-Mandelbrot distribution to the letters in Lichtenberg, *Sudelbücher*, booklet H, no. 15

Rank	<...>	n_x	NP_x	Rank	<...>	n_x	NP_x
1	e	86	73.29	14	o	14	11.36
2	n	49	60.71	15	w	12	10.20
3	i	45	50.81	16	m	9	9.20
4	s	34	42.93	17	z	8	8.32
5	r	33	36.57	18	k	7	7.54
6	t	32	31.39	19	b	6	6.86
7	d	29	27.12	20	p	5	6.26
8	h	26	23.59	21	v	5	5.72
9	a	25	20.64	22	ß	5	5.25
10	l	21	18.15	23	f	4	4.82
11	g	20	16.04	24	ä	3	4.44
12	u	19	14.24	25	ü	3	4.09
13	c	15	12.69	26	ö	1	3.78
a = 3.2395			n = 26				DF = 22
b = 15.7076			X^2 = 16.4824				P = 0.79

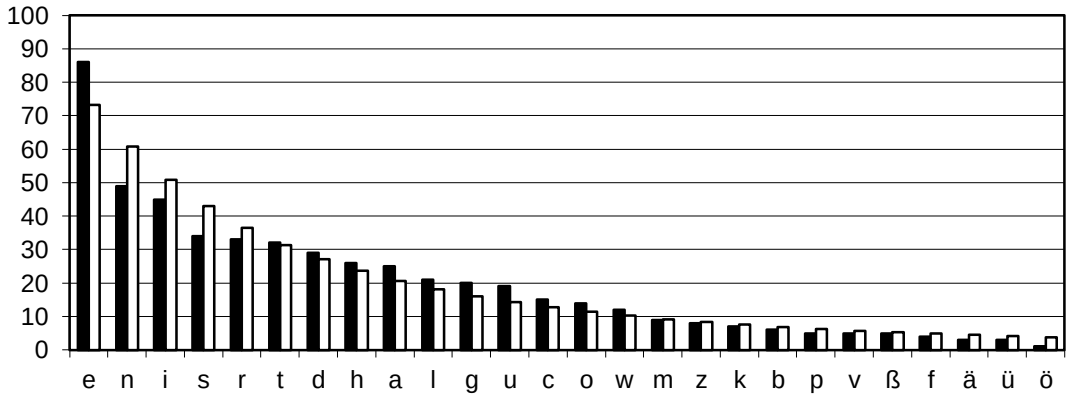


Figure 9.1. Concerning the data in Table 9.1

The test result is very good; the graphical diagram shows that noteworthy deviations between observed and calculated values only occur with the five most frequent letters. All in all, there is a very high degree of conformity between observation and model. However, the result cannot be generalized: this distribution cannot be fitted to the file concerning Pestalozzi's *Hühner, Adler und Mäuse*, but applying the negative hypergeometrical distribution is successful (P = 0.73).

For rank-frequency distributions of letters in German and English cf. Best (2005i,j), Meier (1967: 334) and Nasvytis (1953: 68), in French Nasvytis (1953: 68), in Spanish Meier (1967: 334).

9.2 Rank-Frequency distribution of phonemes

It should be illustrated by two examples. 1) Fitting of the Zipf-Mandelbrot distribution to phonemes in Lichtenberg, *Sudelbücher*, booklet H, no. 15 as shown in Table 9.2.

Table 9.2

Rank	/.../	n_x	NP_x	Rank	/.../	n_x	NP_x
1	e	58	56.31	19	o	8	7.25
2	n	44	45.30	20	ts	8	6.81
3	t	36	37.42	21	m	8	6.41
4	R	32	31.55	22	k	7	6.05
5	i	24	27.04	23	ʃ	7	5.72
6	d	23	23.50	24	o:	6	5.42
7	s	20	20.66	25	h	5	5.14
8	l	20	18.34	26	ŋ	5	4.89
9	a	19	16.41	27	u:	5	4.65
10	f	13	14.80	28	b	5	4.44
11	g	13	13.43	29	a:	4	4.24
12	v	12	12.25	30	oi	2	4.05
13	i:	12	11.24	31	p	2	3.88
14	X	12	10.35	32	au	2	3.72
15	z	11	9.57	33	ü:	2	3.57
16	u	10	8.89	34	ü	1	3.43
17	e:	10	8.28	35	ö:	1	3.29
18	ai	9	7.73				
a = 1.5834			n = 35			DF = 31	
b = 5.7889			X^2 = 10.5946			P = 0.9998	

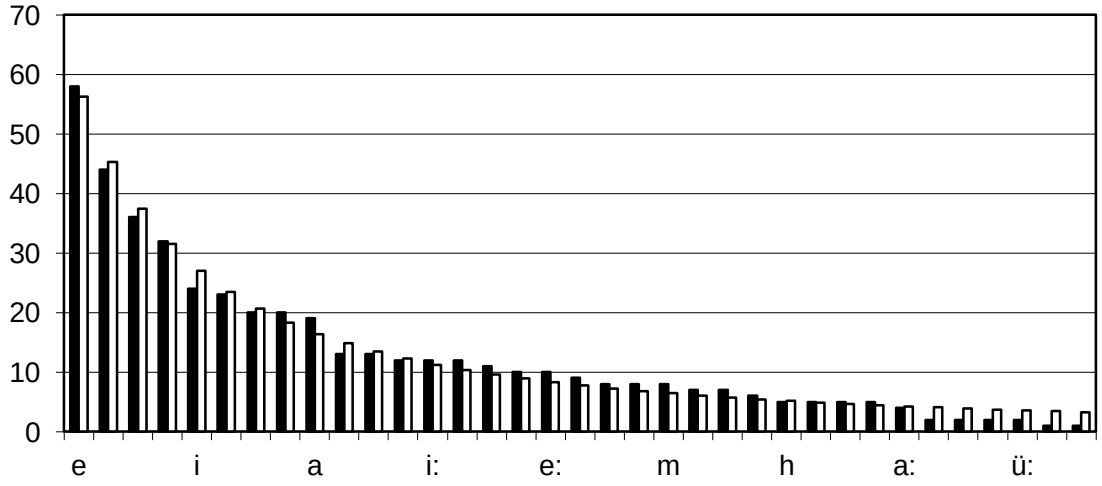


Figure 9.2. Concerning the data in Table 9.2

The phoneme system applied is presented in Best (2001g). The fitting of the Zipf-Mandelbrot distribution for phonemes is considerably better than that for

letters. This distribution can also be fitted with $P = 0.71$ to the file of Pestalozzi's *Hühner, Adler und Mäuse*.

The following file comprises the 7,952 highly frequent word forms in German with 43489083 phonemes; however, here the negative hypergeometric distribution must be fitted.

2) Fitting of the negative hypergeometric distribution to the phonemes in German

Table 9.3

Rank	/.../	n_x	NP_x	Rank	/.../	n_x	NP_x
1	n	4529990	4965451.13	21	ç	855890	755764.58
2	ə	3879296	3483699.49	22	b	731736	702864.14
3	t	3319525	2910077.83	23	a:	589795	652087.99
4	R	3164439	2559123.71	24	k	565780	603290.60
5	ʔ	2588080	2305783.58	25	o	524134	556348.39
6	d	2249715	2106565.31	26	ʃ	475115	511156.55
7	s	2002945	1941549.97	27	h	461768	467626.55
8	i	1687264	1800047.80	28	o:	428442	425684.33
9	a	1427403	1675691.77	29	u:	421855	385268.94
10	l	1348925	1564397.98	30	au	345857	346331.66
11	e:	1196739	1463397.64	31	x	307938	308835.54
12	e	1188053	1370730.95	32	ŋ	300178	272755.41
13	m	1097702	1284962.25	33	p	277713	238078.46
14	i:	1085444	1205009.88	34	ü:	185626	204805.53
15	ai	1034425	1130039.80	35	ü	109062	172953.42
16	f	1027417	1059396.21	36	oi	104951	142558.92
17	z	966011	992554.91	37	j	104731	113685.66
18	g	900992	929090.99	38	ä:	91697	86436.27
19	u	892869	868655.90	39	ö:	84931	60976.10
20	v	886821	810960.85	40	ö	47829	54386.06
$K = 2.9446$		$n = 40$	$M = 0.7230$	$X^2 = 620748.80$			$C = 0.0143$

(acc. to. Kaeding/ Ortmann in: Schulte 1979: 40f. K, M, n: parameters of the distribution; here, we must widely do without captions concerning the axes)

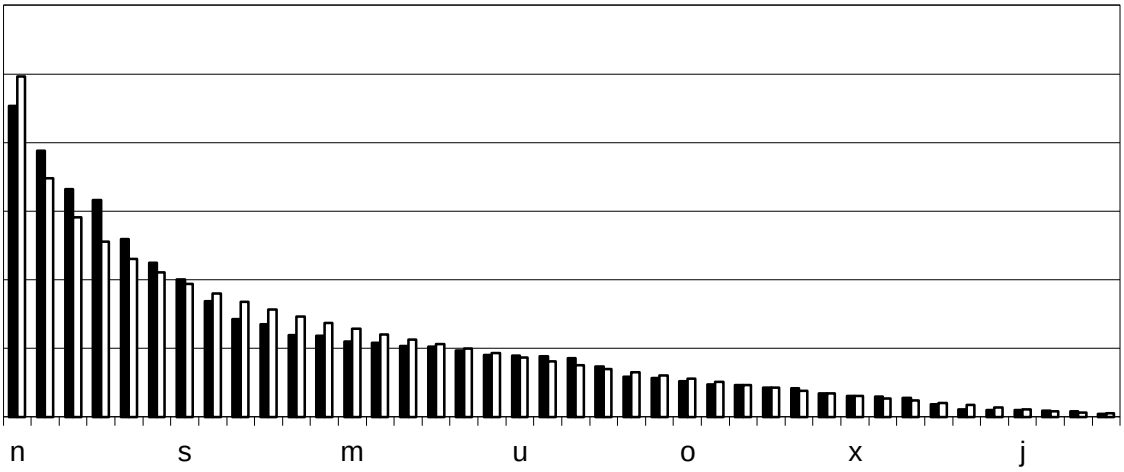


Figure 9.3. Concerning the data in Table 9.3

In contrast to the previous file this depiction presents some allophones as well as the glottal stop as phonemes; however, the affricates are missing. But at this point the problem of phonological evaluation will not be dealt with.

The Zipf-Mandelbrot distribution could not be fitted to those phonemes taken from a very comprehensive corpus. Instead, the possible models obviously are the negative hypergeometric distribution and – with almost the same result – the Polya distribution.

The fitting of the negative hypergeometric distribution that seems to be successful in the graphical diagram has to be deemed weak because of C = 0.0143; however, it need not be considered fully unsuccessful. Probably, that relatively weak result has two causes: Those data originate from one corpus that comprises a mix of a lot of texts; they only deliver the phonemes of the most frequent words. Schulte (1979: 40f.) presents a further phoneme statistics that is based on a small text corpus of just 3,308 word forms comprising 238,543 phonemes from twelve language teaching materials for speech-hearing disabled persons; the fitting of the negative hypergeometric distribution shows a considerably better result with C = 0.0082.

If you test the same distributions by means of the rank-frequency distribution of English sounds (Herdan 1966: 21), results are even a little poorer (negative hypergeometric distribution: C = 0.0165; Polya distribution: C = 0.0175). Fittings do not become better in the distributions in the German and English languages, if the rank-frequency distributions for consonants and vowels are analyzed separately. Altmann & Lehfeldt (1980: 143ff.) apply the geometric distribution to the rank-frequency interrelation in Gujarati.

As to the absolute and relative frequencies of sounds cf. Meier's "100,000-sound counting" (1967: 250 ff.) and a test of those data in Best (2004/05) with a plea for Altmann's model for arbitrary rankings (Altmann 1993; cf. p. 90); a comparison of the values in Meyer's counting of the phonemes in the *Erlkönig* can be found with Grotjahn (1979: 182 ff.). For further more recent comprehensive analyses see: Grzybek & Kelih (2003); Grzybek, Kelih & Altmann (2004).

Words in texts can be analyzed in the same way:

9.3 Rank-frequency distribution of word

From the beginning, Zipf's law was applied to words in lexica being sorted acc. to frequency ranks. But it can also be fitted to words of individual texts; this will be demonstrated by the following examples taken from Lichtenberg's *Sudelbücher*; such a short text was chosen intentionally to prevent the tables from becoming too confusing (as to Goethe's *Erlkönig* see Altmann 1988a: 73 as well as Altmann & Altmann 2005: 72f.; for comparable lists concerning texts and corpora in different languages see Baayen 2001: 291ff.)

Table 9.4
Rank-frequency list of words in Lichtenberg, *Sudelbücher*, booklet H, no. 15

Rank	Form of word	Frequency	Rank	Form of word	Frequency
1	der	5	38	haben	1
2	die	4	39	hielt	1
3	und	3	40	hielten	1
4	das	2	41	Letztere	1
5	des	2	42	Leuten	1
6	für	2	43	mehr	1
7	in	2	44	Meinungen	1
8	ist	2	45	Mitte	1
9	Philosophie	2	46	Naturlehre	1
10	von	2	47	Neuern	1
11	zu	2	48	Religion	1
12	Absicht	1	49	sammeln	1
13	Akademie	1	50	schwächsten	1
14	als	1	51	sind	1
15	also	1	52	so	1
16	am	1	53	solche	1
17	andere	1	54	Skeptikers	1
18	Anhängern	1	55	Stoikers	1
19	anzuraten	1	56	strengen	1
20	auf	1	57	um	1
21	ausgegeben	1	58	Ungewißheit	1
22	betrachten	1	59	unser	1
23	Betrachtung	1	60	unsere	1
24	bloß	1	61	verdient	1
25	da	1	62	Verstand	1
26	daß	1	63	wahrscheinlich	1
27	dieses	1	64	was	1
28	eigentlich	1	65	wer	1
29	eine	1	66	werden	1
30	Entdeckungen	1	67	will	1
31	finden	1	68	wir	1
32	gemacht	1	69	wird	1

33	Geschichte	1	70	worden	1
34	gewiß	1	71	Zeit	1
35	gezogen	1	72	Zuverlässigkeit	1
36	Gleichgültigkeit	1	73	zwischen	1
37	größten	1	Σ		90

Table 9.5
Fitting the Zipf-Mandelbrot distribution to the words in Lichtenberg, *Sudelbücher*, booklet H, no. 15.

Rank	n_x	NP_x	Rank	n_x	NP_x
1	5	3.56	38	1	1.02
2	4	3.02	39	1	1.01
3	3	2.68	40	1	1.00
4	2	2.45	41	1	0.99
5	2	2.27	42	1	0.98
6	2	2.12	43	1	0.97
7	2	2.01	44	1	0.96
8	2	1.91	45	1	0.95
9	2	1.83	46	1	0.94
10	2	1.75	47	1	0.93
11	2	1.69	48	1	0.92
12	1	1.63	49	1	0.92
13	1	1.58	50	1	0.91
14	1	1.54	51	1	0.90
15	1	1.49	52	1	0.89
16	1	1.46	53	1	0.89
17	1	1.42	54	1	0.88
18	1	1.39	55	1	0.87
19	1	1.36	56	1	0.87
20	1	1.33	57	1	0.86
21	1	1.30	58	1	0.85
22	1	1.28	59	1	0.85
23	1	1.26	60	1	0.84
24	1	1.23	61	1	0.83
25	1	1.21	62	1	0.83
26	1	1.19	63	1	0.82
27	1	1.18	64	1	0.82
28	1	1.16	65	1	0.81
29	1	1.14	66	1	0.81
30	1	1.13	67	1	0.80
31	1	1.11	68	1	0.80
32	1	1.10	69	1	0.79
33	1	1.08	70	1	0.79
34	1	1.07	71	1	0.78
35	1	1.06	72	1	0.78

36	1	1.04	73	1	1.08
37	1	1.03	Σ	90	90
$a = 0.4312$			$n = 73$		$FG = 52$
$b = 1.1744$			$X^2 = 3.808$		$P \approx 1$

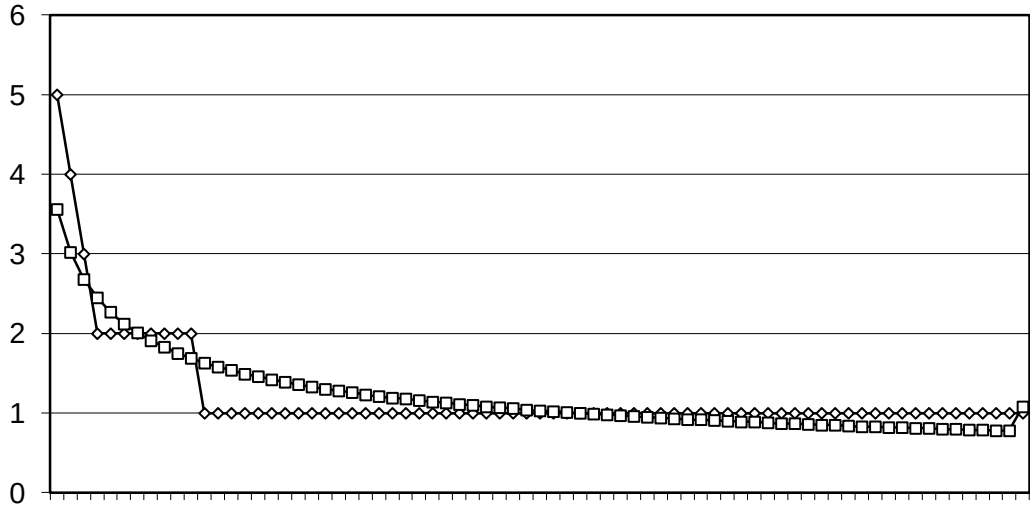


Figure 9.5. Concerning the data in Table 9.5

The deviations between the two lines are based on the alternative that a word either occurs or does not occur; so, just integer values are entered for the observations, whereas the values calculated for the fitting are decimal numbers.

The same good result is obtained by fitting the Zipf-Mandelbrot distribution to Pestalozzi's *Hühner, Adler und Mäuse*.

In his analysis of German texts Billmeier (1969: 45) comes to a really negative appreciation of his results, which probably has two causes: He partly deals with text mixes, partly with very comprehensive works like Kant's *Kritik der reinen Vernunft*, where problems must be expected downright due to missing adequate homogeneity.

The Zipf-Mandelbrot distribution is an especially well studied linguistic law, which can also be used in other different forms of communication (music, art) as well as in economy, biology etc. (Altmann 1988a: 69ff.; Orlov 1982: 118, 142f.; Grzybek 2001). In more recent analyses (Uhlířová 1995; Knüppel 2001) it could be demonstrated that it also applies when just words of one length class of a text is analyzed, e.g. three-syllable words.

10. Discovery of Linguistic Laws III: The Law of Diversification

The object of studies concerning the so-called law of diversification is the formal or functional-semantic differentiation of an entity (Altmann 1985; Rothe [ed.] 1991; Altmann 1996, 2005). The different forms (= diversification at expression level) or functional-semantic validities of an entity appear in compliance with a linguistic law derived and justified by Altmann (1991). In literature examples occur again and again that were taken from completely different contexts and allow validations of the law of diversification. Apart from data especially collected for this section the following remarks include some more of such examples.

As an example of formal diversification the realization of the plural of German nouns can be taken in the form of nine allomorphs (Rothe 1991a: 10). A text was used to determine their frequencies; it was found that the hyper-Poisson distribution could be fitted to the values observed (this text did not include the umlaut as the sole plural marker as it is the case in "Brüder" - German for brothers – and is therefore not mentioned in it). -{0} stands for the zero allomorph as it occurs in "[die] Mädchen-{0}" = German for the girls.)

10.1 Diversification at expression level (formal diversification)

a) Diversification of the allomorphs of the plural of nouns

Table 10.1

Fitting of the 1-displaced hyper-Poisson distribution to the plural of German nouns in: Frank Borsch: Zensiert oder unkontrollierbar? *ai-Journal 5/ 2000, 6-9*

Rank	Plural allomorphs	n_x	NP_x
1	-{en}	33	31.75
2	-{e}	28	29.66
3	-{0}	24	24.50
4	-{n}	20	18.13
5	-{s}	12	12.16
6	umlaut + -{e}	6	7.45
7	umlaut + -{er}	5	4.21
8	-{er}	4	4.14
a = 7.1324	b = 7.6344	FG = 5	X^2 = 0.782	P = 0.98

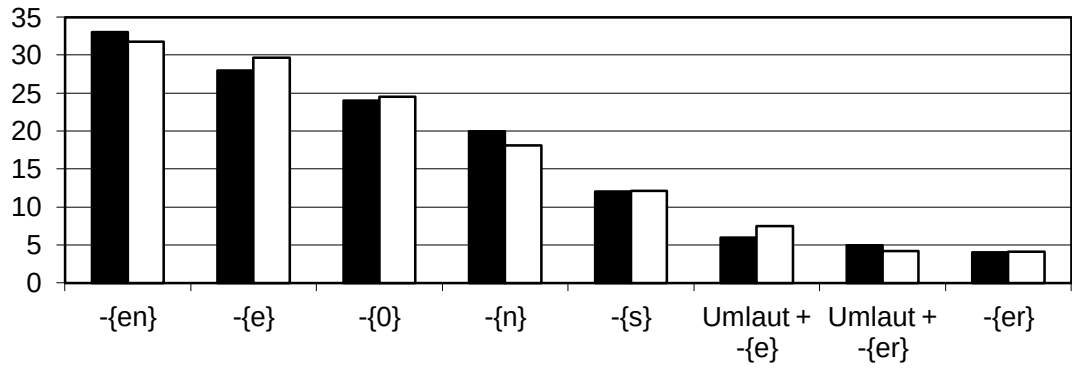


Figure 10.1 Concerning data in Table 10.1

When analyzing Heinrich von Kleist's letters Brüers & Heeren (2004) used the geometric distribution, because in some cases too few allomorphs and therefore too few classes were observed for the use of the hyper-Poisson distribution. Meuser, Schütte & Stremme (2008) fitted the negative hypergeometric distribution to 21 short stories by W. Schnurre.

b) The diversification of compound classes

Using the example of a total of ten advertisement texts Sowinski (1979: 110) analyzed how often different types of compounds were used. Those types are listed in the following table; it is demonstrated that the 1-displaced negative binomial distribution can be fitted to the file:

Table 10.2
Fitting of the 1-displaced negative binomial distribution to the types of compounds

Rank	Type	n_x	NP_x
1	Noun + noun	117	112.33
2	Adjective + noun	16	21.36
3	Noun + adjective	13	10.51
4	Verb + noun	8	6.23|
5	Adjective + noun	7	4.00|
6	Verb + adjective	3	9.57|
$k = 0.2396$ $p = 0.2061$ $FG = 1$ $X^2 = 3.555$ $P = 0.06$			

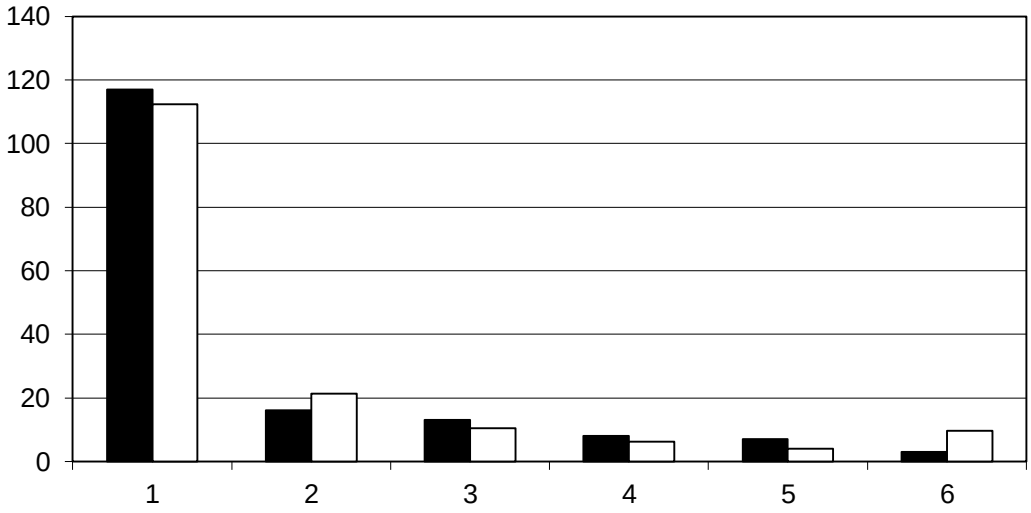


Figure 10.2. Concerning Table 10.2

Sowinski (1998: 67f.) mentions further data, but the types of compounds are not listed completely; so they are not appropriate for a test.

c) The diversification of "pray" in Shakespeare's dramas

Busse (1999: 486) analyzed the different occurrences of "pray" in all Shakespearean dramas. The 1-displaced mixed Poisson distribution can be fitted to the data observed.

Table 10.3
Fitting of the 1-displaced mixed Poisson distribution
to the different occurrences of "pray" in the Shakespearean works

Rank	Occurrence	n_x	NP_x
1	I pray you	253	250.10
2	pray you	134	134.86
3	[pray (other)]	132	125.05
4	pray	101	111.29
5	I pray thee	71	77.12
6	I pray	54	42.89
7	pray thee	21	19.88
8	pray ye	3	7.90
9	we pray you	3	2.75
10	I pray them	1	1.16
a = 2.7813		α = 0.6456	FG = 6
b = 0.2229		X^2 = 7.889	P = 0.25

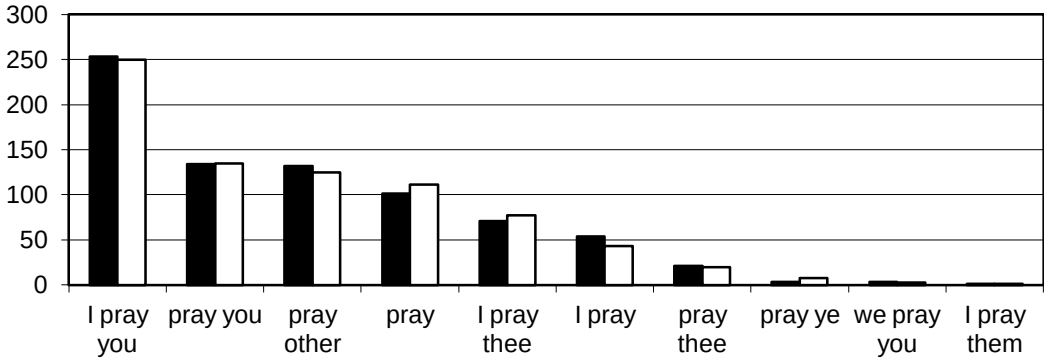


Figure 10.3. Concerning Table 10.3

Fitting is successful though data originate from a mix of many texts.

d) On the diversification of modal particles like German "eben", "halt" or the like

The use of modal particles by people in West and East Berlin was analyzed by Dittmar & Bredel (1999: 158f.). As an example of modal particles the use of different particles by a person from East Berlin (key: B 01 OF) is presented.

Table 10.4
Fitting of the positive Poisson distribution
to the distribution of the modal particles

Rank	Particle	n_x	NP_x
1	ebent	65	63.67
2	halt	42	39.86
3	eben halt	11	16.63
4	eben	9	6.84
a = 1.2519		FG = 2	P = 0.25

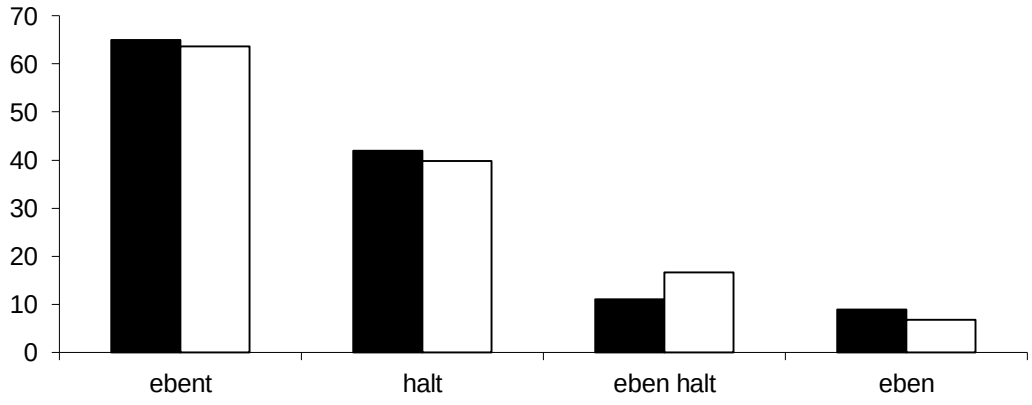


Figure 10.4. Concerning the data in Table 10.4

e) On the diversification of proper names

The occurrence of adjectival transfers of the type "Lang – Lange – Langen – Langer" is demonstrated by Seibicke (1982: 174) for four German big cities (Dresden, Hanover, Cologne, Munich). In all four cases the 1-displaced Hirata-Poisson distribution can be fitted; the same applies to the data for Göttingen collected from a telephone directory.

Table 10.5
Fitting of the 1-displaced Hirata-Poisson distribution to types of names

Rank	Type of name	n_x	NP_x
1	Lange	118	118.00
2	Langer	30	30.00
3	Lang	20	16.20
4	Langen	1	4.80
a = 0.3592	b = 0.2922	C = 0.0000	

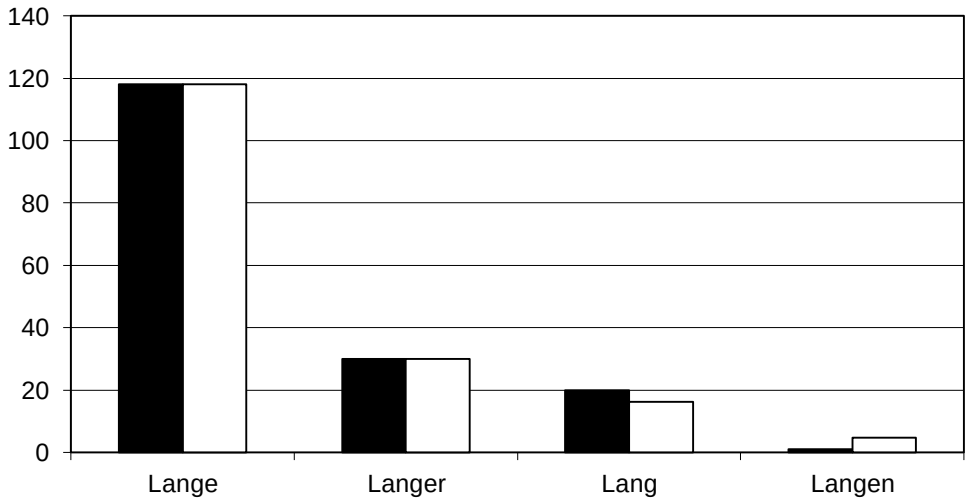


Figure 10.5. Concerning the data in Table 10.5

As to the files for the other towns there are differences with the order of the names, if they are listed according to frequency as it is done here. However, this does not affect the possible choice of the same distribution. In other cases different models had to be taken in case of substantially more comprehensive files with names (Best 2006a).

10.2 Functional-semantic diversification

Examples of functional-semantic diversification can e.g. be observed in word formation (cf. Altmann, Best, & Kind 1987). An Early Modern High German type of word formation was modeled by Best (1990) by means of the 1-displaced negative binomial distribution. The subject were nouns of the type "geschöpf" (=

creature), "gesetz" (= law), "gesicht" (= face) in Albrecht Dürer's bequest, whose occurrence is listed in the following table according to their functional statuses. The 1-displaced hyper-Poisson distribution can be fitted to those data with even better results than with the negative binomial distribution (both distributions are among the potential manifestations of the law of diversification: Altmann 1991: 39f.)

Table 10.6
Fitting of the 1-displaced hyper-Poisson distribution to the functional statuses of an Early Modern German type of word formation (Best 1990)

Rank	Functional statuses	n_x	NP_x
1	idiomatized	12	11.33
2	Collective nouns	11	11.79
3	Nomina patientis	11	10.27
4	with pleonast. affix	6	7.70
5	Abstracts	6	5.06
6	Simplicia	4	2.96
7	Nomina agentis	1	1.56
8	Nomina instrumenti	1	1.33
a = 5.3656	b = 5.1584	FG = 5	X^2 = 1.326	P = 0.93

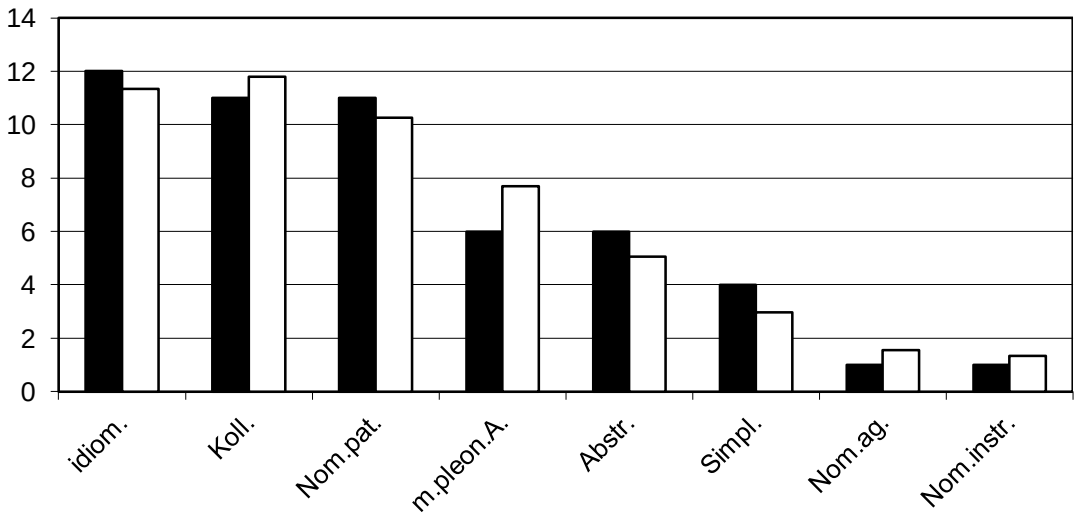


Figure 10.6. Concerning the data in Table 10.6

10.3 Two-dimensional diversification of the French particle "que"

Linguistic entities can be involved in several diversifications at the same time, as Hug (2001: 219) shows by his data collected. On the basis of a corpus he showed which types of word are realized by the French "que" and what the groups of

words are which are followed by it in texts. The disambiguation of this ambiguous word was his objective. However, the data are also productive for the analysis of the diversification, as we can see in the subsequent tables.

Table 10.7
Fitting of the 1-displaced hyper-Poisson distribution to the use of the French word "que" (occurrence after previous words)

Rank	Occurrence after	n_x	NP_x
1	Verb	501	504.84
2	Noun	408	391.07
3	Adverb	290	299.49
4	Adjective	223	226.78
5	Conjunction	165	169.81
6	PastPartic	128	125.75
7	DemPron	98	92.11
8	Preposition	69	66.74
9	Proper Noun	50	47.84
10	Negation etc.	38	33.94
11	Indf/InterAdj	22	23.82
12	PrPersSuj	10	16.55
13	PrPersDisj	9	11.38
14	RelInterPron	6	7.74
15	Numer.Card.	6	5.22
16	Nominal	3	3.48
17	IndefPron	2	2.30
18	PrpersCpl	1	1.50
19	Art *au, du*	1	0.97
20	DefiniteArt	1	0.63
21	Ord./PrPoss	1	0.40
22	Interjection	1	0.65

$a = 67.2658$ $b = 86.8338$ $FG = 17$ $X^2 = 7.316$ $P = 0.98$

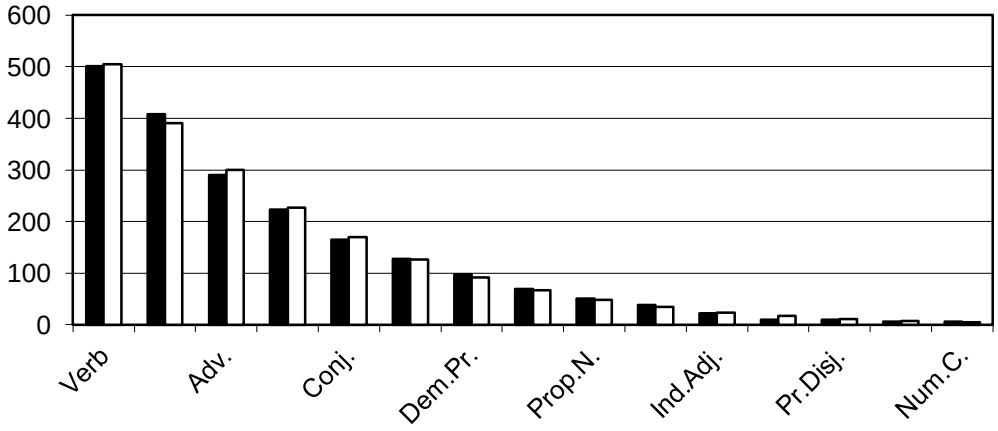


Figure 10.7. Concerning the data in Table 10.7

In the graphical diagram just the first 15 ranks are taken into consideration. The categorizations were adopted from Hug (2001: 219).

The second diversification presented by Hug (2000) is defined by the various applications of the particle.

Table 10.8

Fitting of the modified negative binomial distribution to the uses of *que* (data taken from the room document of the address in Prague on 25-8-2000 on the occasion of the *QUALICO IV*)

Rank	Occurrence as	n_x	NP_x
1	Simplecj.	870	866.13
2	Cpndcj.	325	322.07
3	RelPron.	318	313.90
4	Compar.	212	232.29
5	Restrict.	173	144.65
6	Emphat.	54	79.90
7	Correlat.	49	40.38
8	IntPron.	25	19.06|
9	Imperat.	6	8.51|
10	Quantit.	1	6.10|
k = 6.1974	p = 0.7292	α = 0.6684	X^2 = 17.742 C = 0.0087

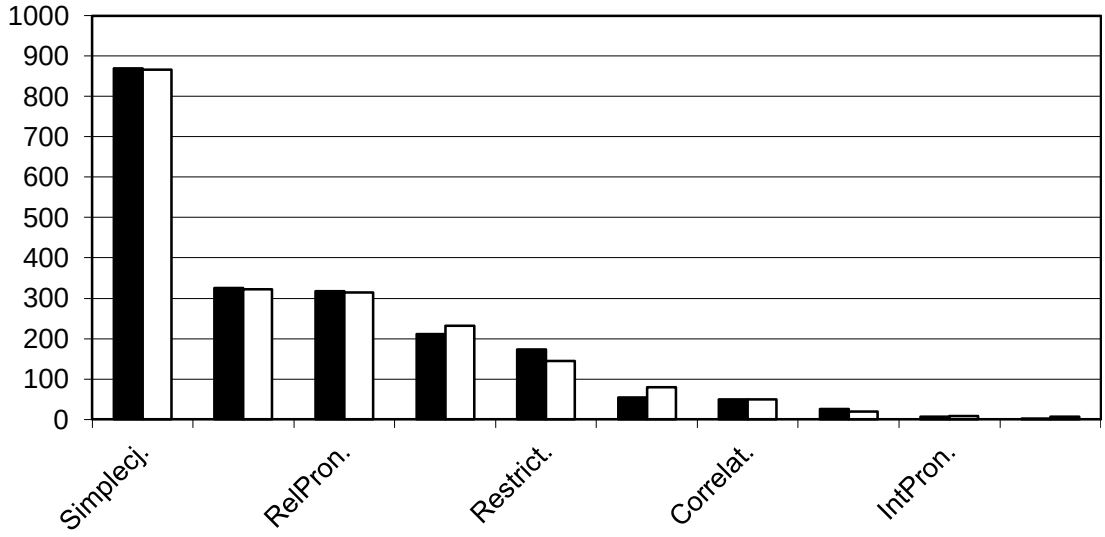


Figure 10.8. Concerning the data in Table 10.8

So, the example of the French "que" shows that one and the same entity can simultaneously diversify with different dimensions and each of those diversifications passes off in compliance with the theory.

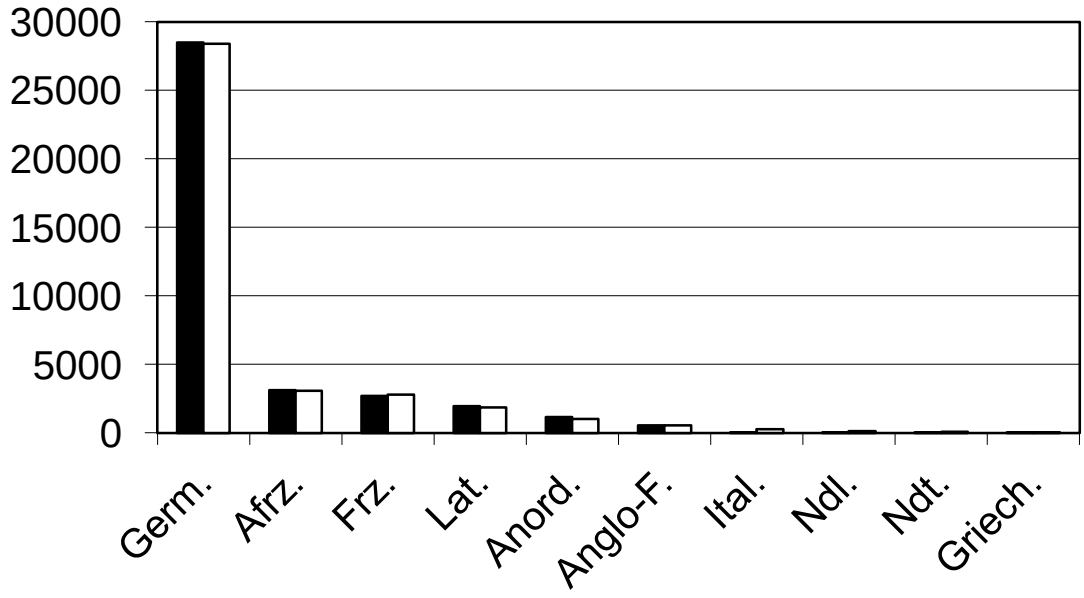
10.4 Diversification of the vocabulary according to the original language

A further type of diversification can be observed when the objective of the analysis is finding the original language of the vocabulary of a language or text (Rothe 1991a: 30). In Finkenstaedt & Wolff (1973: 119) this is called the "etymological spectrum" of a language. Wolff (1969: 156) collected such data for English press texts; each occurrence of a word (token) was taken into account. In the following table words whose origin could not be determined were omitted; further to that, name derivations were omitted as well.

Table 10.9
Fitting of the negative binomial-Poisson distribution to words of different origins of languages in English press texts

x	Origin	n_x	NP_x
1	Germanic	28496	28397.30
2	Old French	3111	3067.95
3	French	2696	2803.86
4	Latin	1966	1839.12
5	Old Nordic	1163	1017.50
6	Anglo-French	565	526.84
7	Italian	51	269.43
8	Dutch	41	137.77
9	Low German	35	70.07
10	Greek	33	35.31
11	Scandinavian	15	17.63
12	Spanish	10	8.74
13	German	3	4.30
14	Hindustani	3	2.11
15	Anglo-Irish	2	1.03
16	Gaelic	2	0.50
17	Icelandic	2	0.24
18	Portuguese	2	0.11
19	Arabic	1	0.05
20	Danish	1	0.02
21	Flemic	1	0.01
22	Scottish	1	0.11
k = 2.6551		p = 0.8732	FG = 11
a = 1.6791		X^2 = 351.492	C = 0.0092

The following graphic diagram only presents the most frequent languages of origin, because the scale given does not allow that the others are made visible.



Germ. = Germanic - Afrz. = Old French – Frz. = French – Lat. = Latin – Anord. = Old Nordic – Anglo-F. = Anglo-French – Ital. = Italian – Ndl. = Dutch – Ndt. = Low German – Griech. = Greek

Figure10.9. Concerning data in Table 10.9
(incl. the meanings of the abbreviations)

The following table indicates the fitting of the Altmann model for arbitrary rankings

$$y_x = \frac{\binom{b+x}{x-1}}{\binom{a+x}{x-1}} c, \quad x = 1, 2, 3, \ldots$$

(Altmann 1993: 62) to the etymological spectrum of the German language (acc. to Körner 2004: 29; just loanwords, without the "foreign words" formed in German). Calculation is implemented with $c = y_1$; a and b are the parameters of the model.

Table 10.10
Körner's data base: complete evaluation of the datable vocabulary of 16781 words with 5244 of them being loanwords in *Duden. Herkunftwörterbuch* (2001).

x	Origin	n_x	NP_x	x	Origin	n_x	NP_x
1	Latin	2031	2031.00	17	Portuguese	5	1.23
2	French	1424	1225.32	18	Celtic	4	0.81
3	Low German	545	743.93	19	Eskimo	3	0.53
4	English	519	454.45	20	Gothic	3	0.35

5	Italian	286	279.28	21	Indian	3	0.24
6	Greek	144	172.64	22	Africaans	2	0.16
7	Dutch	87	107.33	23	Icelandic	2	0.11
8	Slavic	44	67.10	24	Malayan	2	0.07
9	Spanish	43	42.18	25	Chinese	1	0.05
10	Cant	41	26.66	26	Finnish	1	0.03
11	North German.	12	16.93	27	Hebrew	1	0.02
12	Japanese	10	10.81	28	Hunnish	1	0.02
13	Hungarian	10	6.94	29	Ladin	1	0.01
14	Turkish	8	4.47	30	Persian	1	0.01
15	Yiddish	6	2.90	31	Polynesian	1	0.00
16	Arabic	5	1.89	32			
a = 100.8222 b = 60.0335 D = 0.9856							

The graphical diagram elucidates the good result:

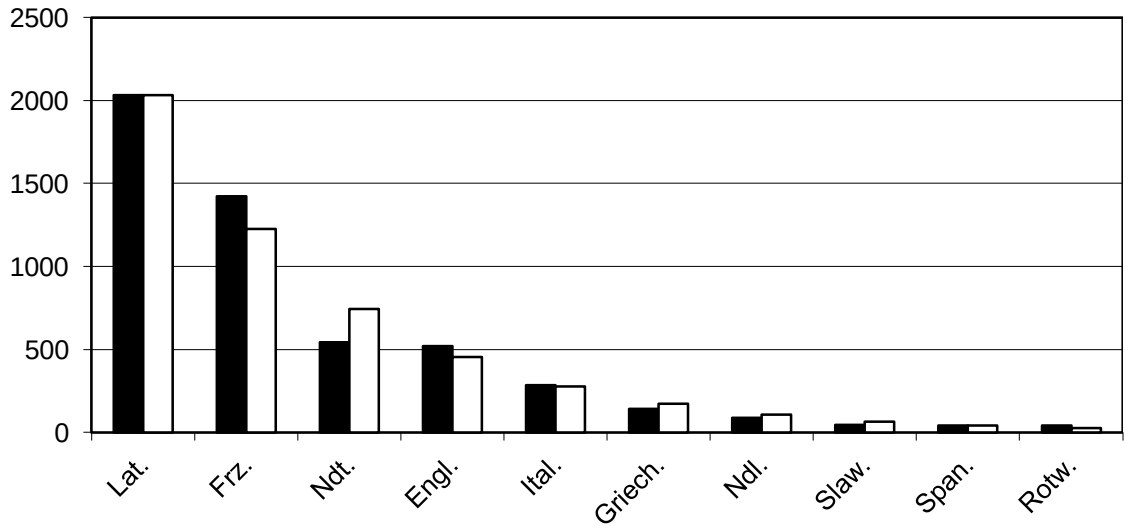


Figure 10.10. Concerning the data in Table 10.10

Further data of etymological spectra concerning the German language are found in Best (2001b: 11, 14; 2005f), as to the English lexicon in Finkenstaedt & Wolff (1973: 119f.), Best (2005f) and the Turkish language in Best (2005; 2008).

The examples show that a high number of very different phenomena complies with the assumptions concerning the law of diversification. This is the point where it must be pointed out that such phenomena also include distributions of entities that have already been demonstrated as the realization of a different linguistic law. Thus, Schweers & Zhu (1990) have already treated the distributions of word types in different languages as a phenomenon of diversification. Here, it is indicated that linguistic laws seen so far as something different can indeed be traced back to a common basic mechanism (Wimmer & Altmann 2006).

11. Discovery of Linguistic Laws IV: Martin's Law

When looking for a word in a lexicon to find out its meaning, this word looked for is explained by other words. When looking up "Sessel" (= armchair) in the *Duden, Deutsches Universalwörterbuch (²1989)* the basic explanation for that word is: "Sitzmöbel" (= seating furniture). If the explaining word "seating furniture" does not help you, you look it up and get to know "Möbel" (= furniture). Then, for "furniture" you get: "Einrichtungsgegenstand" (= furnishing), followed by the explanation "Gegenstand" (= object). Thus, you can build up "Sequences of definitions" for arbitrary cues (Sambor & Hammerl 1991) of the type "armchair – seating furniture – furniture – furnishing – object" that can be identified by the fact that more specific terms are replaced with more and more general ones. If that process is applied to many words, that application results in levels beginning with specific words (e.g. like above "armchair") and end with very general terms (like here "object"). In their development from the very special to the very general those levels are occupied by fewer and fewer terms. It seems that such observations were made by Martin (1974) for the French lexis for the first time; that is why subsequent analyses speak about "Martin's law(s)". Martin's law expresses that specific proportions based on definition chains exist; in their simplest form they appear as a geometric distribution (Wimmer et al. 1994: 101).

Schierholz (1991: 41-44) resumed examinations of German material concerning Martin's law in the form of tables; the compilation shows that results relatively strongly depend on the dictionaries used. A problem of his analysis was that the dictionary mainly used turned out to be inadequate; that made it impossible to obtain a confirmation for this law on the basis of the German language

Finally, on a different lexical basis and by an improved preparation of the sequences of definitions Hammerl (1991b) could show that a very good conformity with the coherence he called "Martin's law I" could be achieved. He defined the connection between the number of terms and the rank number of that level by means of the formula

$$y_{x+1} = \frac{c}{(x+1)^a} \cdot y_x^b .$$

y_{x+1}: number of terms at level $x+1$
y_x: number of terms at the next lower level
a, b and c: constants

The testing of the model by Hammerl (1991b: 57) came to results presented in Table 11.1 and Figure 11.1.

Table 11.1
Number of terms y_x per level of terms *x* in a German dictionary

x	y_x	$y_{x\ calculated}$
1	1000	1000
2	512	519.5
3	183	197.9
4	72	75.3
5	31	31.5
6	14	14.7
7	9	7.6
8	5	4.3
9	3	2.6
10	2	1.7
11	1	1.2
a = 1.4710 *b* = 0.5631 *c* = 29.4456 *D* = 0.9997		

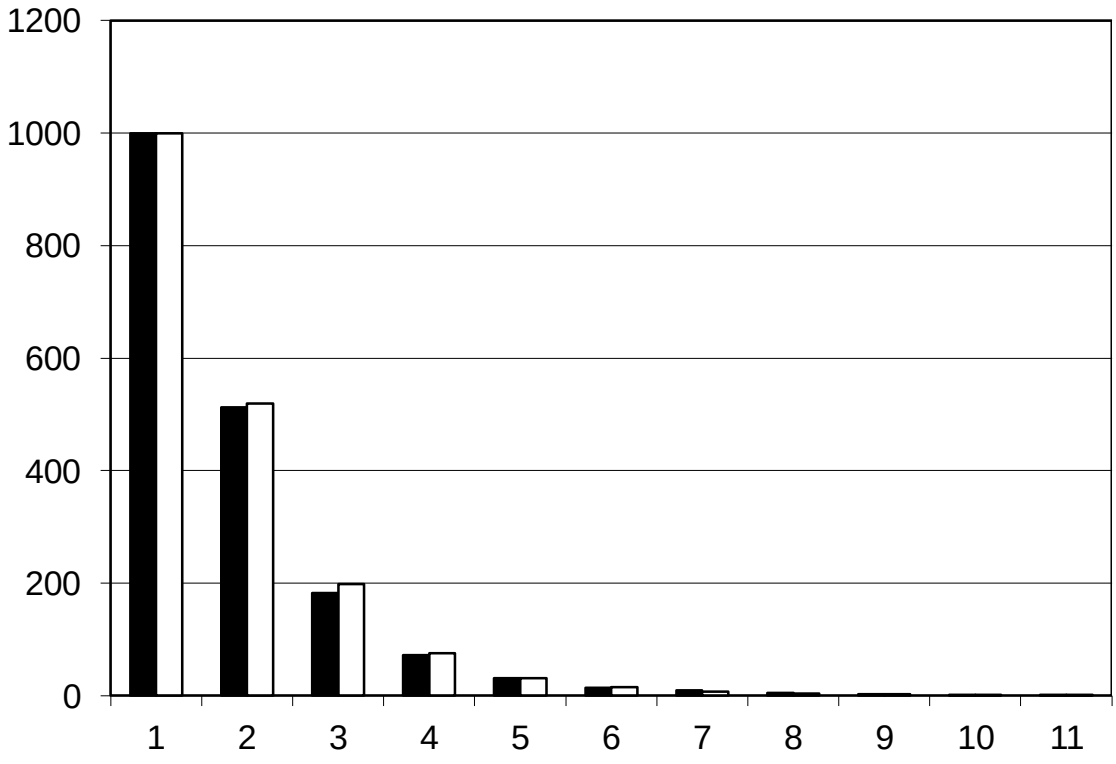


Figure 11.1. Fitting Hammerl's formula to German vocabulary

So, testing Martin's law has proved to be as successful as it was for French and Polish (Hammerl 1991b: 57).

12. Discovery of Linguistic Laws V: Menzerath-Altmann's Law

So far, - with the exception of Martin's law – the subject has always been distributions acting out within a linguistic level and realizing how often units or classes of units of the same kind occur in the text or lexicon. This can be seen as the "horizontal" perspective of language. The Menzerath-Altmann law processes a different perspective of language: the "vertical" one that connects linguistic levels with each other. A decisive step towards that was the measurement of the duration of sounds, syllables and words in Spanish; they resulted in the formulation of general and specific quantitative laws by Menzerath & de Oleza (1928: 68ff.). The most important ones were as follows (marking as in the original):

"1. The average duration of the sound in the word becomes *smaller*, when the number of sounds in the word *increases*." (68)
"4. A sound becomes *shorter* when the number of syllables of the word increases." (70)
"6. The average duration of syllables in general *decreases* when the number of syllables of the word … *increases*." (71)
„9. The average duration of a word *increases* when the number of sounds in a word *increases*." (73)

Menzerath (1954: 100f.) detected a "law-based ratio" between the number of syllables and their lengths in a word; he worded the following "economy rule": "The larger the entity, the smaller the components!" Altmann & Schwibbe (1989: 5) worded this relationship a little more linguistically: The longer a linguistic construct, the smaller its constituents (Altmann & Schwibbe 1989: 8ff.). The relevant law was derived first by Altmann (1980; also see Altmann & Schwibbe 1989: 6f.); in its simplest form it says:

$$y = ax^b.$$

In the meantime, a lot of tests concerning the Menzerath-Altmann law have been implemented; many of those are found in Altmann & Schwibbe (1989: 37ff.). Finally, Hřebíček (1997; 2000) – using the example of the first eight sections of a Turkish text – could prove that that law was effective over all linguistic levels from the sentence aggregate (= group of sentences having a lexeme in common, today called "hreb"; Hřebíček 1997: 31) being the biggest unit down to the level of phonemes. Köhler (1984) proposes to interpret that law as a consequence of the human language processing process; it also proves to be successful in applications beyond linguistics (Altmann & Schwibbe 1989: 84ff.).

An example of the Menzerath-Altmann law is presented by Gerlach (1982: 95), who tested the hypothesis of "The longer the word, the shorter its

morphs" on the basis of a fully analyzed dictionary. This hypothesis combines the level of words with that of its constituents, the morphs. Fitting the function $y = ax^b$ to Gerlach's data the result is presented in Table 12.1 and Figure 12.1:

Table 12.1
Decrease of morphic length n_x when word length x increases

x	n_x	f_x
1	4.53	4.29
2	3.25	3.46
3	2.93	3.06
4	2.78	2.80
5	2.65	2.61
6	2.58	2.47
a = 4.2854	b = -0.3074	D = 0.95

x: word length (in morphs);
n_x: observed average length of morphs (number of phonemes per morph);
f_x: calculated length of morphs[15];
D: determination coefficient, producing a good result by $D \geq 0.80$.

With D = 0.95 the determination coefficient confirms – as the following diagram does – a very good conformity between model and observation:

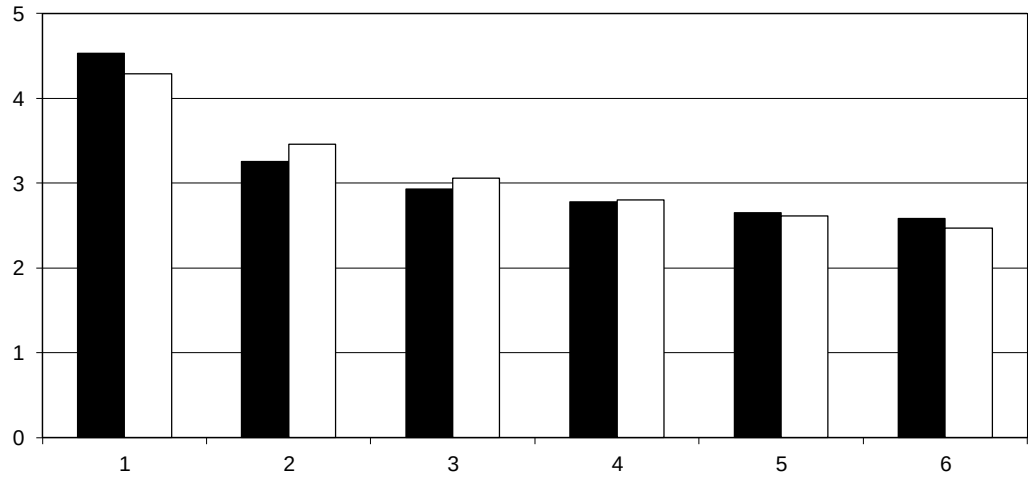


Figure 12.1. Decrease of morphic length n_x when word length x increases

Gerlach (1982: 100f.) achieved an even better fitting result by a more complex version of the Menzerath-Altmann law. Polikarpov (2006) pleaded in favor of a

[15] Those calculations required that NLREG was applied instead of the Altmann-Fitter (1997), because the observations were no integer values.

more differentiated handling of the law in case of different morph classes (prefixes / suffixes, roots) and their positions in the word.

Interestingly enough, the Menzerath-Altmann law does not only apply to the relationships of construct and constituent on the side of word forms in language, but also to the ratio of length and number of meanings of a term. For that, the relationship between word length x (measured by the number of letters per word) and the number of meanings of the words in German n_x, also produced from a dictionary (Altmann & Schwibbe 1989: 69; newly calculated), are presented in Table 12.2 and Figure 12.2.

Table 12.2
Decrease of polysemy n_x with word length x increasing

x	n_x	f_x	x	n_x	f_x
2	4.64	3.85	9	1.89	1.96
3	2.89	3.21	10	1.68	1.87
4	2.14	2.82	11	1.75	1.79
5	2.81	2.55	12	1.80	1.72
6	2.36	2.35	13	1.40	1.66
7	2.22	2.20	14	1.93	1.61
8	2.46	2.07	15	1.60	1.56
a = 5.2446 b = -0.4476 D = 0.81					

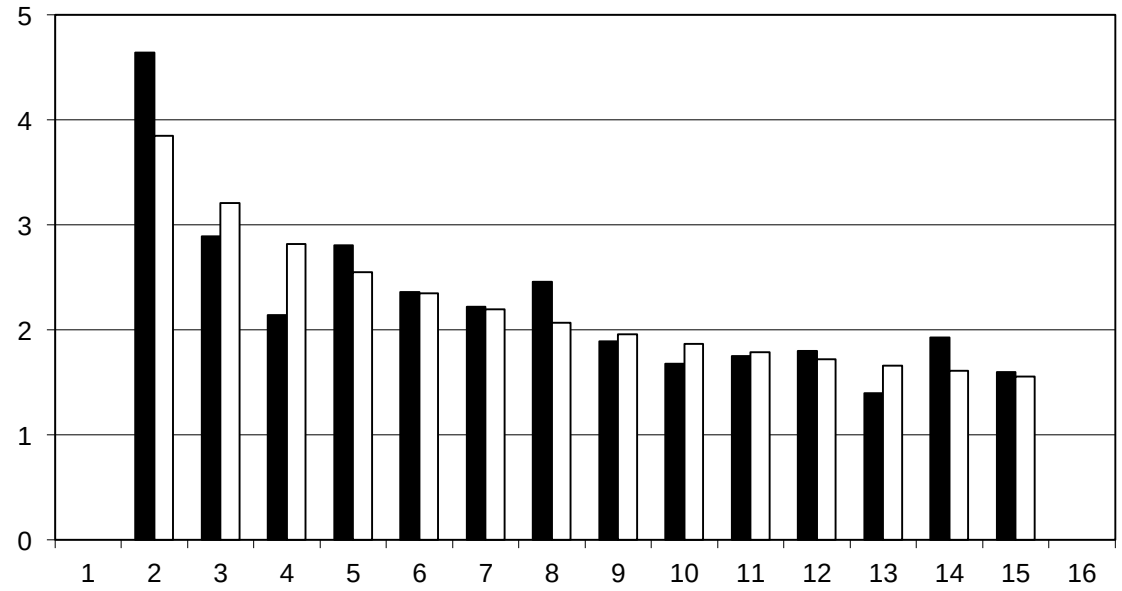


Figure 12.2. Decrease of polysemy n_x with word length x increasing

With D = 0.81 conformity between observation and theory is acceptable.

Even better results were found by the analyses of the other 10 examined languages (Altmann & Schwibbe 1989: 69-71).

However, further analyses concerning the Menzerath-Altmann law also showed that the correlation between construct and constituent does not manifest

that the increasing size of the construct results in a permanent decrease of the size of the constituent. To cope with such conditions a generalization of that law proposed by Altmann (1980: 3) puts itself forward: "The length of the component is a function of the length of language constructs"; it proved to be applicable in tests of some German and Turkish texts. For demonstration a short prose text, "Fugen" by Guntram Vesper (In: G. Vesper. 1985. *Kriegerdenkmal ganz hinten*. Frankfurt: Fischer Taschenbuch Verlag. S. 18-21) is taken to present the ratio of word length (measured in syllables) and syllable length (measured in phonemes, beginning with the consideration of all words (tokens). The model used to fit the two following text files is:

$$y = ax^b e^{(\frac{c}{x} + dx)}$$

Table 12.3
Decrease of syllable length n_x with word length x (token) increasing

x	n_x	f_x
1	2.88	2.88
2	2.62	2.58
3	2.65	2.69
4	2.70	2.71
5	2.57	2.62
6	2.60	2.45
7	2.14	2.23
a = 0.6423 b = 1.5548 c = 1.7933 d = -0.2912 D = 0.88		

(key in analogy to page 101; the curve in the following diagram shows the calculated course of the values f_x.)

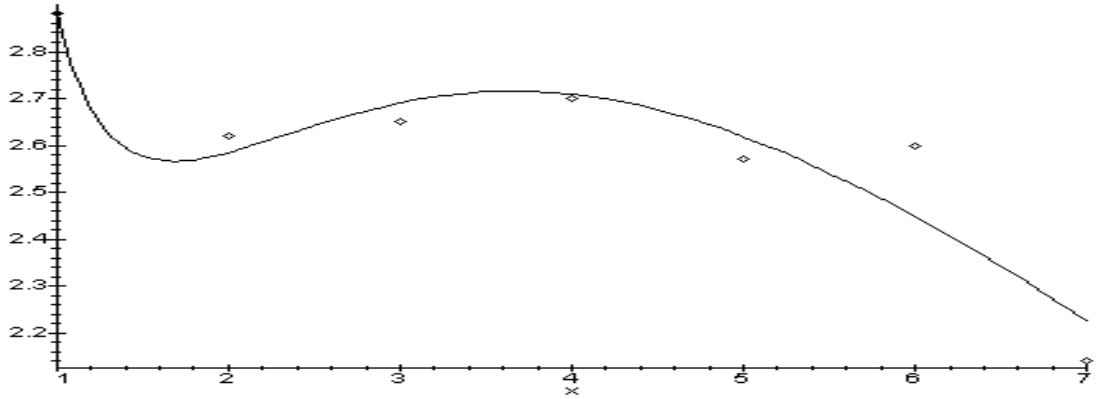


Figure 12.3. Decrease of syllable length n_x with word length x (token) increasing

A slightly better result is achieved for word form types (cf. Table 12.4 and Fig.12.4):

Table 12.4
Decrease of syllable length n_x with word length x (types) increasing

x	n_x	f_x
1	3.37	3.37
2	2.69	2.65
3	2.66	2.71
4	2.72	2.72
5	2.57	2.62
6	2.60	2.45
7	2.14	2.22
a = 0.4550	b = 1.7833 c = 2.3199	d = -0.3163 D = 0.95

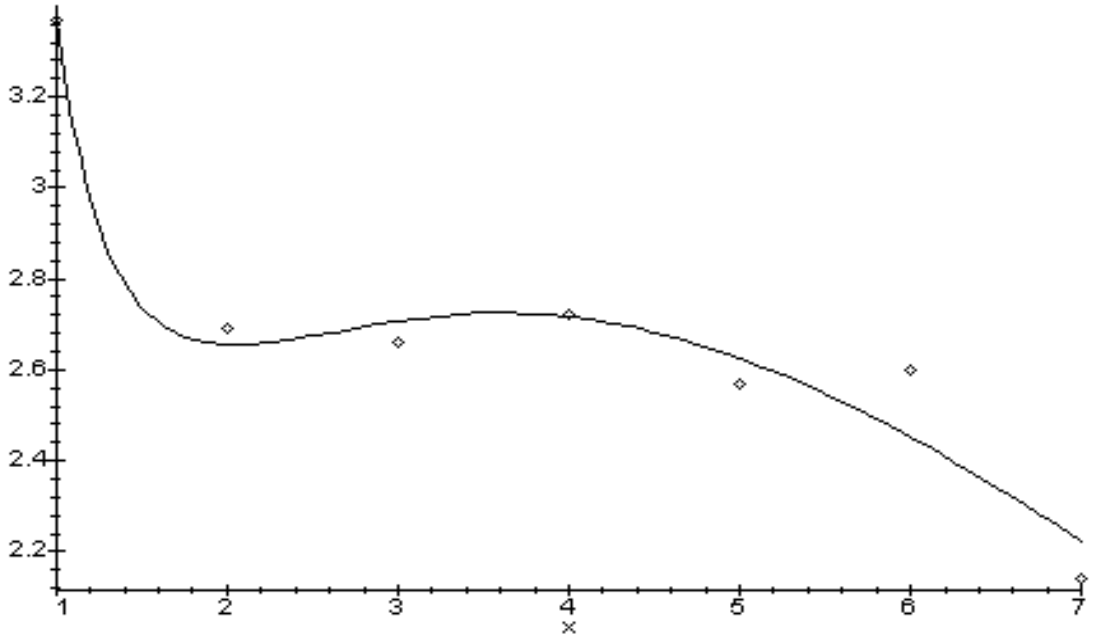


Figure 12.4. Decrease of syllable length n_x with word length x (types) increasing

Even better results are achieved if the rarely used six- or seven-syllable words are omitted.

(For further reference concerning possible forms of the Menzerath-Altmann law see: Fickermann, Markner-Jäger & Rothe 1984: 120. For further detailed reviews of the law see Asleh & Best 2004/05. An overview can be found in Cramer 2005; an evaluation of the significance of Menzerath for quantitative linguistics can be found in Best 2006l).

13. Discovery of Linguistic Laws VI: The Law Found by Zwirner, Zwirner & Frumkina (Law on Text Blocks)

This law describes how often a specific unit occurs in sections of a text. So, e.g. you can ask yourself how often the conjunction "and" occurs in text blocks of specific length. The probably first analyses of that kind are found with Zwirner & Zwirner (1935, 1938); they concentrated on the frequency of sounds in German in text sections of 100 sounds each. In 1962, Frumkina (Altmann & Burdinski 1982: 148f.) analyzed the occurrence of certain words in Pushkin's works in 1000-word blocks. Her assumption that units were governed by the Poisson distribution could no longer be maintained in that form; Brainerd (1972: 18f.), however, achieved convincing results on the basis of the same hypothesis for the distribution of the article in 50-word blocks in English texts taken from different text types. Altmann & Burdinski (1982: 150ff.), however, proposed the negative hypergeometric distribution as the basic model in relation to which the Poisson distribution as well as the binomial and the negative binomial distributions are only special cases (also see Altmann 1988a: 176f.).

This law that is mostly named after Frumkina has proved to be successful in this new form in a number of cases. Apart from examinations of sounds by Zwirner & Zwirner (1935, 1938) and a first attempt concerning syntax (Köhler 2001) there seem to be only a few tests concerning German texts; they refer to "das", "bis" and "in" in Siegfried Lenz, Deutschstunde (Altmann & Burdinski 1982: 158ff.). As far as tests could be performed, the negative hypergeometric distribution and the negative binomial distribution turned out to be appropriate models. To extend the empirical basis in the German language the distribution of five words in 50- and 100-word blocks was examined in the story "Dazugehören" by Jägersberg. Without the heading the story comprises 8005 words, which results in 160 fifty-word blocks, conforming to 80 100-word blocks; the excessive five words and the heading were not taken into consideration. In all ten cases the negative hypergeometric distribution turned out to be a successful model. The negative binomial distribution often permits even better results, but does not fit to one of the files. The example taken here is the occurrence of "und" in 50- and 100-word blocks. The other units are only presented in tables, but without diagrams.

The following tables begin with the text blocks with $x = 0$ in which "and" does not occur. $x = 1$ stands for text blocks, in which "and" occurs once etc.

The negative hypergeometric distribution is defined as follows:

$$P_x = \frac{\binom{-M}{x}\binom{-K+M}{n-x}}{\binom{-K}{n}}, \quad x = 0, 1, 2, ..., n$$

Fitting the negative hypergeometric distribution to the occurrences of "and" in 50-word text blocks (Otto Jägersberg, Dazugehören. In: ders., *Der letzte Biß*. Zürich: Diogenes 1977, p. 113-164.) yields the results presented in Table 13.1 and Figure 13.1.

Table 13.1
Fitting the negative hypergeometric distribution to 50-word text blocks

x	n_x	NP_x
0	30	29.67
1	47	49.97
2	51	44.36
3	20	25.26
4	10	9.08
5	2	1.66
K = 14.0358		X_2^2 = 2.444
M = 4.5441		P = 0.29
n = 5.0000		

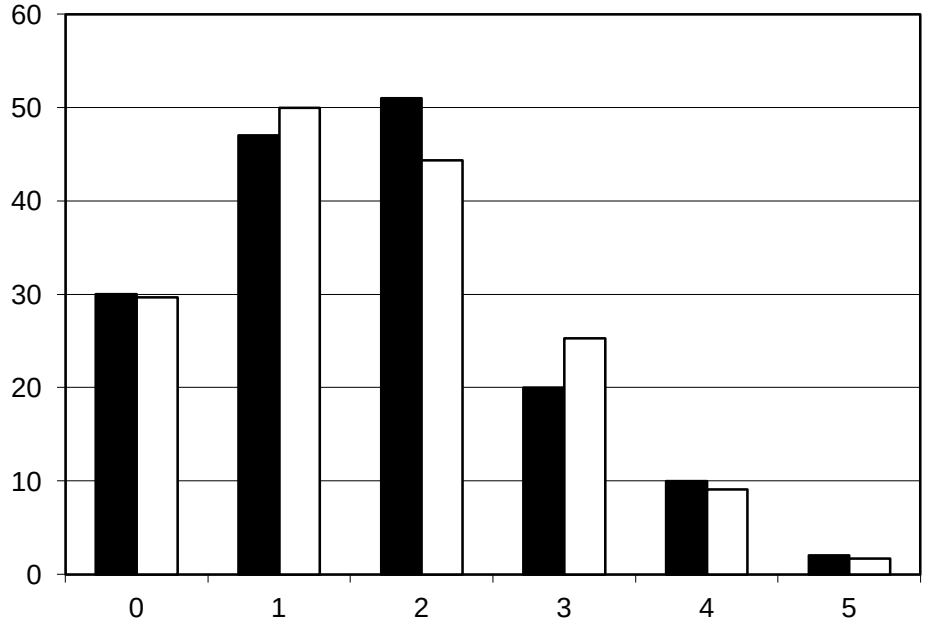


Figure 13.1. Fitting the negative hypergeometric distribution to the occurrences of "and" in 100-word text blocks (Otto Jägersberg, Dazugehören. In: ders., *Der letzte Biß.* Zürich: Diogenes 1977, p. 113-164.)

Table 13.2
Fitting the negative hypergeometric distribution to 100-word text blocks

x	n_x	NP_x
0	3	2.42
1	7	8.63
2	14	15.74
3	25	19.05
4	16	16.66
5	9	10.74
6	3	4.97
7	1	1.51
8	2	0.28
K = 21.9064		X_4^2 = 4.492
M = 8.8995		P = 0.34
n = 8.0000		

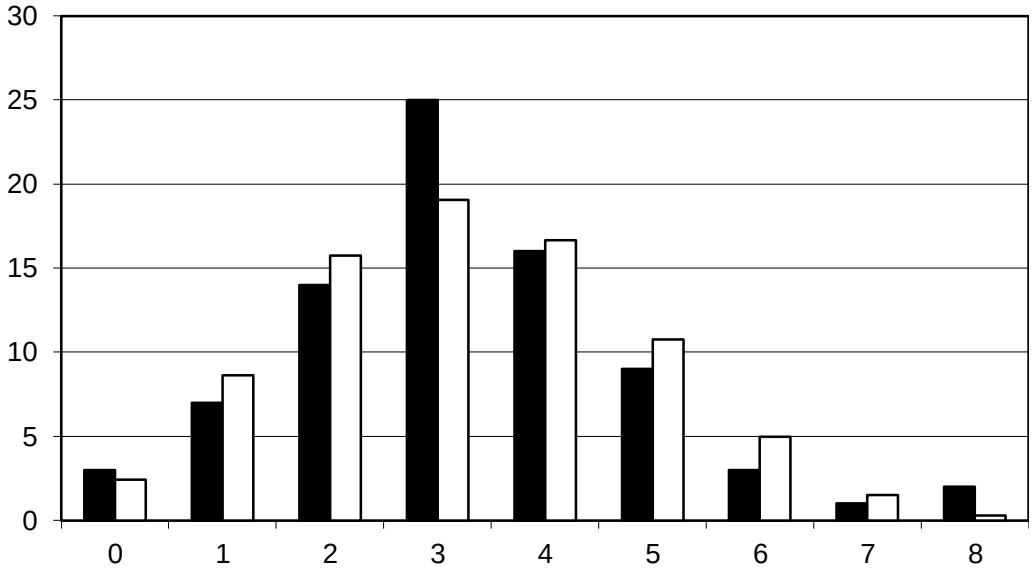


Figure 13.2 Concerning data in Table 13.2

The fitting of the negative hypergeometric distribution to the files of the other words “nicht“, “er“, “ich“ and “Franz“ (= hero of the story) examined is successful in all cases:

The fitting of the negative hypergeometric distribution to the occurrence of “nicht“, “er“, “ich“ and “Franz“ in 50-word text blocks (Otto Jägersberg, Dazugehören. In: ders., *Der letzte Biß*. Zürich: Diogenes 1977, p. 113-164.) is presented in Table 13.3

Table 13.3

	"nicht"		"er"		"ich"		"Franz"	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	70	72.45	80	79.95	105	108.02	57	59.16
1	58	52.63	44	43.12	27	22.11	66	60.06
2	24	24.23	20	21.55	9	11.79	26	30.59
3	5	8.23	12	9.78	5	7.39	9	8.93
4	1	2.06|	2	3.90	12	4.84|	2	1.26
5	1	0.35|	1	1.30|	1	3.12|		
6	1	0.05|	0	0.32|	0	1.85|		
7			1	0.08|	1	0.88|		
K =	16.0420		8.0268		2.1661		16.5712	
M =	2.2726		1.0034		0.2321		3.9615	
n =	6.0000		7.0000		7.0000		4.0000	
X^2 =	2.027		1.627		3.628		1.834	
FG =	1		2		1		1	
P =	0.15		0.44		0.06		0.18	

and 100-word blocks in Table 13.4

Table 13.4

The fitting of the negative hypergeometric distribution to the occurrence of "nicht", "er", "ich" and "Franz" in 100-word text blocks (Otto Jägersberg, Dazugehören. In: ders., *Der letzte Biß*. Zürich: Diogenes 1977, p. 113-164.)

x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	13	15.52	20	21.16	36	34.85	8	13.55
1	31	24.08	23	20.16	18	14.82	30	22.41
2	20	20.40	12	15.74	9	9.42	20	20.28
3	10	12.08	16	10.89	4	6.56	13	13.07
4	2	5.41	3	6.68	4	4.70	3	6.61
5	2	1.88|	4	3.51	3	3.39	4	2.74
6	0	0.50|	1	1.45|	3	2.40	2	1.34
7	0	0.09|	1	0.41|	1	1.65		
8	1	0.01|			1	1.08		
9	1	0.03|			0	0.64|		
10					1	0.49|		
K =	28.1384		5.5190		3.2294		47.3405	
M =	5.3150		1.3798		0.4925		6.3739	
n =	9.0000		7.0000		11.0000		14.0000	
X^2 =	5.825		5.859		2.310		7.738	
FG =	2		3		6		3	
P =	0.05		0.12		0.89		0.05	

Thus, testing the law by Zwirner, Zwirner & Frumkina succeeded in five further units in the German language.

In Zwirner & Zwirner (1935: 44) a table can be found in which the authors indicate the distribution of the German sound [œ] in 200 text blocks; each of them includes 100 sounds. In compliance with Fumkina's assumptions the Poisson distribution

$$P_x = \frac{e^{-a}a^x}{x!}, \; x = 0,1,2,...$$

can be fitted to the file presented below:

Table 13.5
The distribution of [œ] in text blocks

x	n_x	NP_x
0	180	178.51
1	18	20.29
2	2	1.20
a = 0.1137	X_1^2 = 0.808	P = 0.37

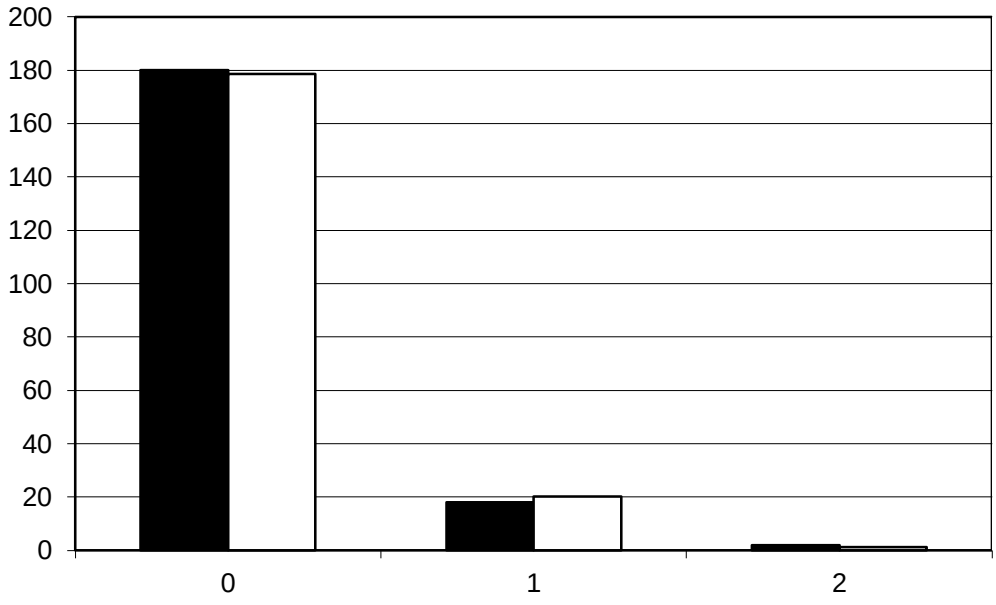


Figure 13.5. Concerning data in Table 13.5

The relevant data concerning [b] and [ə] are found in Zwirner & Zwirner (1938); test result concerning them in Best (2005h: 6).

Thus, the law found by Zwirner, Zwirner & Frumkina got a number of further confirmations; its first discovery seems to have been successful, see

Zwirner & Zwirner (1935, 1938). A further analysis on the basis of a low-German text can be found in Suhren (2002), whose data – apart from others – are partly reproduced in Best (2005h, 2006e) together with new calculations.

The number of [k]s at the beginning of lines in Baudelaire's sonnets was analyzed by Knauer (1971: 198); it is also governed by the Poisson distribution (Best 2006h).

14. Discovery of Linguistic Laws VII: The Law of Vocabulary Dynamics

The occurrence of new words not used up to a certain point in the text can be analyzed in two ways: 1. You either analyze word by word whether an unused one occurs, or 2. you form text blocks as in the previous section and find out how many new words occur in a certain text block. In both cases you get an impression of the vocabulary dynamics in the relevant text which is also subjected to laws. The process is to be introduced by the example of Büchner's *Der hessische Landbote*; it is based on text blocks and conforms to the vocabulary dynamics in texts using children's language (p. 17-18). Again,

$$y = ax^b$$

is used as the model. First, text block by text block is analyzed to indicate how many word form types are new. A clear trend is elucidated:

Table 14.1
Increase in new word form types per 250 tokens
in G. Büchner, *Der hessische Landbote*

Text block (250 tokens each)	New types (observed)	New types (calculated)	Text block (250 tokens each	New types (observed)	New types (calculated)	Text block (250 tokens each	New types (observed)	New types (calculated)
1	165	156.36	7	77	80.42	13	72	65.09
2	107	123.39	8	72	76.83	14	54	63.46
3	103	107.42	9	77	73.80	15	63	61.98
4	103	97.37	10	65	71.19	16	53	60.63
5	95	90.22	11	89	68.91	17	62	59.39
6	85	84.77	12	67	66.89			
a = 156.3614 b = -0.3417 D = 0.91								

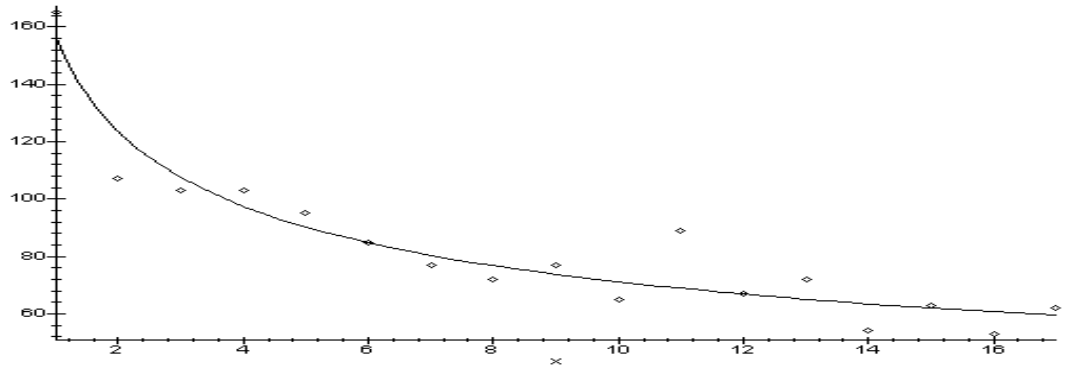


Figure 14.1. Concerning the data in Table 14.1

Including the author's name and the title the text comprises 4275 words (tokens) of 1417 word form types. 4250 running words were taken into consideration from the beginning of the text; the remaining 25 were not considered.

In the following table and diagram the new word form types were added text block by text block; the same model is fitted:

Table 14.2
Increase in new word form types per 250 tokens in
G. Büchner, *Der hessische Landbote* (cumulated values)

Text block (250 tokens each)	New types (observed)	New types (calculated)	Text block (250 tokens each)	New types (observed)	New types (calculated)	Text block (250 tokens each)	New types (observed)	New types (calculated)
1	165	169.86	7	735	730.99	13	1177	1162.91
2	272	285.67	8	807	807.99	14	1231	1229.38
3	375	387.20	9	884	882.62	15	1294	1294.67
4	478	480.44	10	949	955.19	16	1347	1358.88
5	573	567.96	11	1038	1025.97	17	1409	1422.09
6	658	651.18	12	1105	1095.16			
a = 169.8604 b = 0.7500 D = 0.9994								

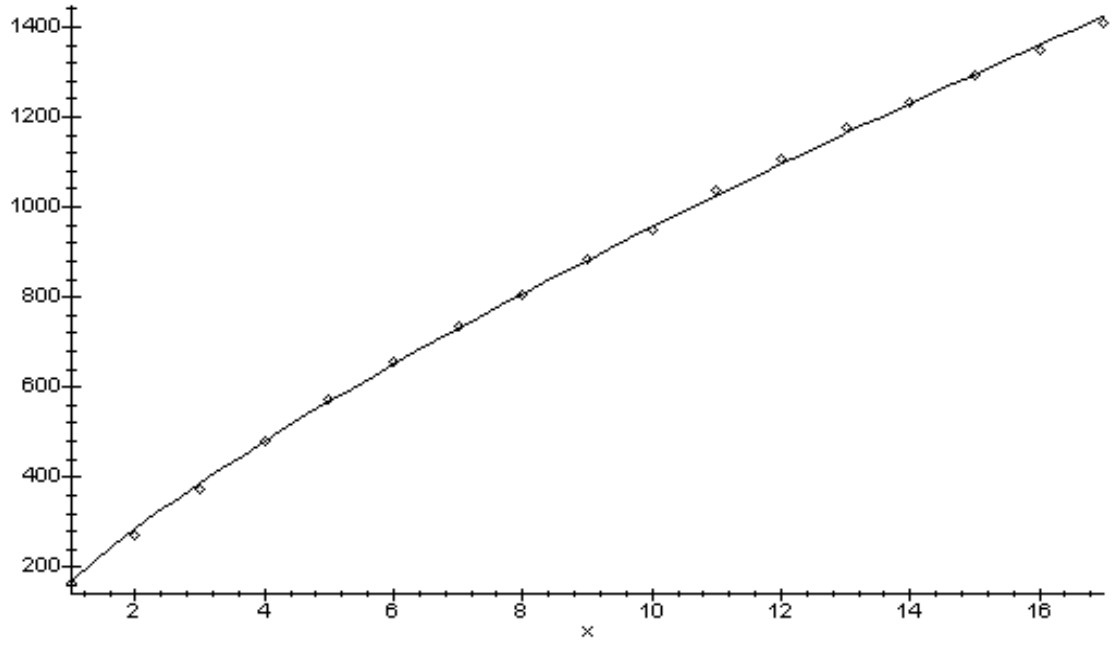


Figure 14.2. Concerning the data in Table 14.2

The laws of vocabulary dynamics belong to the central subjects in quantitative linguistics; they are of theoretical interest, but also relevant for problems of style, e.g. when reflections on the poverty or richness of vocabulary in certain works are required (see Best 2004b,c, 2007; Fan 2006; Wimmer 2005, Wimmer & Altmann 1999 on vocabulary dynamics).

If text length (words per text) is related to the number of its different words, that ratio is also subject to a linguistic law (Best 2004c, 2007).

15. Discovery of Linguistic Laws VIII: The Logistic (Piotrowski) Law

So far, just regularities in the system and use of language that become noticeable synchronously have been dealt with. However, linguistic change does not proceed chaotically either, but also follows specific laws.

To allow language to change any member of the linguistic community must use an innovation. That can then be adopted by other individuals, but it can also be rejected. Should it be accepted and then be used further, it can spread slowly at the beginning, but then more quickly due to the growing number of contact persons, until it either reaches a tolerated degree of satiation or has become prevalent completely. Thus, a typical course of linguistic change processes comes about. It can be observed again and again, that linguistic change phenomena are tolerated by the linguistic community only up to a certain degree; this e.g. applies to the adoption of words of foreign origin, which alter the vocabulary of a language to a limited degree only (Best & Altmann 1986; Best 2001; 2001b). In texts they usually achieve only low shares in the overall vocabulary (Müller-Hasemann 1983). Other alterations fully prevailed: the 2nd ps. ind. present tense with modal verbs is now in all cases -{st} (darfst, sollst, willst); the former form -{t} (darft, solt, wilt) disappeared completely. Further to that, the form "was" was replaced completely with "war" (Best 1983), and "ward" was replaced widely with "wurde" (Best & Kohlhase 1983). Sometimes, new forms are deleted again from the language; this can be seen by the e-epithesis of the $1^{st}/3^{rd}$ ps. sg. ind. present tense with strong verbs: forms like "floh-e", "sah-e" are no longer acceptable (Imsiepen 1983). Accordingly, three types of linguistic changes are presented: 1. The change where former forms are completely replaced with new ones (= complete linguistic change); 2. The one where new forms prevail to some degree (= incomplete linguistic change) and 3. Such a change where new forms come up, spread and then disappear completely or partly (= reversible linguistic change).

15.1. Complete Linguistic Change

During the discussion of a proposal by Piotrovskaja & Piotrovski (1974) Altmann and others (1983: 107) developed and reviewed the function

$$p_t = \frac{1}{1+ae^{-kt}}$$

that was called "Piotrovski's law" then. That law is a law of growth that has been known as "logistic law" since Verhulst (1838, 1845) and Pearl (1926)[16]. This is

[16] For prehistory see Best & Zhu (2006). Many thanks to R. Schimming, Greifswald, for the relevant information. On the logistic model from the mathematical point of view: Banks 1994: 27ff. The history of statistics in Germany since the 18th century is dealt with by S. Köhler (1994).

the version of the logistic law for a complete linguistic change where all old forms were replaced with new ones. Take the change from "wilt" to "willst" as the first example (for the principles and problems of data acquisition see Best 1983; the following data concerning the modal verbs have been supplemented to Best 1983):

Table 15.1
The replacement of -{t} with -{st} in the 2nd ps. sg. ind. present tense of "wollen"[17]

t	Period	-{t}	-{st}	f_t	p_t
1	1450-1479	853	3	0.0035	0.0039
2	1480-1509	683	1	0.0015	0.0088
3	1510-1539	950	4	0.0042	0.0195
4	1540-1569	792	27	0.0330	0.0427
5	1570-1599	426	45	0.0955	0.0909
6	1600-1629	341	56	0.1411	0.1831
7	1630-1659	421	247	0.3698	0.3343
8	1660-1689	266	332	0.5552	0.5295
9	1690-1719	178	362	0.6741	0.7161
10	1720-1749	69	437	0.8636	0.8497
11	1750-1779	59	558	0.9044	0.9268
12	1780-1809	10	645	0.9847	0.9660
13	1810-1839	7	494	0.9860	0.9845
a = 565.2653 k = 0.8069 D = 0.9966					

f_t: observed portion and p_t: calculated portion of new forms in the relevant period; a, k: parameters. This linguistic change also results in excellent D = 0.99 to follow the formula

$$p_t = \frac{1}{1+565.2653e^{-0.8069t}}.$$

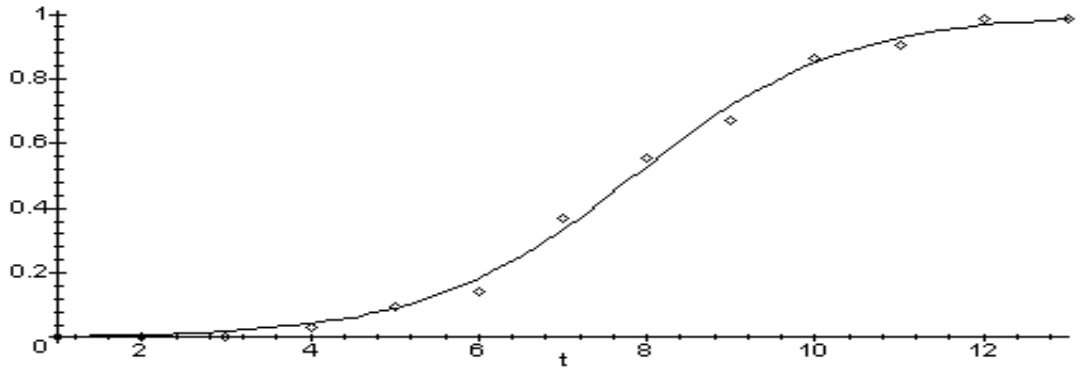


Figure 15.1 Concerning the data in Table 15.1

[17] New calculations for modal verbs were performed by means of NLREG in 2006.

Second example: For the relevant linguistic change from “solt“ to “sollst“ an equally good compliance between theory and observation can be achieved:

Table 15.2
The replacement of -{t} with -{st} in the 2nd ps. sg. ind. present tense of “sollen“

t	Period	-{t}	-{st}	f_t	p_t
1	1420-1449	270	7	0.0253	0.0114
2	1450-1479	865	13	0.0148	0.0202
3	1480-1509	804	7	0.0086	0.0358
4	1510-1539	986	64	0.0610	0.0624
5	1540-1569	825	79	0.0874	0.1068
6	1570-1599	467	126	0.2125	0.1768
7	1600-1629	343	181	0.3454	0.2784
8	1630-1659	313	206	0.3969	0.4093
9	1660-1689	203	198	0.4938	0.5545
10	1690-1719	144	237	0.6220	0.6909
11	1720-1749	33	276	0.8932	0.8006
12	1750-1779	55	334	0.8586	0.8782
13	1780-1809	10	408	0.9761	0.9283
14	1810-1839	5	338	0.9854	0.9588
a = 156.2492		k = 0.5856		D = 0.9852	

Data supplemented to Best 1983. Compliance with D = 0.99 is very high again, which is shown by the diagram below:

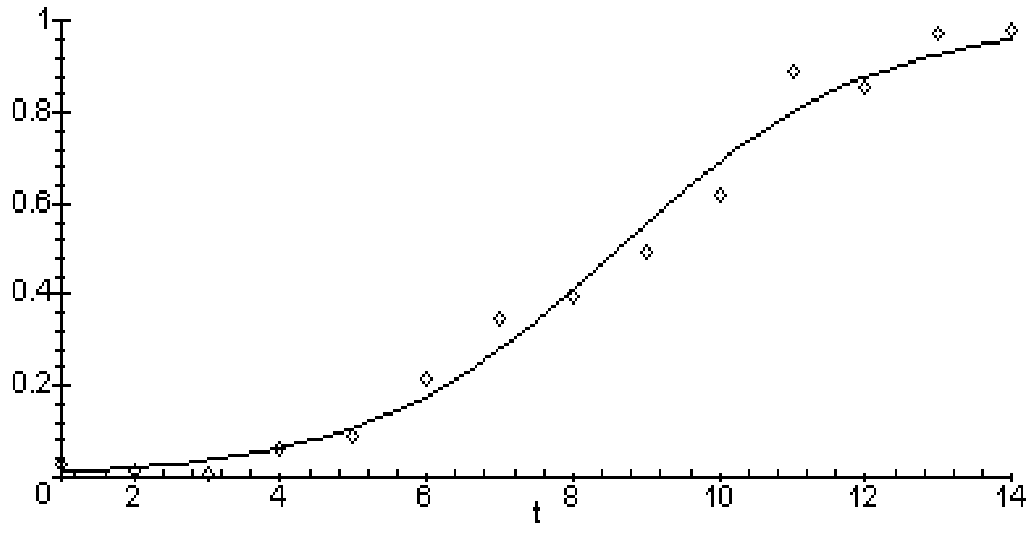


Figure 15.2. Concerning the data in Table 15.2

Third example: The same change from -{t} to -{st} in the 2nd ps. sg. ind. present tense takes place with the verb “dürfen“: “darft“ to “darfst“. The attempt in Best (1983) to prove that this linguistic change follows the logistic law was not suc-

cessful due to lacking satisfactory data. Since then further observations could be collected:

Table 15.3
The replacement of -{t} with -{st} in the 2nd ps. sg. ind. present tense of "dürfen"

t	Period	-{t}	-{st}	f_t	p_t
1	1426-1447	7	3	0.3000	0.2646
2	1448-1469	10	8	0.4444	0.4242
3	1470-1491	35	40	0.5395	0.6013
4	1492-1513	19	47	0.7121	0.7554
5	1514-1535	11	120	0.9160	0.8634
6	1536-1557	-	105	1.0000	0.9283
a = 5.6910		k = 0.7166		D = 0.9592	

Thus it can be shown that this process of linguistic change conforms to the logistic law with a good D = 0.96.

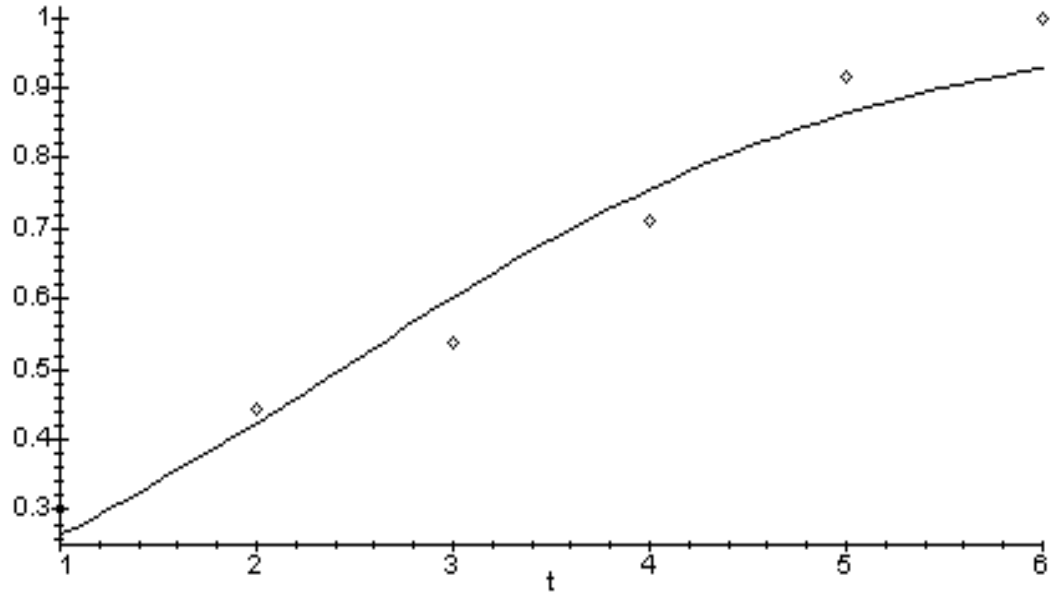


Figure 15.3. Concerning the data in Table 15.3

It is highly remarkable that the replacement of -{t} with -{st} in the 2nd ps. sg. ind. present tense is the same process, but requires a different long time with the different verbs. The change from "ward" to "wurde" turned out to be especially long-drawn-out (Best & Kohlhase 1983). However, in all cases observed the logistic law turns out to be appropriate to model that process.

15.2 Incomplete Linguistic Change

Altmann (1983: 61) developed the function

$$p_t = \frac{c}{1+ae^{-kt}}$$

for the incomplete linguistic change; here, the parameter *c* indicates the limit value approaching the linguistic change.

As the first example of this kind of linguistic change loan word formations ending with -{ität} were analyzed as presented by Eisenberg (1998: 274; based on Munske 1988: 63). They evolved as follows:

Table 15.4
On the spreading of words ending with -{ität}

t	century	n_t	n_t (cumulated)	p_t
1	15.	9	9	9.9257
2	16.	30	39	27.7750
3	17.	29	68	66.6765
4	18.	50	118	122.2096
5	19.	46	164	167.4602
6	20.	31	195	190.3240
a = 60.1664		c = 203.6792	k = 1.1257	D = 0.99

Here $n_{t:}$ is the observed number of loans. The data for the 14th and 15th centuries were summarized. The diagram shows the excellent compliance of observation and model.

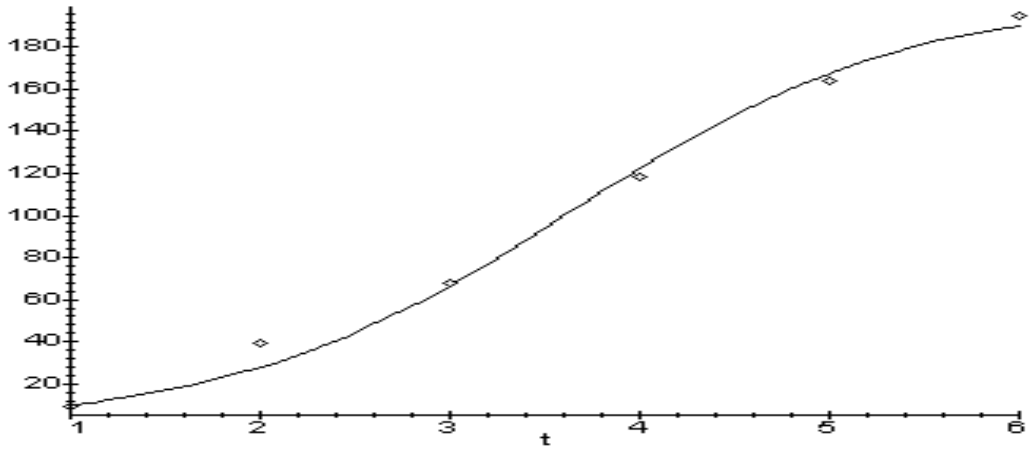


Figure 15.4. Concerning the data in Table 15.4

The model also proved successful for the spreading of words ending with *-ical* (Best 2002) or *-ion* (Körner 2002) and beginning with *therm(o)-* (Best 2005g).

As a further example we'd like to present the process of adoption of Greek words into German; that process was already presented in Best & Altmann (1986) and tested successfully. At this point we'd like to present data from analyses by Kirkness (1991: 302f.); this allows a further test for a different set of data gained from Greek foreign words, which also has the advantage of being more comprehensive.

Table 15.5
The adoption of Greek words into German

t	century	n_t	n_t (cumulated)	p_t
1	15.	24	24	26.7118
2	16.	138	162	96.1267
3	17.	81	243	241.9827
4	18.	128	371	373.3967
5	19.	60	431	427.2100
6	20.	7	438	441.9847
a = 67.7782		c = 446.6496	k = 1.4612	D = 0.97

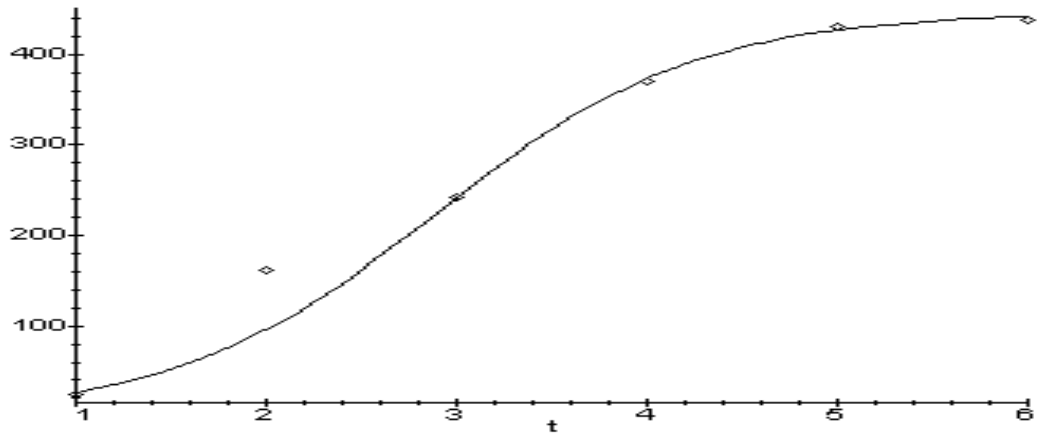


Figure 15.5. Concerning the data in Table 15.5

This is the typical progression of foreign word adoption into a language as it was documented in several cases (Best & Altmann 1986; Best 2001, 2001b, 2003, 2004, 2005k, 2006k; Hentschel 1995; Körner 2004; collection of papers: Kelih & Best 2014). It can also be assumed for the especially disputed influence of English on German at present; that linguistic change begins very slowly, but then becomes active appreciably in the 17th century and can be considered for the present to be the phase of strongest growth unless serious disturbances occur; that development should fade out in some centuries at the earliest (Best 2001, 2003; Körner 2004).

The following two tables and diagrams show the influence of English on German in a comparison with that of Latin and French, as it results from the analysis of two appropriate dictionaries. The following development was found

in Best (2003: 10) on the basis of *Deutsches Fremdwörterbuch* (Kirkness 1988), which was worked out in the first half of the 20th century:

Table 15.6
The adoption of words into German from the Latin, French and English languages

		Latin		French		English	
t	cent.	n_t (cum.)	p_t	n_t (cum.)	p_t	n_t (cum.)	p_t
1	12.	13	14.54	4	0.1338		
2	13.	71	51.59	24	0.8131		
3	14.	186	176.56	30	4.9344		
4	15.	443	539.60	50	29.6364	1	0.5314
5	16.	1379	1257.79	195	167.6518	2	3.3609
6	17.	1902	1995.26	695	715.9256	19	20.5356
7	18.	2390	2383.31	1558	1548.9307	105	103.9995
8	19.	2545	2519.39	1936	1915.4293	287	287.5352
9	20.	2561	2559.97	1971	1992.9770	398	397.7846
a =		634.5436		91325.4018		5127.1609	
c =		2575.9298		2008.9849		428.3872	
k =		1.2812		1.8052		1.8512	
D =		0.9996		0.9994		≈ 1.00	

(acc. to Kirkness 1988; just cumulative values. See Best (2001, 2003, 2004a). In the table the computational result for English only concerns the period from the 15th to the 20th century; in the diagram 0 data were added for the 12th, 13th and 14th centuries each to fit the curve to the two other languages; thus, the parameter a changes to a = 1323712.01. The other values remain almost unaltered).

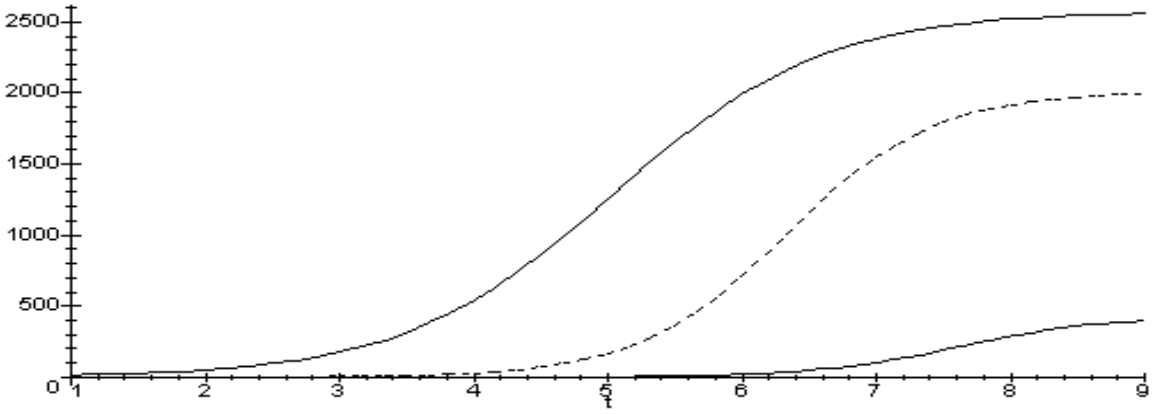


Figure 15.6. Concerning the data in Table 15.6

On the *t*-axis t = 1 stands for the 12th century, t = 2 for the 13th century, etc. The lower line represents English, the one in the middle French, the upper one Latin. The number of the foreign words found is indicated on the *y*-axis.

Körner (2004) presents the trends that can be obtained about 50 years later from the *Duden. Das Herkunftswörterbuch* ([3]2001):

Table 15.7
The adoption of words into German from

		Latin		French		English	
t	cent.	n_t (accum.)	p_t	a = 1323712.01	p_t	n_t (accum.)	p_t
1	8.	2	6.57				
2	9.	3	13.92				
3	10.	4	29.40				
4	11.	268	61.57	5	0.68	1	0.01
5	12.	270	126.86	10	2.34	1	0.03
6	13.	290	252.95	19	8.04	1	0.10
7	14.	347	474.91	30	27.36	1	0.37
8	15.	759	808.46	132	90.26	1	1.38
9	16.	1292	1207.09	271	270.95	3	5.17
10	17.	1548	1571.39	617	646.40	13	19.12
11	18.	1857	1831.26	1099	1081.02	73	68.29
12	19.	2001	1985.66	1363	1343.03	216	217.23
13	20.	2031	2067.63	1424	1444.61	519	518.86
a =		692.6991		7546.0427		562522.44	
c =		2146.2605		1490.6980		1047.4253	
k =		0.7546		1.2374		1.3222	
D =		0.99		0.99		0.99	

Those data were corrected only slightly. In the diagram the calculation of French resulted in a = 308948.71 and that of English in a = 29700660 to correctly position the curves of these languages of origin in relation to Latin.

Here t in t = 1 stands for the 8th century. The arrangement of the languages is the same as in the previous diagram. The trend line for English is a bit steeper. The comparison of the two dictionaries shows that in the meantime the influence of English has become stronger; however, in the lexicon Latin and French are still represented more strongly.

In his further development of a proposal by Piotrovski, Bektaev & Piotrovskaja (1985: 68ff.) Tuldava (1998: 136ff.) showed that the growth of the dictionary of the Estonian literary language can also be understood as a process following the logistic law in the form of incomplete linguistic change.

Relevant data concerning the growth of the English vocabulary are presented by Wermser (1976: 27; Basis: *OED = The Oxford English Dictionary*), which allows new testing of Tuldava's hypothesis:

Table 15.8
On the growth of the English vocabulary

t	Period	n_t	n_t (cumulative)	p_t
1	before 1450	19518	19518	17194.21
2	1450-1469	985	20503	19345.85
3	1470-1489	1462	21965	21682.91
4	1490-1509	1087	23052	24201.07
5	1510-1529	1466	24518	26890.99
6	1530-1549	3313	27831	29737.95
7	1550-1569	2717	30548	32721.81
8	1570-1589	3807	34355	35817.28
9	1590-1609	6045	40400	38994.59
10	1610-1629	4171	44571	42220.53
11	1630-1649	2867	47438	45459.79
12	1650-1669	3367	50805	48676.47
13	1670-1689	2249	53054	51835.66
14	1690-1709	1977	55031	54904.92
15	1710-1729	1522	56553	57855.56
16	1730-1749	1110	57663	60663.57
17	1750-1769	1670	59333	63310.21
18	1770-1789	1647	60980	65782.17
19	1790-1809	2467	63447	68071.44
20	1810-1829	3209	66656	70174.92
21	1830-1849	4797	71453	72093.75
22	1850-1869	4032	75485	73832.63
23	1870-1889	3374	78859	75399.04
24	1890-1909	887	79746	76802.49
25	1910-1929	448	80194	78053.88
26	1930-1949	221	80415	79164.89
27	1950-1957	56	80471	80147.49
a = 4.7028		c = 86842.3162	k = 0.1493	D = 0.99

Tuldava's hypothesis can also be confirmed by the example of the English language. The diagram again shows an effect that obviously was discovered first by Köhler (1986: 137ff.) as oscillation in the lexicon by the example of the interaction of length and frequency in case of synchronous analysis (for this, see contributions in Hammerl [ed.] 1990.) However, Grzybek & Altmann (2002) consider this to be an artifact. This effect occurs here again as oscillation in the growth of the English vocabulary:

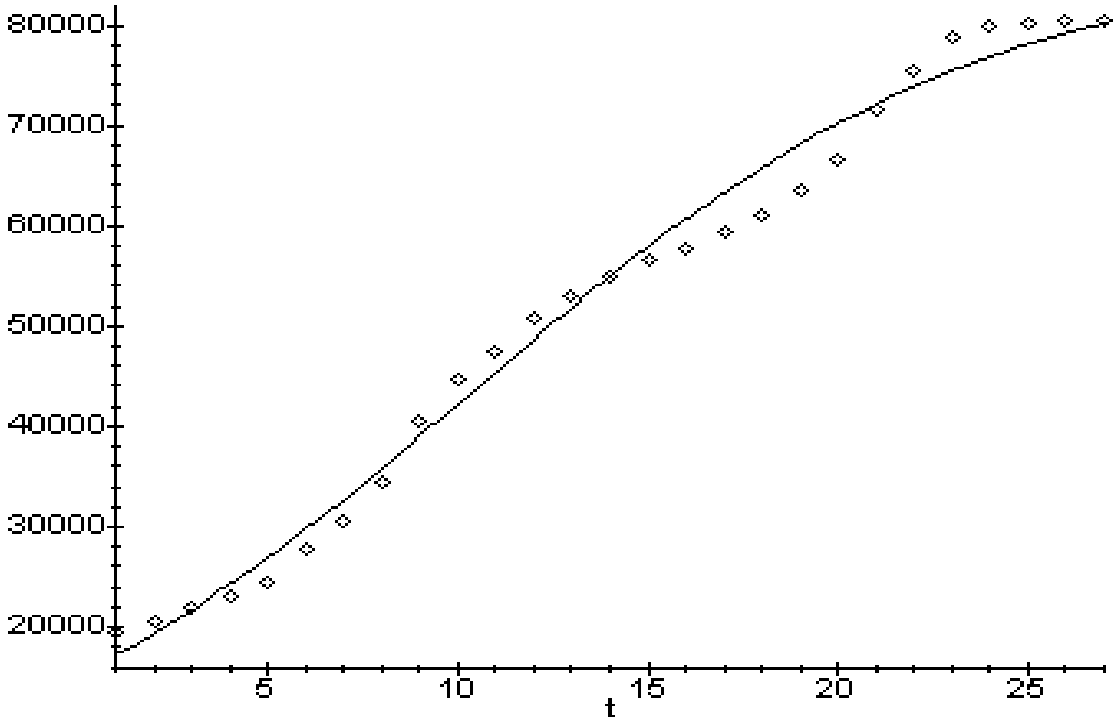


Figure 15.8. Concerning the data in Table 15.8

This is the increase in the vocabulary of the English language as it is presented on the basis of a lexicon (*OED*). The logistic law models the process excellently: $D = 0.99$; the values observed (dots), however, obviously oscillate systematically around the curve calculated (drawn-through line).

However, this subject can also be approached in another way. Wolff (1969: 136) analyzed the words occurring in press texts with respect to the linguistic epoch of the English language from which they originated. Most words already originated from the Old and Middle English periods. The 15345 different words in the analyzed press texts of the 20th century are composed as follows: 7268 words originate from the Old English period, 4237 from the Middle English one, 750 from the late Middle English and further 504 from the 15th century. This means that 12759 of those words were part of the English vocabulary in the 15th century. The increase in the following centuries is relatively low, but follows the logistic law precisely. The following table includes the results from the 15th century onwards:

Table15.9
On the growth of the English vocabulary

t	x	n_t (cumulated)	p_t
1	15th cent.	12759	12785.30
2	16th cent.	14128	14095.89
3	17th cent.	14846	14791.33
4	18th cent.	15077	15133.80
5	19th cent.	15323	15296.25
6	20th cent.	15345	15371.93
a = 0.4521 c = 15436.64 k = 0.7794 D = 0.998			

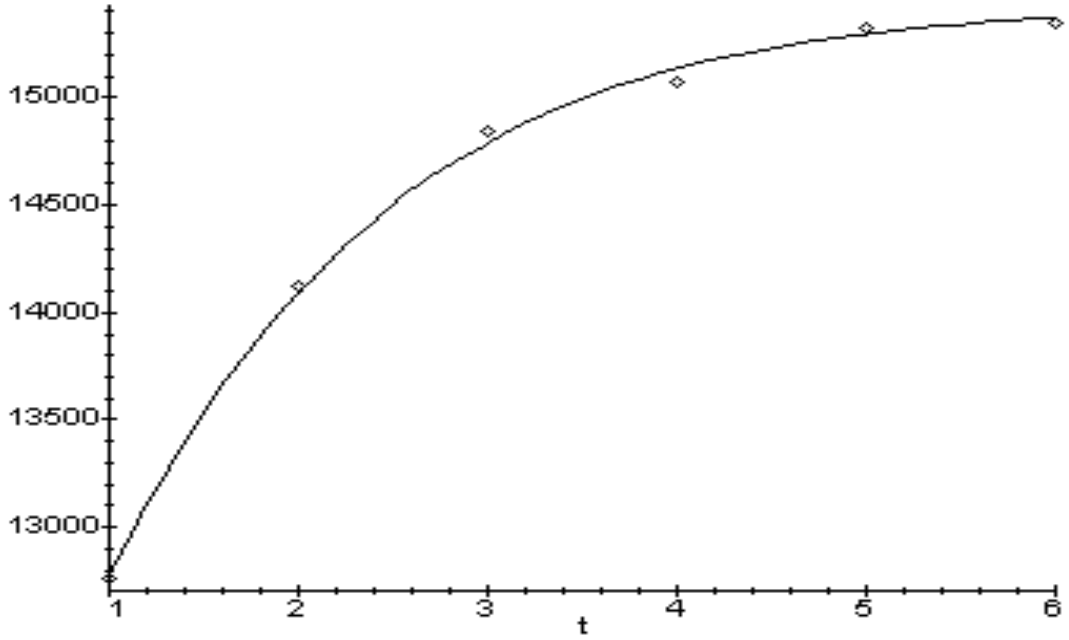


Figure 15.9. Concerning the data in Table 15.9

By the example of the computer vocabulary in German it can be shown that the extension of certain vocabulary components follows the same regularities (Best 2006j).

Alterations to the vocabulary of a language do not only refer to its increase, which certainly is an especially remarkable characteristic of the lexical development, but also to its decrease.

Data concerning vocabulary losses in English are mentioned by Dike (1935: 364: "I classified 16,018 obsoletisms: 1126 OE; 3005 ME; 8612 Early Modern; 3275 Later Modern (1660ff.)". To gain a file from that the periodization of the English language acc. to Viereck, Viereck & Ramisch (2002: 70) was taken as the basis. If it is assumed that all later vocabulary losses still belonged to the inventory of the English lexicon at the beginning of the Old English period (about 700), the following overview results:

Table 15.10
On vocabulary losses in English

t	Time	Loss	Still available	pt
1	700		16018	16029.98
6	1200	1126	14892	15981.85
9	1500	3005	11887	11837.73
10.5	1650	8612	3275	3312.24
13.3	1930	3275	0	48.67
	a = 2.1804	c = 16030	k = -1.5889	D = 0.99

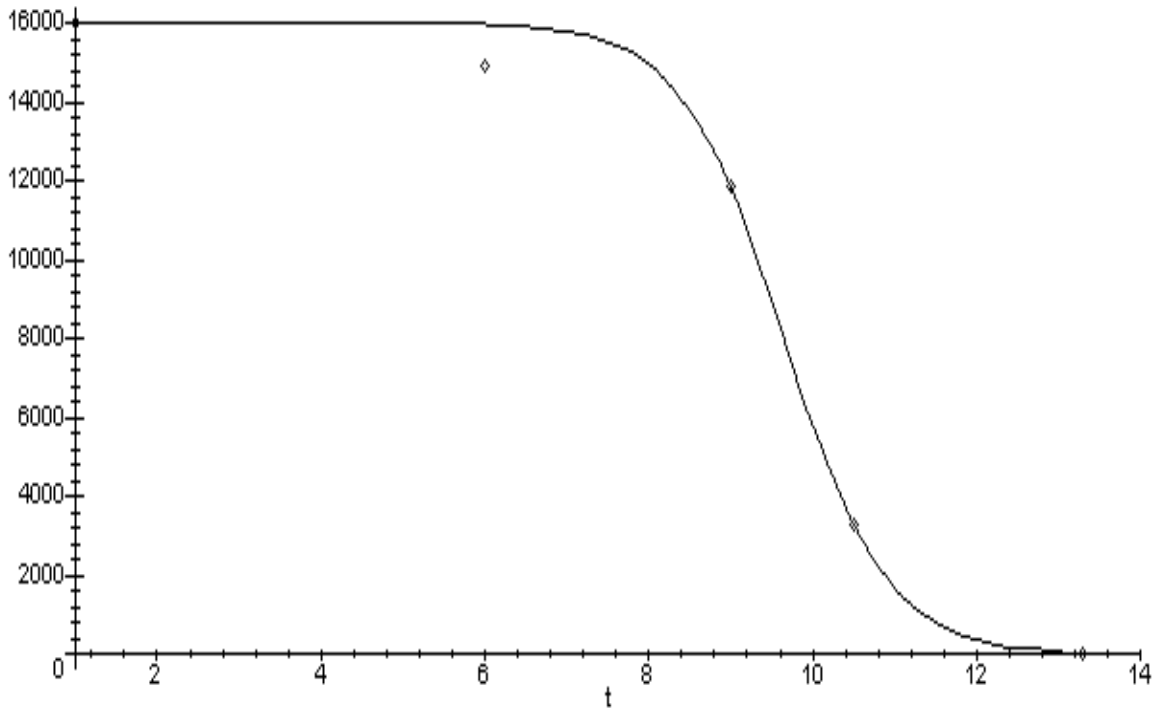


Figure 15.10. Concerning the data in Table 15.10

The loss of individual words can be modeled in a similar way; this succeeded by using the example of "Elektronengehirn" (electronic brain). Wichter (1991) and Busch (2004) present findings concerning the development of the terms "Elektronengehirn", "Computer" and "Rechner" (computer); "Computer" and "Rechner" percolate at first, but lose ground during the so-called public phase (1981-1994) at the latest, probably due to the rivalry of "PC". "Elekronengehirn" is one of the early terms of computer vocabulary; it can already be found in 1950 in the magazine *DER SPIEGEL* (Wichter 1991: 9). Its development is shown by Busch (2004) by means of a *Stern* corpus; the logistic model can be fitted to those data (Best 2006j):

Table 15.11
The development of "Elektronengehirn"

Phase	*t*	obs.	obs.*	calc.
1968-1972	1	13	65.00	64.85
1973-1980	7	11	28.21	28.36
1981-1994	19	2	2.04	0.72
from 1994	26	0	0	0.07
$a = 0.1659$ $c = 80$ $k = -0.3422$ $D = 0.99$				

obs.*: converted such as if 100 texts were evaluated for each phase.

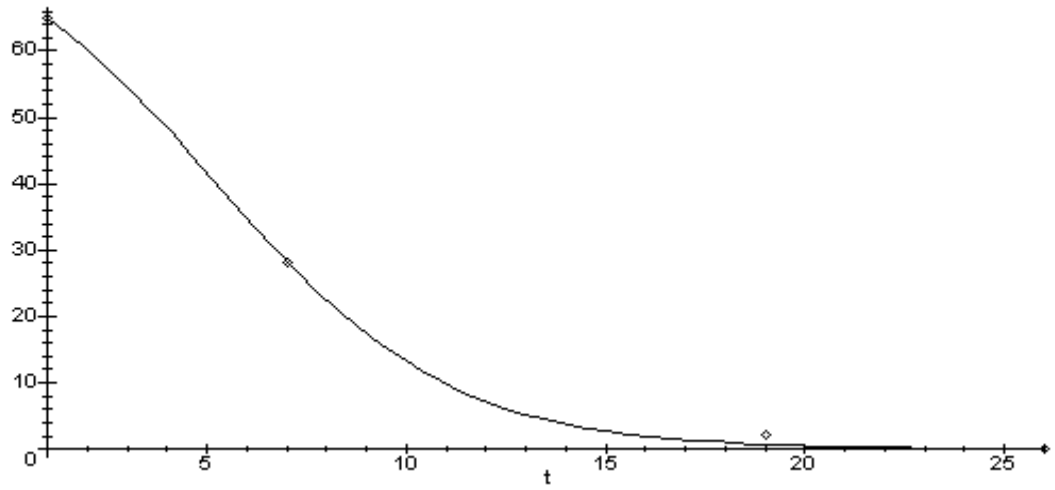


Figure 15.11. Concerning the data in Table 15.11

The observation that words become obsolete again and again and vanish eventually from a language is nothing special (Cherubim 2002; Osman 1992). However, the case of "Elektronengehirn" has been the only single word so far about which we got to obtain data about the course of its disappearance.

Observations concerning the new formation and loss of English words for "Wasserläufe" (water courses) can be found in Rundblad (2000).

The loss rates of vocabularies in different languages have turned out to be variable. The glottochronology attempts to exploit the decay rates to gain information about the degree of relationship of languages. We'd like to refer to the relevant publications by Arapov & Cherc (1983: 35ff.) and Piotrovski, Bektaev & Piotrovskaja (1985: 71ff.) only; for critical comments Altmann et al. (1983: 104-106).

15.3 Linguistic Changes incl. Increase and Decrease (Reversible Linguistic Change)

The third form of linguistic change is shown by an example being presented for the increase of a new phenomenon, which then, however, decreases fully or part-

ly; the reverse case can also be observed. Examples of that can e.g. be observed with the use of Christian names that often come into vogue and are then forced back by other names (Koß 1990: 88). Such mainstream flows can also be observed in other areas. Imsiepen (1983) analyzed the increase and decrease of the e-epithesis in the 1st/3rd person sg. ind. present tense of strong verbs in German. That e-epithesis, however, can also be observed with the auxiliary verb "sein" (to be) with the forms "wase" and "ware". The specialty of that process is that the e-epithesis has disappeared completely with all verbs apart from one and has only prevailed almost completely with "werden" (= to be or to become): „ward" (= was or became) is hardly used and limited to a few applications that are restricted to a purposely ironical or archaic use. Instead, just the form "wurde" is used.

Imsiepen (1983) presented the process of the e-epithesis in two ways: once separated for the growth and decrease phases, once as a homogeneous process. For the overall process Altmann (1983: 61f.) derived

$$p_t = \frac{1}{1+ae^{-kt+ct^2}}$$

as a model. Following Imsiepen (1983) Best, Beöthy & Altmann 1990 established and exemplified processes of data smoothing and also presented a further generalization of that law.

The following presentation deals with the e-epithesis with the auxiliary "sein" (= to be), first separated for the two phases. In both cases the formula can be used for the incomplete linguistic change; the difference is that the parameter *k* is negative when the e-epithesis grows, but positive when it decreases. In a further step the phases of growth and decrease of the epenthetic –e are presented as a homogeneous process.

Let's begin with the growth phase of the epithetic –e (cf. Table 15.12).

Table 15.12
The growth phase of the e-epithesis with the auxiliary "sein": "wase/ware"

t	Period	-{0}	-{e}	f_t	p_t
1	1440-1459	873	4	0.0046	0.0014
2	1460-1479	4514	4	0.0009	0.0020
3	1480-1499	3693	2	0.0005	0.0030
4	1500-1519	3794	11	0.0029	0.0044
5	1520-1539	3419	60	0.0172	0.0064
6	1540-1559	4097	102	0.0243	0.0093
7	1560-1579	3082	62	0.0197	0.0136
8	1580-1599	2738	65	0.0232	0.0198
9	1600-1619	2499	74	0.0288	0.0288
10	1620-1639	2294	80	0.0337	0.0417
11	1640-1659	2733	99	0.0350	0.0601
12	1660-1679	2970	265	0.0819	0.0857
13	1680-1699	2639	414	0.1356	0.1209
		a = 1058.1291	k = 0.3831	D = 0.92	

-{0}: number of forms without epithetic -{e}.

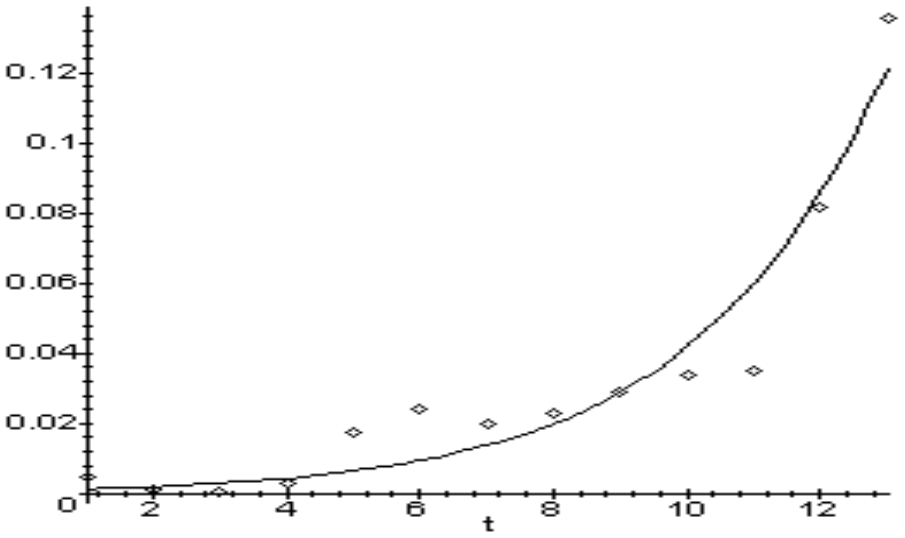


Figure 15.12. Concerning the data in Table 15.12

Table 15.13
The phase of decrease of the e-epithesis with the auxiliary "sein": "wase/ware"

t	Period	-{0}	-{e}	f_t	p_t
1	1700-1719	2644	373	0.1236	0.1444
2	1720-1739	2461	199	0.0748	0.0643
3	1740-1759	3551	85	0.0234	0.0272
4	1760-1779	2620	51	0.0191	0.0113
5	1780-1799	2815	9	0.0032	0.0046
a = 2.4120		k = -0.8989		D = 0.94	

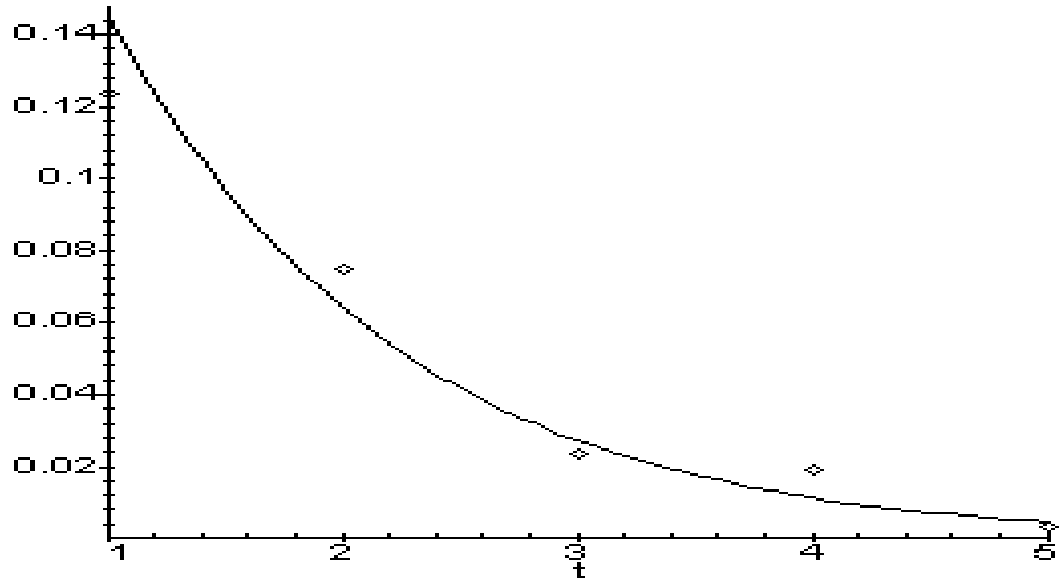


Figure 15.13. Concerning the data in Table 15.13

It can be seen that the decrease of the e-epithesis proceeds much more quickly than its growth. This could already be observed with strong verbs. In contrast to the strong verbs including the e-epithesis in approx. 50% of all forms at the end of the 17th century the auxiliary "sein" does not seem to achieve a portion of 15 % with all relevant forms in the preterit. The e-epithesis can be observed for a really long time with verbs whose radicals end on <h> as it is the case in "sahe" (= saw), "flohe" (fled").

Thus, it is proved that – as it is the case with the replacement of -{t} with -{st} in the 2nd ps. sg. ind. present tense with modal verbs – one and the same process can proceed with different verbs at different times at different intensity and at a different length of time.

If growth and decrease phases of the e-epithesis with the auxiliary "sein": "wase/ware" are modeled as a uniform process, the result is as follows:

Table 15.14
Growth and decrease phases of the e-epithesis with the auxiliary "sein": "wase/ware"

t	Period	-{0}	-{e}	f_t	p_t
1	1440-1459	873	4	0.0046	0.0000
2	1460-1479	4514	4	0.0009	0.0000
3	1480-1499	3693	2	0.0005	0.0000
4	1500-1519	3794	11	0.0029	0.0000
5	1520-1539	3419	60	0.0172	0.0000
6	1540-1559	4097	102	0.0243	0.0001
7	1560-1579	3082	62	0.0197	0.0006
8	1580-1599	2738	65	0.0232	0.0030
9	1600-1619	2499	74	0.0288	0.0110
10	1620-1639	2294	80	0.0337	0.0300
11	1640-1659	2733	99	0.0350	0.0615
12	1660-1679	2970	265	0.0819	0.0957
13	1680-1699	2639	414	0.1356	0.1151
14	1700-1719	2644	373	0.1236	0.1084
15	1720-1739	2461	199	0.0748	0.0796
16	1740-1759	3551	85	0.0234	0.0447
17	1760-1779	2620	51	0.0191	0.0189
18	1780-1799	2815	9	0.0032	0.0060
a = 20275164700		c = 0.1367		k = 3.6224	D = 0.86

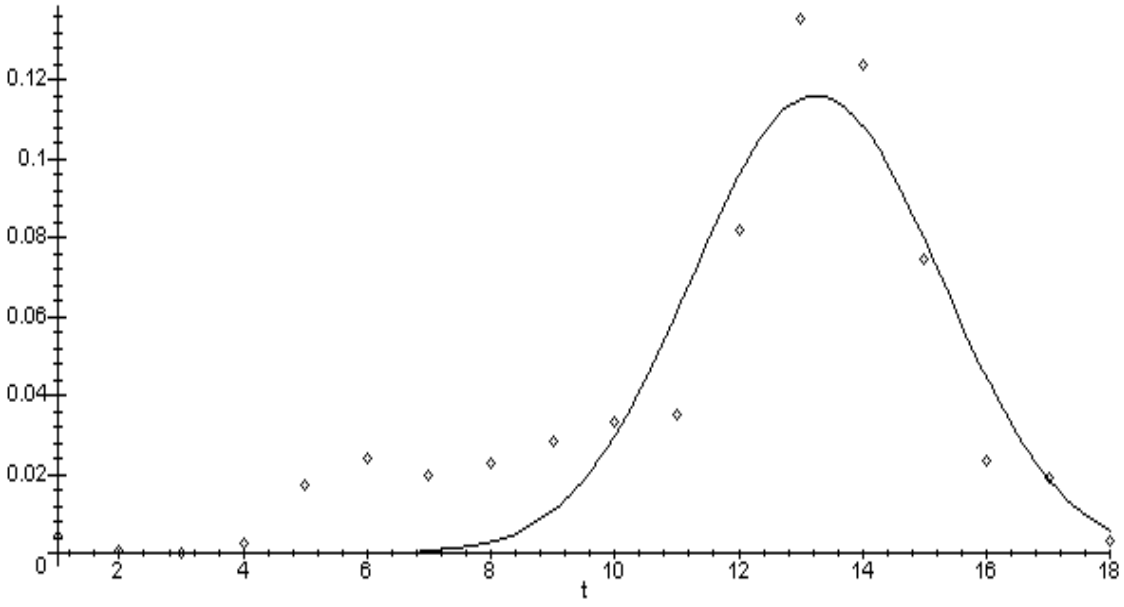


Figure 15.14. Concerning the data in Table 15.14

A reversible linguistic change could also be modeled for the words "Computer" and "Rechner" (= data processor) in Best (2006j).

In the meantime, the logistic law has proved to be successful in its different forms in a lot of cases of intralinguistic changes, loanwords, expansion and reductions of a language (syntactical change: Best 2002a; Busch 2002; Yu 2002); it is an appropriate model for growth as well as decay processes. In some cases, it can also be corroborated for reversible processes that they conform to the expansion of the law. As it is the case with "sein" they can first include growth followed by decay. However, the inverse case – first decay, then growth again – was already observed: in connection with the identification of vowel length in the spelling of German (Best 2003b) and the development of the medium word length in German (Best 2006i; also see Kromer 2006: 207). Interest in the logistic law is also increased by the fact that it is also followed by a language acquisition process of most different kinds (Best 2003a,b, 2006c). This is to be proved by some examples in this chapter.

15.4 The Logistic Law in Language Acquisition

When children start to build up their vocabulary, that process runs the same course that was already presented in connection with the build-up of a language, here the English language. Wagner, Altmann & Köhler (1987: 138 f.) gave a theoretical reason for vocabulary acquisition following the logistic law in its form for incomplete linguistic change. Best (2003a) showed that data concerning vocabulary acquisition of German as well as English children support that assumption of Wagner, Altmann & Köhler; in addition, it is demonstrated that this also applies to other linguistic acquisition processes like the increase in the quantity and speed of speaking.

A further example is offered by Morley (²1967). She analyzed a group of 114 English-speaking children whose linguistic development was normal with respect to the occurrence of the first words, the point of time when the first 2- or 3-word sentences were formed and the time when they developed a mode of speaking being understandable by foreigners as well. Then, the results of that analysis were validated to find out whether they conformed to the logistic law for incomplete linguistic changes. Morley examined that group of children whom she found representative to obtain a comparison with the linguistic acquisition by articulation-disturbed children. Data of individual children (up to three children were concerned) could not be validated. The maximum number of children was used for the parameter *c*.

Table 15.15
First occurrence of individual words with children with undisturbed linguistic development (acc.to Morley ²1967: 433)

t	Period (in months)	Number of children with first word	Number of children with first word (cumulative)	Number of children with first word (calculated)
1	6-8	8	8	15.87
2	9-10	31	39	39.22
3	11-12	42	81	71.20
4	13-14	9	90	94.82
5	15-16	4	94	105.50
6	17-18	9	103	109.26
9	19-24	6	109	110.95
12	25-30	2	111	111.00
a = 19.6335		*c* = 111	*k* = 1.1864	*D* = 0.96

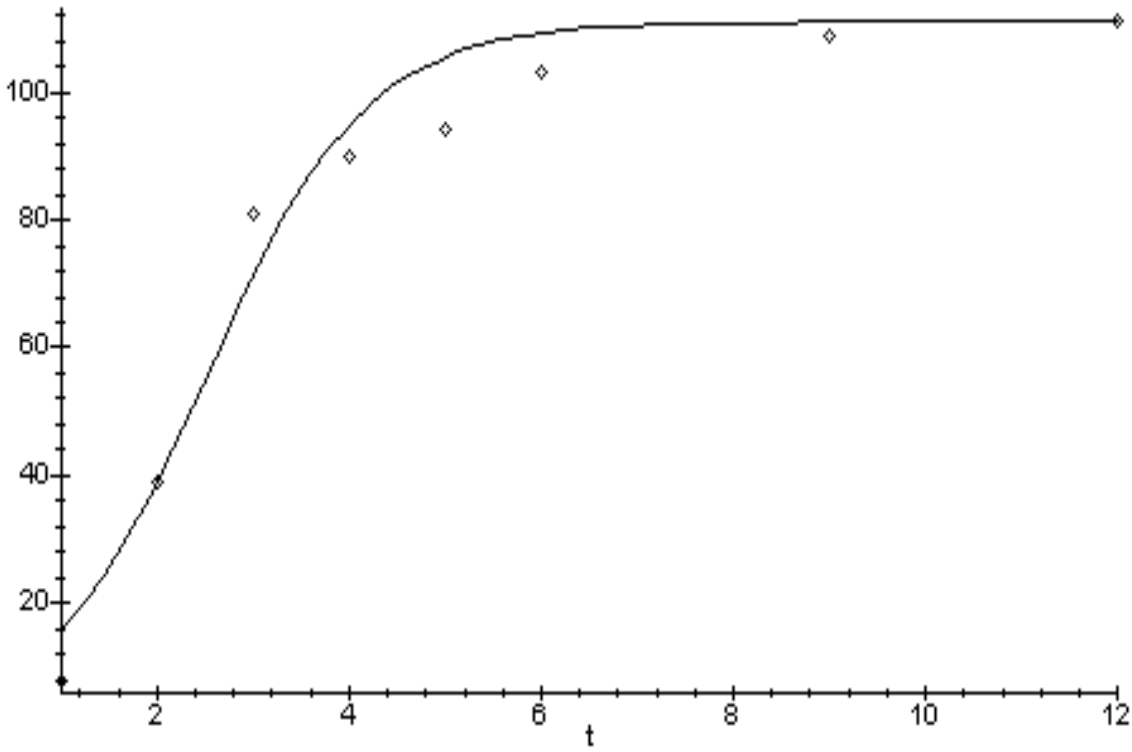


Figure 15.15. Concerning the data in Table 15.15

It can be seen that most children use their first words at the age of nine to twelve months; the development of all children conforms to the logistical law because of the very good D = 0.96.

Table 15.16
First occurrence of children's two- to three-word sentences
with undisturbed language development (acc. to Morley ²1967: 433)

t	Period (in months)	Number of children with first 2- to 3-word sentences	Number of children with first 2- to 3-word sentences (cumulative)	Number of children with first 2- to 3-word sentences (calculated)
2	9-10	1	1	1.67
3	11-12	9	10	4.94
4	13-14	6	16	13.81
5	15-16	9	25	33.61
6	17-18	45	70	63.45
9	19-24	30	100	109.05
12	25-30	5	105	111.89
15	31-36	6	111	112.00
18	37-42	1	112	112.00
$a = 613.82$		$c = 112$	$k = 1.1146$	$D = 0.98$

Most children begin with short sentences in their second year of life; that development again conforms very well to the logistic law with $D = 0.98$.

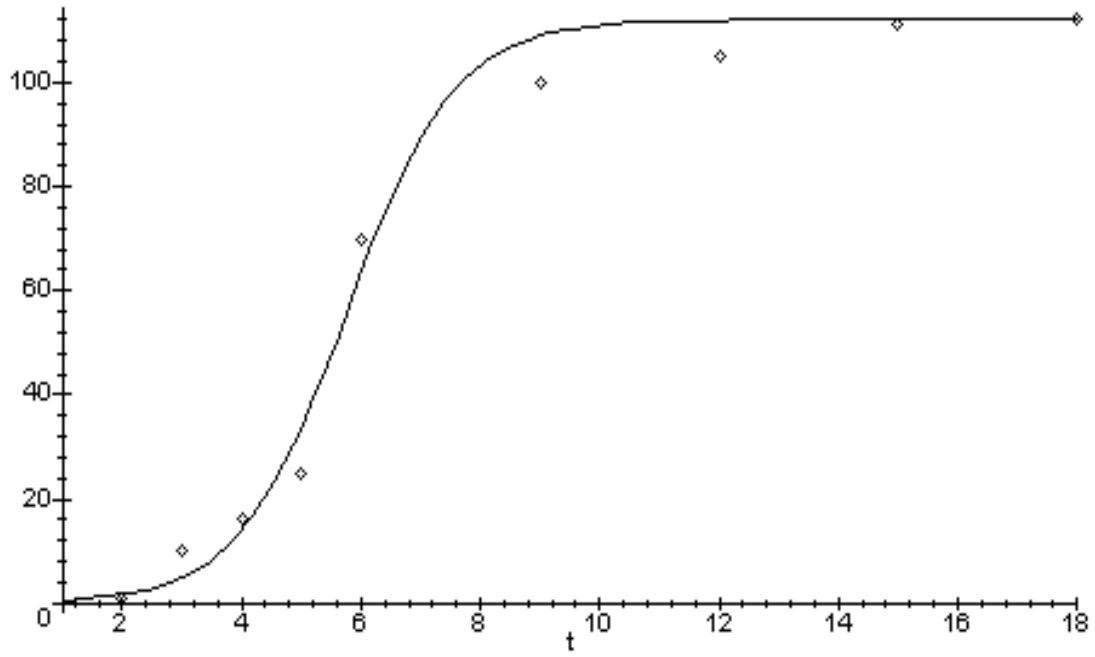


Figure 15.16. Concerning the data in Table 15.16

Table 15.17
First occurrence of speaking by children with undisturbed linguistic development with speaking being understandable to foreigners (acc. to Morley ²1967: 433)

t	Period (in months)	Number of children speaking understandably	Number of children speaking understandably (cumulative)	Number of children speaking understandably (calculated)
2	9-10	1	1	8.66
3	11-12	6	7	12.85
4	13-14	4	11	18.70
5	15-16	7	18	26.49
6	17-18	36	54	36.26
9	19-24	24	78	71.32
12	25-30	9	87	96.51
15	31-36	8	95	106.68
18	37-42	4	99	109.80
21	43-48	2	101	110.67
24	49-54	5	106	110.91
27	55-60	2	108	110.98
37	61-80	3	111	111.00
a = 28.2941		c = 111	k = 0.4365	D = 0.95

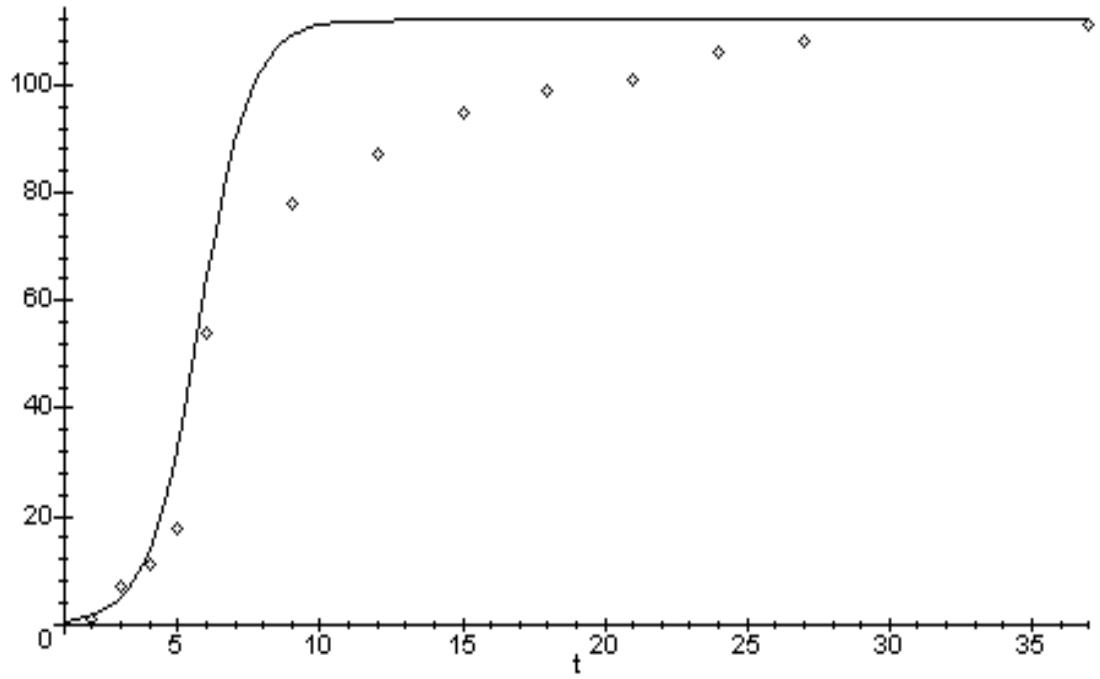


Figure 15.17. Concerning the data in Table 15.17

Again, the logistic law achieves the value $D = 0.95$ and proves to be a good model to describe language acquisition. If the logistic law is tested on the basis of those data presented by Morley (²1967: 434) for articulation-disturbed children, results are even slightly better. Kegel (³1987: 39-41) describes the comparison of both groups of children.

The following graphical presentation permits the comparison of the three different acquisition processes with those children with normal development of language (cf. Figure 15.18):

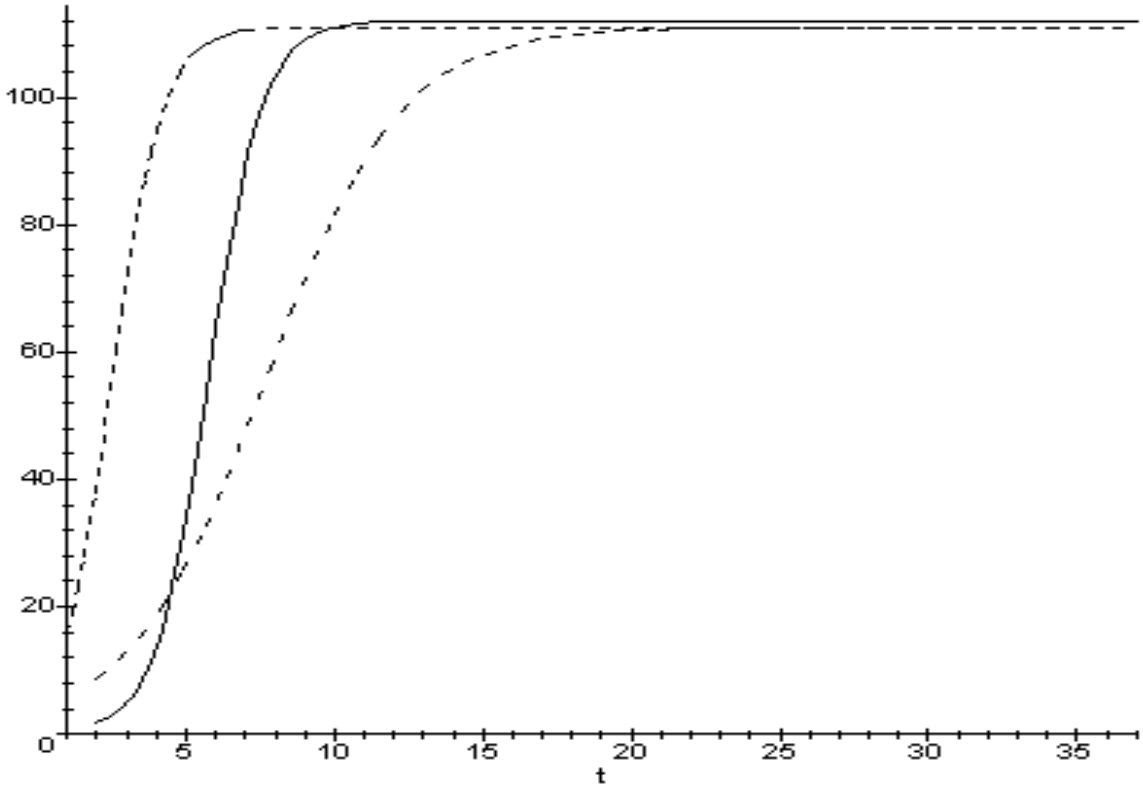


Figure 15.18

The left line shows the use of first words, the uninterrupted line in middle the first 2-and 3-word sentences and the right one the beginning of understandable speaking.

Apart from the acquisition of sounds (Best 2006c) it seems that linguistic acquisition processes run acc. to the logistic law (word, sentence and text lengths or the like); however, states reached in a specific acquisition phase are obviously – at least usually – subjected to the same laws that were also found for the language used by adults (Best 2003a,b; 2006c).

16. Summary

The presentation of the different regularities can only impart a preliminary impression. There is a multitude of analyses concerning the circumstances presented; reference was made to quite a number of them. Though relatively few examples were presented for the individual laws, it should have become clear that despite the multitude of factors influencing the composition of language (system and use of language) it always becomes clear that the effects of laws developed on the basis of theoretical reflections can always be confirmed.

Maybe that with respect to text structuring several continuous regularities become apparent. If frequency distributions of classes of units of the same kind are seen as "horizontal" structuring, they all seem to be governed by the distribution law $P_x = g(x)\ P_{x-1}$, whose first form of $g(x) = a/x$ leads to the Poisson distribution and called the Čebanov-Fucks law (Piotrovski, Bektaev & Piotrovskaja 1985, 254). The specific form of $g(x)$ can appear differently, but the basic regularity is always one and the same. Possibly, distributions of length classes of linguistic units (e.g. words of a different number of syllables) and categories or functions of the same units (e.g. word classes) can be seen as principally of the same structure.

If the rank-frequency distribution of individual units is analyzed instead of classes of units at a specific linguistic level, they are obviously governed by the negative hypergeometric distribution (Phonemes in text corpora: Kaeding, Schulte) or the Zipf-Mandelbrot distribution (letters, lexemes, phonemes in individual texts; cf. lexeme frequencies: Knüppel 1997, Uhlířová 1995). Often, Altmann's model concerning arbitrary rankings (Altmann 1993) proves to be appropriate, sometimes even superior.

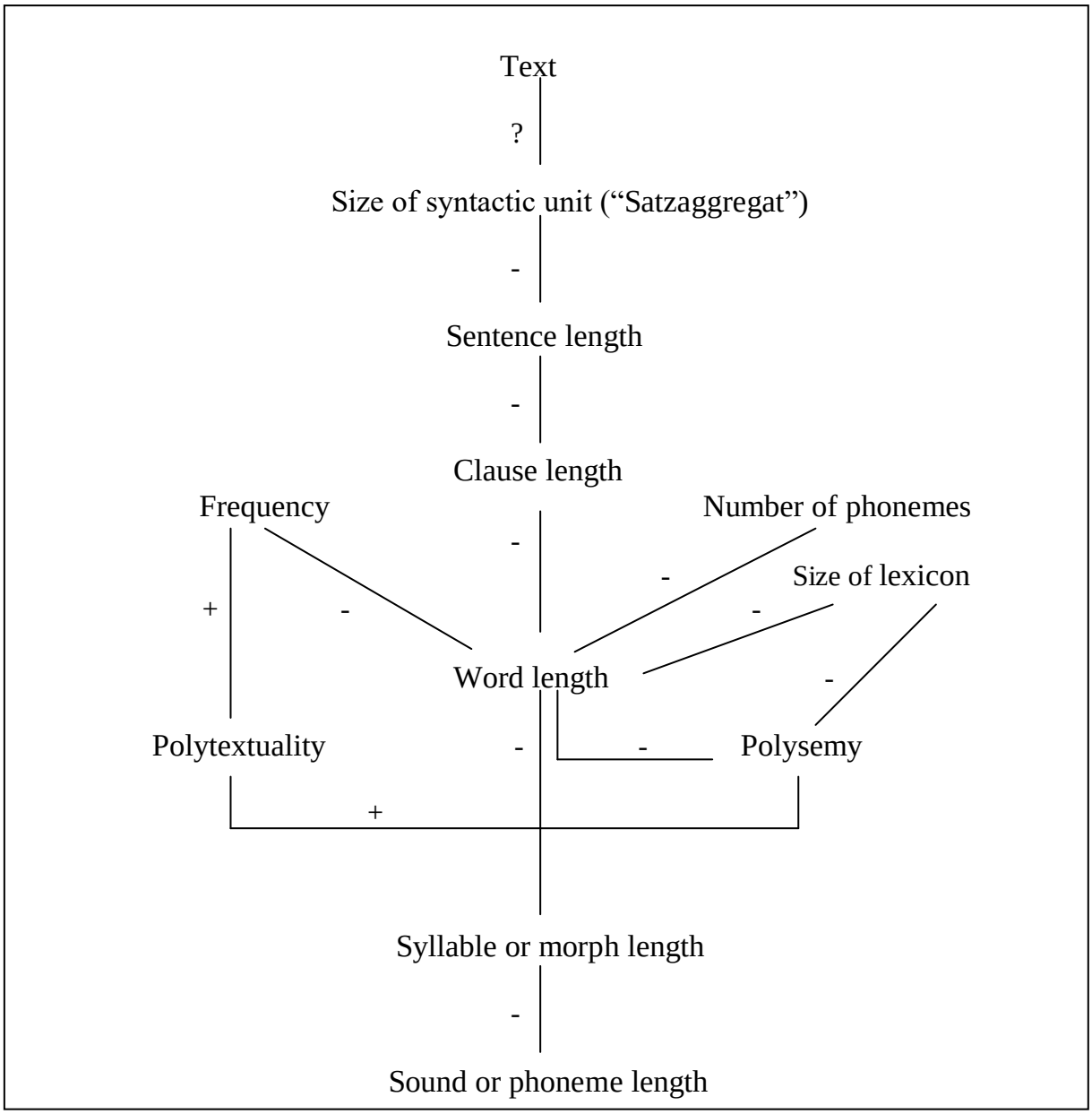
Instead of remaining within one linguistic level it is also possible to analyze the interaction of entities across linguistic levels and also the dependence of the size of a construct on the size of its constituent, a phenomenon being known as the Menzerath-Altmann law (Altmann & Schwibbe 1989). This "vertical" structuring obviously has a very general character. Thus, Hřebíček (1997) using eight chapters of a Turkish text could demonstrate that the Menzerath-Altmann law continuously determines the text structuring from the biggest units to the smallest ones.

Finally, Köhler's conception of "linguistic synergetics" in which a control circuit of linguistic entities is developed and which has an effect partly within one level, partly across the levels provides a further form of the structuring of texts as well as the lexicon (Hoffmann & Krott 2002; Köhler 1986; Köhler 1999) and refers it to the requirements of language to be met by speaker and listener.

The following model integrates the "vertical" structuring of language acc. to the Menzerath-Altmann law (Hřebíček 1997) and Köhler's control circuit (Köhler 1986: 74) influencing the level of words. It shows how system units interact with each other. Minus signs express: the bigger *x*, the smaller *y*; positive signs accordingly mean: the bigger *x*, the bigger *y*. So e.g.: the longer the sentences, the shorter the clauses; however, the shorter the clauses, the longer the words. The higher the frequency of the words, the shorter they are etc. The ques-

tion mark between “text“ and “syntactic unit“ indicates that Hřebíček (1997: 31) does not consider it clarified whether or not syntactic units can be deemed immediate text constituents.

Excerpt of a functional model of language



(This model neglects locally limited deviations as presented on page 103f.)

Explanations concerning the functional model:
“Size of lexicon “: number of lexemes included in the lexicon of the language.
“Number of phonemes “: number of phonemes in the inventory of a language.
“Polytextuality “: number of texts in which a certain word occurs.
“Polysemy”: multiple meanings of a sign

"Syntactic unit": the sentences of a text comprising a certain lexeme (Hřebíček 1997: 31). So, each sentence belongs to a number of syntactic units being identical to the number of lexemes.

All relations provided with the negative or positive sign have been validated empirically and can therefore be deemed rather ascertained. In almost all cases it is about a connection of the form

$$y = ax^b .$$

(On page 103f. there are data where "word" as a construct and "syllable" as a constituent are in a regular connection and where a somewhat more complex form of the law mentioned is used.)

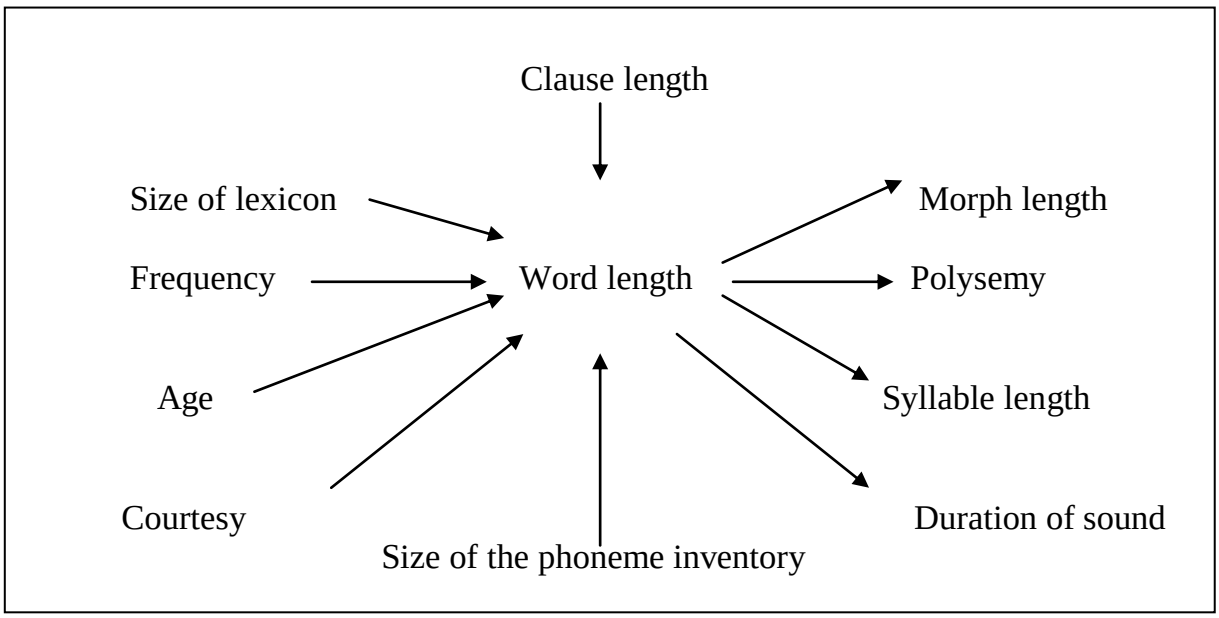
At the level of words Köhler's control circuit is implied in a more simplified way. In that control circuit operators indicating how strong the relevant interaction is are missing. Further to that "system requirements", i.e. the needs that the members of a linguistic community want to have met by the language are missing as well.

A different perspective results when reflections proceed from the lexeme or the word as a central unit of the linguistic system as well as the use of language. In the course of research several law hypotheses were established; the list on p. 29 can be supplemented, amongst others with reference to the age of the units:

- The older words are, the shorter they are and the more frequently they are used (Miyajima 1992: 228; Sanada-Yogo 1999: 244).
- The older words are, the more polysemous they are (Sanada-Yogo 1999: 244f.).

With word length being in the centre the following interaction field can be established (Best 2006g: 27; arrows: assumed influential direction):

Scheme of some influences between word length and other entities



The list can be continued (Baayen 2001: 195; Baker 1950; Saporta 1963: 72, foot note 15); it shows that words are integrated into a multitude of regular relations. More or less comprehensive empirical validations exist for almost all listed hypotheses and some further ones. However, an expansion of empirical data in more languages is urgently wanted in many cases. It can be assumed that other linguistic entities like syntactic constituents, clauses and sentences are in the center of comparably manifold relations, even if evidences are less frequent at present than with words.

What remains to be done is to imagine that at all linguistic levels[18] a similar number of many – or even more – interactions between entities are present, as it is indicated here by using the example of the word level. But even the word level is not complete; therefore, Ziegler & Altmann (2001) pointed out that word length also has an effect on the number of synonyms: the shorter a word, the greater the number of its synonyms.

It will be a future task to validate those concepts further and either verify or falsify them. However, even now some plausibility can be assigned when thinking of the multitude of successful tests with respect to the different laws being found in the previous sections and given literature.

A problem that so far could only be solved partly is the interpretation of the parameters integrated into the mathematical formulation of the different linguistic laws. In several cases it is fully clear what the meaning is. For example, the parameter *c* used with the incomplete change of language can only be understood as the boundary value against which the process analyzed strives.

In other cases there are quite plausible assumptions for the interpretation of the parameters; it e.g. concerns *a* and *b* in the formula

$$g(x) = a/(b+x)$$

(= approach to the hyper-Poisson distribution) as diversification and unification force or as influence by speaker and listener (or as influence by the language community); in many cases acceptable conclusions result. However, the conditions being the basis for *g(x)* taking exactly that form and under which a different one can be determined retrospectively, but – for theoretical reasons - not always in advance.

Example: With respect to word length distributions of many present German texts (measured in syllables) the positive negative binomial distribution can be fitted on the one hand; it can be derived from the approach of

$$g(x) = (a+bx)/\, cx$$

(Best 1997b). Almost the same good results accrue from tests in which the hyper-Poisson distribution is used, as could be found by renewed calculations. Irrespective of the interpretation of the parameters it must be noted that the forms of nu-

[18] For some relations among letter properties see Altmann 2008, 160.

merator and denominator are different in both approaches, though texts are completely identical. At present, it remains one of the unsolved problems that at least for part of the cases the specific form of the approach can be determined only retrospectively, and this is not always unambiguous. Maybe, the already mentioned project implemented in Graz will result in progress (Kelih, Grzybek, & Stadlober 2003).

Finally, linguistic change processes have to be referred to; so far, it could be seen that all processes supported by a sufficient number of data follow one of the forms of the logistic law. This concerns linguistic change of the following kind: loans, expansion and decrease of the vocabulary of a language, alterations to the use of individual lexemes, analogies in morphology, the enforcement of orthographic conventions, phonetic changes as well as various syntactical ones. But the analysis of different processes during the acquisition of the mother tongue by children shows that they almost always conform to the logistic law. Obviously that law has a large range in linguistics (Best 2003a,b, 2006c). In language acquisition it is just the structure of the sound system that can be modeled better by a different law (Best 2006c).

The tested laws have proved successful to a very high degree. But in individual cases it must also be stated that a model may fail. Example: in the Lappish language (Sámi) there are some press texts for which the model used for word length distribution cannot be used (Bartens & Best 1997). Behaghel (1900) analyzed the allomorphs -{e} and -{0} in the dative singular of German masculine and neutral nouns; for them, a model cannot be indicated (Best 2006m) yet. However, such cases are rare to such an extent that they do not mean a problem to the theory. Possible causes may be analytic errors, lack of data, influential factors and the competition by different regularities.

17. Theory

The principal target of each science is to develop theories in its subject of research. They are the prerequisite for explanations. If you want to explain why something is as it is laws are needed. "Without laws there are no theories, and without theories there are no explanations" (Köhler & Altmann 1986: 254; Köhler & Altmann 2005: 30ff.). If there is a law, you can say: With boundary conditions taken into account an issue must be such as it was observed.

A theory can be determined as a system of "law-like statements about the connections and dynamics of the entities of language" (Altmann 1978: 2). This can also be worded in the other way round: "No laws, no science" (Bunge 1967, I: 318), or even more detailed: Where laws cannot be spotted, there is no theory, and where there is no theory, there is no real scientific cognition. A sketch of which conditions are to be met by linguistics as a science was developed by Altmann (1987; 1993). All this is backed by Bunge's opinion (1977: 17): "Everything abides by laws" (Altmann 1985a: 7).

This approach was mainly intended to show by means of some examples that and how you can obtain laws in linguistics. This is no assumption that all proposals presented are laws already. A law becomes a law when it finds its substantiation and systematic integration in the context of a theory and is additionally validated sufficiently. Those conditions are not met by all presented law hypotheses in the same way: some of them are tested rather well, others hardly. As to the Menzerath-Altmann law for example many successful tests exist, as to the Martin's law, however, only a few. In addition, it must always be expected that further empirical analyses result in a revision of findings already deemed quite secured.

A linguistic theory to the effect that is a system of all law hypotheses can only be seen as a remote target. An approach to such a linguistic theory was first presented by Köhler (1986). Already earlier regular interactions between linguistic entities were disclosed, like e.g. Zipf's laws (Prün 1999) and Baker's presentation of a "law of abbreviation" and "law of sense-increase" (Baker 1950). The particular element of Köhler's conception of linguistic synergetics is that a full control circuit was developed; it consists of the connection of several mathematically framed linguistic laws (cf. Köhler 1986: 74). Each of those connections is founded theoretically and successfully tested as to its consequences.

The central idea of linguistic synergetics is self-regulation that implements language in reaction to those requirements it has to meet again and again:

"Permanent changes of the system surroundings as well as system-inherent target conflicts result in permanent changes of the optimum that are only very rarely accepted factually, though the communication system permanently puts forth an effort to achieve its adequacy ideal. As a compromise between disturbances and adaptations as well as competing system needs a steady state equilibrium is developed.

It is this interaction of permanently changing requirements of the system surroundings on the one hand and adaptation reactions of the linguistic system on the other that ... represents the motor of linguistic evolution“ (Leopold 1998: 1).

It was proved in several analyses that this conception is very viable. Köhler who analyzed the German lexis and Hammerl (1991) in his analyses of the Polish lexis came to basically the same results; Sanada-Yogo (1999) applied those principles on the Japanese lexis. Krott (1994) conferred the conception to morphology, Köhler (1999) to syntax. Leopold (1995) discussed whether or not musical texts could be modeled accordingly. Köhler & Martináková-Rendeková (1998) dealt with the synergetic modeling of language and music. Even if not all law hypotheses were tested in the works mentioned, it can be shown already now that ideas developed for German lexis could be conferred to other languages and linguistic levels and also to other communication levels – music in this case – with a reasonable chance of success.

18. Perspectives

The philosopher Julian Nida-Rümelin who lectured in Göttingen at that time commented on basic assumptions in physics in an interview in the journal *Charakter* (year's issue 10, no. 2, February 1996, page 20) as follows: "One of its essential elements is the probabilistic character of the world in which we live. The mechanistic view of the world of the past century is obviously wrong. This has already been taught by quantum physics; quite recently, the chaos theory has supervened."

It is time to add part of this view of the world to our disciplines and react accordingly. The great number of tables, diagrams and mathematical models in this publication may be understood as a plea for that. A stronger consideration of the possibilities offered to the philologies by statistics is urgently required.

Which perspectives can then be nominated for our sciences?

With respect to the remarks in the section "Theory" the following perspectives can be outlined:

1. Even in case of relatively well tested hypotheses a low number from the overall number of all languages has been taken into consideration. It would at least be desirable to test proposed laws on as many typological and/or genetically different languages as possible. An especially widely researched field is that of word length distributions (the project run in Göttingen alone has already analyzed more than 4,000 texts from approx. 50 languages with respect to that problem: approx. 1,400 German, 169 Low German, 16 Palatine and 24 Swiss-German texts, in addition 2,466 texts from foreign languages. Effective date: 2002). However, so far no single language belonging to the language groups in Central and South Africa has been analyzed. Almost the same applies to languages in other regions (Pacific, South America,...).

2. Testing a law hypothesis on further material again and again results in new aspects leading to modifications or even generalizations of the original assumptions. It has already been pointed out that possibly just one law could govern word lengths and word classes. Wimmer & Altmann (2005, 2006) bring forward their arguments exactly in this sense, but are beyond the scope outlined here and show that linguistic laws known so far can be derived from a homogeneous basis.

3. Apart from some others the law of linguistic changes (logistic law) in its different forms has not been tested sufficiently. Further testing is needed. So far, only a few cases of reversible linguistic change have been worked up so well that the modified form of the logistic law could be tested. However, a further problem arises: How can findings concerning linguistic changes be linked to those concerning the structure of the linguistic system? The state of the system is – after all – the result of previous linguistic change processes.

Outlook: Linguistic change produces a multitude of different entities (e.g. the plural allomorphs of German nouns, the distribution of words in word classes, the distribution of the vocabulary according to different languages of origin), which in several cases evidently follow the law of diversification. Leopold

(1998: 99ff.) treats the Piotrowski law (= logistic law) as a model for linguistic fitting processes within the scope of Köhler's theory of linguistic synergetics. The trigger of linguistic change processes is often found in disturbances of an equilibrium state of the system existing that are overcome by fitting processes in a new equilibrium.

4. In almost all areas of literary analysis and linguistic research statistical surveys are already reasonable as tools to come to precise findings, even if that is obviously not recognized always and accepted even less. However, many analyses presenting their subjects in the form of statistical surveys governed by the wish for precision and representativity in such a way that the material gained cannot be used very much and definitely less than this would have been possible when using improved methods.

5. A further aspect of future research must be to continue testing approaches to a linguistic theory as they are found in Köhler's control circuit and mainly develop them such that as many as possible of law hypotheses proposed so far can be integrated. If that conception of linguistic synergetics is followed further, the long-term objective is a linguistic theory that is understood to be a system of interacting laws.

6. Independently, which part they play within the scope of an intended linguistic theory statistical surveys of linguistic phenomena are always useful, either to satisfy humans' curiosity concerning their languages or help put badly founded attitudes – e.g. concerning the question whether certain stylistic features have already been accepted by the linguistic community or the question whether the language could be ultimately doomed – on a better foundation and if possible confirm or revise them. This aspect of quantitative linguistics should be accepted as well.

In a publication already mentioned above (cf. p.4), which could have become trend-setting for the humanities if it had been considered appropriately, we can find the following statement: "The p r o p e r application of the statistical method always comes to results which are remarkable in some way or other... Figures prove that they represent undisputable facts that m u s t be taken into consideration" (Thumb 1911: 3). There is nothing left to be added.

References

Aichele, Dieter (2005). Das Werk von W. Fucks. In: Köhler, Altmann & Piotrowski (Eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch* (pp. 152-158). Berlin: de Gruyter.

Altmann, Gabriel (1978). Towards a theory of language. In: Altmann, Gabriel (Ed.), *Glottometrika 1* (pp. 1-25). Bochum: Brockmeyer.

Altmann, Gabriel (1980). Prolegomena to Menzerath's Law. In: Grotjahn, Rüdiger (Ed.), *Glottometrika 2* (pp. 1-10). Bochum: Brockmeyer.

Altmann, Gabriel (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best & Kohlhase (Eds.), *Exakte Sprachwandelforschung* (pp. 59-90).

Altmann, Gabriel (1985). Semantische Diversifikation. *Folia Linguistica XIX, 177-200.*

Altmann, Gabriel (1985a). Sprachtheorie und mathematische Modelle. Christian-Albrechts-Universität Kiel, *SAIS Arbeitsberichte. H. 8, 1-13.*

Altmann, Gabriel (1987). The levels of linguistic investigation. *Theoretical Linguistics 14, 227-239.*

Altmann, Gabriel (1988). Verteilungen der Satzlängen. In: Schulz, Klaus-Peter (Ed.), *Glottometrika 9* (S. 147-169). Bochum: Brockmeyer.

Altmann, Gabriel (1988a). *Wiederholungen in Texten.* Bochum: Brockmeyer.

Altmann, Gabriel (1991). Modelling diversification phenomena in language. In: Rothe (Ed.), *Diversification Processes in Language: Grammar* (pp. 33-46). Hagen: Rottmann.

Altmann, Gabriel (1992). Das Problem der Datenhomogenität. In: Rieger, Burghard (Ed.), *Glottometrika 13* (pp. 287-298). Bochum: Brockmeyer.

Altmann, Gabriel (1993). Phoneme Counts. In: Altmann, Gabriel (ed.), *Glottometrika 14, 54-68*. Trier: Wissenschaftlicher Verlag Trier.

Altmann, Gabriel (1993a). Science and linguistics. In: Köhler, Reinhard, & Rieger, Burghard B. (Eds.), *Contributions to Quantitative Linguistics* (pp. 3-10). Dordrecht: Kluwer.

Altmann, Gabriel (1995). *Statistik für Linguisten. 2., verb. Aufl.* Trier: Wissenschaft-licher Verlag Trier.

Altmann, Gabriel (1996). Diversification Processes of the Word. In: Schmidt, Peter (Ed.), *Glottometrika 15* (pp. 102-111). Trier: Wissenschaftlicher Verlag Trier.

Altmann, Gabriel (2004). Script complexity. *Glottometrics 8, 68-74.*

Altmann, Gabriel (2005). Diversification processes. In: Köhler, Altmann & Piotrowski (Eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch* (pp. 646-658).

Altmann, Gabriel (2008). Towards a theory of script. In: Altmann, Gabriel, & Fan, Fengxiang (eds.), *Analyses of Script. Properties of Characters and Writing Systems* (pp. 149-164). Berlin/New York: Mouton de Gruyter.

Altmann, Gabriel (2015). *Problems in Quantitative Linguistics 5.* Lüdenscheid: RAM-Verlag.

Altmann, Gabriel, & Best, Karl-Heinz (1996). Zur Länge der Wörter in deutschen Texten. In: Schmidt, Peter (Ed.), *Glottometrika 15* (pp. 166-180), Trier: Wissenschaftlicher Verlag Trier.

Altmann, Gabriel, Best, Karl-Heinz, & Kind, Bernd (1987). Eine Verallgemeinerung des Gesetzes der semantischen Diversifikation. In: Fickermann, Ingeborg (Ed.), *Glottometrika 8* (pp. 130-139). Bochum: Brockmeyer.

Altmann, Gabriel, & Burdinski, Violetta. (1982). Towards a Law of Word Repetitions in Text Blocks. In: Lehfeldt, Werner, & Strauss, Udo. (Eds.), *Glottometrika 4* (pp. 147-167). Bochum: Brockmeyer.

Altmann, G., von Buttlar, H., Rott, W., & Strauß, U. (1983). A law of change in language. In: Brainerd, Barron (ed.), *Historical linguistics* (pp. 104-115). Bochum: Brockmeyer.

Altmann, Gabriel, & Lehfeldt, Werner (1973). *Allgemeine Sprachtypologie.* München: Fink.

Altmann, Gabriel, & Lehfeldt, Werner (1980). *Einführung in die Quantitative Phonologie.* Bochum: Brockmeyer.

Altmann, Gabriel, & Schwibbe, Michael H. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim: Olms.

Altmann, Vivien, & Altmann, Gabriel (2005). *Erlkönig und Mathematik.* http:// ubt.opus.hbz-nrw.de/volltexte/2005/325.

Altmann, Vivien, & Altmann, Gabriel (2008). *Anleitung zu quantitativen Textanalyse. Methoden und Anwendungen.* Lüdenscheid: RAM-Verlag.

Androutsopoulos, Jannis, & Schmidt, Gurly (2002). SMS-Kommunikation: Ethnographische Gattungsanalyse am Beispiel einer Kleingruppe. *Zeitschrift für Angewandte Linguistik 36, 49-79.*

Arapov, Michail Viktorovič, & Cherc, Maja Markovna (1983). *Mathematische Methoden in der historischen Linguistik.* Bochum: Brockmeyer.

Asleh, Laila, & Best, Karl-Heinz (2004/05). Zur Überprüfung des Menzerath-Altmann-Gesetzes am Beispiel deutscher (und italienischer) Wörter. *Göttinger Beiträge zur Sprachwissenschaft 10/11, 9-19.*

Baayen, R. Harald (2001). *Word Frequency Distributions.* Dordrecht/Boston/ London: Kluwer Academic Publishers.

Baker, Sidney J. (1950). The Pattern of Language. *The Journal of General Psychology 42, 25-66.*

Ballod, Matthias (2001). *Verständliche Wissenschaft.* Tübingen: Narr.

Bamberger, Richard, & Vanecek, Erich (1984). *Lesen – Verstehen – Lernen – Schreiben.* Wien: Jugend u. Volk/ Frankfurt: Diesterweg.

Banks, Robert B. (1994). *Growth and Diffusion Phenomena. Mathematical Frameworks and Applications.* Berlin u.a.: Springer.

Bär, Jochen A. (2000). Wörter des Jahres 1999. *Der Sprachdienst 44, 1-20.*

Bär, Jochen A. (2001). Fremdwortprobleme. *Der Sprachdienst 45, 121-133; 169-182.*

Bartens, Hans-Hermann, & Best, Karl-Heinz (1997). Word Length Distribution in Sámi Texts. In: Altmann, Gabriel, Mikk, Jaan, Saukkonen, Pauli, &

Wimmer, Gejza (eds.), *Festschrift in Honour of Juh. Tuldava. Journal of Quantitative Linguistics 4, 45-52.*

Bauer, Friedrich L. (1995). *Entzifferte Geheimnisse. Codes und Chiffren und wie sie gebrochen werden.* Berlin/ Heidelberg: Springer.

Becker, Holger (1995). *Die Wirtschaft in der deutschsprachigen Presse.* Frankfurt: P. Lang (= Diss.phil., Bochum 1988).

Bergenholtz, Henning (1989). Probleme der Selektion im allgemeinen einsprachigen Wörterbuch. In: Hausmann, Franz Josef, Reichmann, Oskar, Wiegand, Herbert Ernst, & Zgusta, Ladislav (Eds.), *Wörterbücher. Bd. 5.1* (pp. 772-779). Berlin/ New York: de Gruyter.

Best, Karl-Heinz (1983). Zum morphologischen Wandel einiger deutscher Verben. In: Best & Kohlhase (Eds.), *Exakte Sprachwandelforschung* (pp. 107-118). Göttingen: Edition herodot.

Best, Karl-Heinz (1990). Die semantische Diversifikation eines Wortbildungsmusters im Frühneuhochdeutschen. In: Hřebíček, Luděk (Ed.), *Glottometrika 11* (pp. 107-110). Bochum: Brockmeyer.

Best, Karl-Heinz (1996). Zur Bedeutung von Wortlängen, am Beispiel althochdeutscher Texte. *Papiere zur Linguistik 55, 141-152.*

Best, Karl-Heinz (1996a). Word Length in Old Icelandic Songs and Prose Texts. *Journal of Quantitative Linguistics 3, 97-105.*

Best, Karl-Heinz (1997). Zur Wortartenhäufigkeit in Texten deutscher Kurzprosa der Gegenwart. In: Best, K.-H. (Ed.), *Glottometrika 16* (pp. 276-285).

Best, Karl-Heinz (1997a). Warum nur: Wortlänge? Nicht nur ein Vorwort. In: Best, K.-H. (Ed.), *Glottometrika 16* (pp. V-XII).

Best, Karl-Heinz (1997b). Zur Wortlängenhäufigkeit in deutschsprachigen Presse-texten. In: Best, K.-H. (Ed.), *Glottometrika 16* (S. 1-15).

Best, Karl-Heinz (Ed..) (1997), *Glottometrika 16.* Trier: Wissenschaftlicher Verlag Trier.

Best, Karl-Heinz (1998). Zur Interaktion der Wortarten in Texten. *Papiere zur Linguistik 58, 83-95.*

Best, Karl-Heinz (1999). Quantitative Linguistik: Entwicklung, Stand und Perspektive. *Göttinger Beiträge zur Sprachwissenschaft 2, 7-23.*

Best, Karl-Heinz (2000). Verteilungen der Wortarten in Anzeigen. *Göttinger Beiträge zur Sprachwissenschaft 4, 37-51.*

Best, Karl-Heinz (2000a). Unser Wortschatz. Sprachstatistische Untersuchungen. In: Hoberg, Rudolf, & Eichhoff-Cyrus, Karin (Eds.), *Die deutsche Sprache zur Jahrtausendwende. Sprachkultur oder Sprachverfall?* (S. 35-52). Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag.

Best, Karl-Heinz (2001). Ein Beitrag zur Fremdwortdiskussion. In: Schierholz, Stefan J. (Ed.), in Zusammenarbeit mit Fobbe, Eilika, Goes, Stefan, & Knirsch, Rainer. *Die deutsche Sprache der Gegenwart. Festschrift für Dieter Cherubim zum 60. Geburtstag* (pp. 263-270). Frankfurt u.a.: Lang.

Best, Karl-Heinz (2001a). Kommentierte Bibliographie zum Göttinger Projekt. In: Best, Karl-Heinz (Ed.), *Häufigkeitsverteilungen in Texten* (pp. 284-310).

Best, Karl-Heinz (2001b). Wo kommen die deutschen Fremdwörter her? *Göttinger Beiträge zur Sprachwissenschaft 5, 7-20.*

Best, Karl-Heinz (2001c). Zur Gesetzmäßigkeit der Wortartenverteilungen in deutschen Pressetexten. *Glottometrics 1, 1-26.*

Best, Karl-Heinz (2001d). Zur Länge von Morphen in deutschen Texten. In: Best, K.-H. (Ed.), *Häufigkeitsverteilungen in Texten* (pp. 1-14).

Best, Karl-Heinz (2001e). Wie viele Morphe enthalten deutsche Wörter? Am Beispiel einiger Fabeln Pestalozzis. In: Ondrejovič, Slavomír & Považaj, Matej (eds.), *Lexicographica '99. Sborník na Počest Kláry Buzássyovej* (pp. 258-270). Bratislava: Veda.

Best, Karl-Heinz (2001f). Probability Distributions of Language Entities. *Journal of Quantitative Linguistics 8, 1-11.*

Best, Karl-Heinz (2001g). Silbenlängen in Meldungen der Tagespresse. In: Best, K.-H. (Ed.), *Häufigkeitsverteilungen in Texten* (pp. 15-32).

Best, Karl-Heinz (2001h). Zur Verteilung rhythmischer Einheiten in deutscher Prosa. In: Best **, K.-H.** (Ed.), *Häufigkeitsverteilungen in Texten* (pp. 162-166).

Best, Karl-Heinz (2001i). Wie viele Wörter enthalten Sätze im Deutschen? Ein Bei-trag zu den Sherman-Altmann-Gesetzen. In: Best, K.-H. (Ed.). *Häufigkeitsverteilungen in Texten* (pp. 167-210).

Best, Karl-Heinz (Ed.) (2001), *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2002). The Distribution of Rhythmic Units in German Short Prose. *Glottometrics 3, 136-142.*

Best, Karl-Heinz (2002a). Satzlängen im Deutschen: Verteilungen, Mittelwerte, Sprachwandel. *Göttinger Beiträge zur Sprachwissenschaft 7, 7-31.*

Best, Karl-Heinz (2002b). Der Zuwachs der Wörter auf *-ical* im Deutschen. *Glottometrics 2, 11-16.*

Best, Karl-Heinz (2003). Anglizismen – quantitativ. *Göttinger Beiträge zur Sprach-wissenschaft 8, 7-23.*

Best, Karl-Heinz (2003a). Zur Entwicklung von Wortschatz und Redefähigkeit bei Kindern. *Göttinger Beiträge zur Sprachwissenschaft 9, 7-20.*

Best, Karl-Heinz (2003b). Spracherwerb, Sprachwandel und Wortschatzwachstum in Texten. Zur Reichweite des Piotrowski-Gesetzes. *Glottometrics 6, 9-34.*

Best, Karl-Heinz (2004). Zur Ausbreitung von Wörtern arabischer Herkunft im Deutschen. *Glottometrics 8, 75-78.*

Best, Karl-Heinz (2004a). Das Fremdwort aus der Sicht der Quantitativen Linguistik. In*:* Sigurd Wichter u. Oliver Stenschke unter Mitarbeit v. Manuel Tants (Eds.), *Theorie, Steuerung und Medien des Wissenstransfers* (pp. 89-99). Frankfurt: Lang.

Best, Karl-Heinz (2004b). Wortschatzwachstum. In: *Wissenstransfer und gesellschaftliche Kommunikation. Festschrift für Sigurd Wichter zum 60. Geburtstag* (S. 333-342). Ed. by Albert Busch & Oliver Stenschke. Frankfurt u.a.: Peter Lang.

Best, Karl-Heinz (2004c). Zum Wortschatzwachstum und -umfang in Texten. *Naukovyj Visnyk Černivec'koho Universytetu: Hermans'ka filolohija. Vypusk 206-207, 31-43.*

Best, Karl-Heinz (2004/05). Laut- und Phonemhäufigkeiten im Deutschen. G*öttinger Beiträge zur Sprachwissenschaft 10/ 11, 21-32.*

Best, Karl-Heinz (2005). Diversifikation der Fremd- und Lehnwörter im Türkischen. *Archív Orientální 73, 291-298.*

Best, Karl-Heinz (2005a). Georg Philip Harsdörffer (1607-1658). *Glottometrics 9, 86-88.*

Best, Karl-Heinz (2005b). Gottfried Wilhelm Leibniz (1646-1716). *Glottometrics 9, 79-82.*

Best, Karl-Heinz (2005c). Karl Marbe (1869-1953). *Glottometrics 9, 74-76.*

Best, Karl-Heinz (2005d). Laut- und Phonemhäufigkeiten im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 10/ 11, 21-32.*

Best, Karl-Heinz (2005e). *LinK: Linguistik in Kürze mit einem Ausblick auf die Quantitative Linguistik. 3., überarb. Aufl.* Skript. Göttingen.

Best, Karl-Heinz (2005f). Ein Modell für das etymologische Spektrum des Wortschatzes. *Naukovyj Visnyk Černivec'koho Universytetu: Hermans'ka filolohija. Vypusk 266, 11-21.*

Best, Karl-Heinz (2005g). Quantitative Linguistik: Ein Plädoyer. In: Altmann, Gabriel, Levickij, Viktor, & Perebyinis, Valentina (eds.), *Problemy kvantytatyvnoï linhvistyky/ Problems of Quantitative Linguistics: 76-88.* Černivci: Ruta.

Best, Karl-Heinz (2005h). Sprachliche Einheiten in Textblöcken. *Glottometrics 9, 1-12.*

Best, Karl-Heinz (2005i). Buchstabenhäufigkeiten im Deutschen und Englischen. *Naukovyj Visnyk Černivec'koho Universytetu. Vypusk 231, 119-127.*

Best, Karl-Heinz (2005j). Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten. *Glottometrics 11, 9-31.*

Best, Karl-Heinz (2005k). Turzismen im Deutschen. *Glottometrics 11, 56-63.*

Best, Karl-Heinz (2005l). Wortlänge. In: Köhler, Altmann & Piotrowski (Eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch* (pp. 260-273).

Best, Karl-Heinz (2005m). Morphlänge. In: Köhler, Altmann & Piotrowski (Eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch* (S. 255-260).

Best, Karl-Heinz (2005n). Satzlänge. In: Köhler, Altmann & Piotrowski (Eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch* (pp. 298-304).

Best, Karl-Heinz (2005o). Längen rhythmischer Einheiten. In: Köhler, Altmann & Piotrowski (Eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein inter-nationales Handbuch* (pp. 208-214).

Best, Karl-Heinz (2006). August Friedrich Pott (1802-1887). *Glottometrics 12, 94-96.*

Best, Karl-Heinz (2006a/2007). Diversifikation bei Eigennamen. Diversifikation bei Eigen-namen. In: Grzybek, Peter, & Köhler, Reinhard (Eds.), *Exact Methods in the Study of Language and Text* (pp. 21-31). Berlin/ New York: Mouton de Gruyter.

Best, Karl-Heinz (2006b). Ernst Wilhelm Förstemann (1822-1906). *Glottometrics 12, 77-86.*

Best, Karl-Heinz (2006c). Gesetzmäßigkeiten im Erstspracherwerb. *Glottometrics 12, 39-54.*

Best, Karl-Heinz (2006d). Jean Paul (1763-1825). *Glottometrics 12, 75-77.*

Best, Karl-Heinz (2006e). Quantitative Untersuchungen zum Niederdeutschen und Niederländischen. *Göttinger Beiträge zur Sprachwissenschaft 13, 51-71.*

Best, Karl-Heinz (2006f). Rhythmische Einheiten im Altgriechischen. *Göttinger Beiträge zur Sprachwissenschaft 13, 73-76.*

Best, Karl-Heinz (2006g). Sind Wort- und Satzlänge brauchbare Kriterien der Lesbarkeit von Texten? In: Wichter, Sigurd, & Busch, Albert (Eds.), *Wissenstransfer – Erfolgskontrolle und Rückmeldungen aus der Praxis* (pp. 21-31). Frankfurt/ M.: Lang.

Best, Karl-Heinz (2006h). Karl Knauer (1906-1966). *Glottometrics 12, 86-94.*

Best, Karl-Heinz (2006i). Wortlängen im Deutschen. *Göttinger Beiträge zur Sprach-wissenschaft 13, 23-49.*

Best, Karl-Heinz (2006j). Zum Computerwortschatz im Deutschen. *Naukovyj Visnyk Černivec'koho Universytetu: Hermans'ka filolohija. Vypusk 289, 10-24.*

Best, Karl-Heinz (2006k). Jiddismen im Deutschen. *Jiddistik Mitteilungen 36, 1-14.*

Best, Karl-Heinz (2006l). Paul Menzerath (1883-1954). *Glottometrics 14, 86-98.*

Best, Karl-Heinz (2006m). Otto Behaghel (1854-1936). *Glottometrics 14, 80-86.*

Best, Karl-Heinz (2007). Wortschatzwachstum und -umfang in Texten und Text-korpora. In: Sibyla Mislovičová (ed.), *Jazyk a jazykoveda v pohybe. Zborník štúdií vychádza na počest' Slavomíra Ondrejoviča pri príležitosti jeho životného jubilea* (pp. 422-437). Bratislava: VEDA, vydavatel'stvo SAV.

Best, Karl-Heinz (2008). Das Fremdwortspektrum im Türkischen. *Glottometrics 17, 8-11.*

Best, Karl-Heinz (2015). *Studien zur Geschichte der Quantitativen Linguistik. Band 1.* Lüdenscheid: RAM-Verlag.

Best, Karl-Heinz, & Altmann, Gabriel (1986). Untersuchungen zur Gesetzmäßigkeit von Entlehnungsprozessen im Deutschen. *Folia Linguistica Historica 7, 31-41.*

Best, Karl-Heinz, Beöthy, Erszébet, & Altmann, Gabriel (1990). Ein methodischer Beitrag zum Piotrowski-Gesetz. In: Hammerl, Rolf (Eds.), *Glottometrika 12* (pp. 115-124). Bochum: Brockmeyer.

Best, Karl-Heinz, & Kohlhase, Jörg (1983). Der Wandel von *ward* zu *wurde*. In: Best & Kohlhase (Eds.), *Exakte Sprachwandelforschung* (pp. 91-102).

Best, Karl-Heinz, & Kohlhase, Jörg (Hrsg.) (1983), *Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte.* Göttingen: edition herodot.

Best, Karl-Heinz, & Kotrasch, Brita (2005). Albert Thumb (1865-1915). *Glottometrics 9, 82-84.*

Best, Karl-Heinz, & Zhu, Jinyang (2001). Wortlängenverteilungen in chinesischen Texten und Wörterbüchern. In: Best, K.-H. (Ed.), *Häufigkeitsverteilungen in Texten* (pp. 101-114).

Best, Karl-Heinz, & Zhu, Jinyang (2006). Sprachwandel im Chinesischen. *Archív Orientální 74, 203-214.*

Beutelsbacher, Albrecht (1997). *Geheimsprachen.* München: Beck.

Biggs, N.L. (1979). The Roots of Combinatorics. *Historia Mathematica 6, 109-136.*

Billmeier, Günther (1969). *Worthäufigkeitsverteilungen vom Zipfschen Typ, überprüft an deutschem Textmaterial.* Hamburg: Buske.

Boettcher, Wolfgang, Herrlitz, Wolfgang, Nündel, Ernst, & Switalla, Bernd (1983). *Sprache. Das Buch, das alles über Sprache sagt.* Braunschweig: Wester-mann.

Bohn, Hartmut (1998). *Quantitative Untersuchungen der modernen chinesischen Sprache und Schrift.* Hamburg: Kovač.

Brainerd, Barron (1972). Article use as an indicator of style among English-language authors. In: Jäger, Siegfried (Ed.), *Linguistik und Statistik* (pp. 11-32). Braunschweig: Vieweg.

Braun, Peter (1998). *Tendenzen in der deutschen Gegenwartssprache*. 4. Aufl. Stuttgart: Kohlhammer.

Bright, T. (1588). *Characterie: An Arte of Shorte, Swifte, and Secrete Writing by Character*. London (zitiert n. Wolff 1969: 211)

Brockhaus, der grosse (1957). Sechzehnte, völlig neu bearb. Aufl. in zwölf Bänden. 12. Bd. Wiesbaden: Brockhaus.

Brockhaus Enzyklopädie (1966ff.). Wiesbaden: Brockhaus.

Brüers, Nina, & Heeren, Anne (2004). Plural-Allomorphe in Briefen Heinrich von Kleists. *Glottometrics 7, 85-90.*

Bünting, Karl-Dieter, & Bergenholtz, Henning (1989). *Einführung in die Syntax.* 2., überarbeitete Aufl. Frankfurt/M.: Athenäum.

Bunge, Mario (1967). *Scientific Research. I, II.* Berlin: Springer.

Bunge, Mario (1977). *Treatise on basic philosophy. Vol. 3: Ontology I: The Furniture of the World.* Dordrecht: Reidel.

Busch, Albert (2000). Sprechen über den Computer in Fachsprachen und Gemein-sprache. Eine Handreichung für die schulische Praxis im Hinblick auf die sprachliche Verarbeitung der Computertechnologie in den Medien. In: Busch, Albert, & Wichter, Sigurd (Eds.), *Computerdiskurs und Wortschatz. Corpusanalysen und Auswahlbibliographie* (pp. 205-279). Frankfurt: Peter Lang.

Busch, Albert (2004). *Diskurslexikologie und Sprachgeschichte der Computertechnologie.* Tübingen: Niemeyer. (Habilschrift, Göttingen 2003)

Busch, Andrea (2002). *Zur Entwicklung der Satzlängen in modernen Fachtexten.* Staatsexamensarbeit, Göttingen.

Busse, Ulrich (1999). „Prithee now, say you will, and go about it.“ *Prithee* vs. *pray you* as Discourse Markers in the Shakespeare Corpus. In: Neumann,

Fritz-Wilhelm, & Schülting, Sabine (eds.), *Anglistentag 1998 Erfurt. Proceedings.* (pp. 485-500). Trier: Wissenschaftlicher Verlag Trier.

Bußmann, Hadumod ([2]1990, [3]2002). *Lexikon der Sprachwissenschaft.* Stuttgart: Kröner.

Cassier, Falk-Uwe (2001). Silbenlängen in Meldungen der Tagespresse. In: Best, K.-H. (Ed.), *Häufigkeitsverteilungen in Texten* (pp. 33-42).

Čebanov, Sergej Grigor'evič (1947). O podčinenii rečevych ukladov 'indo-evropejskoj' gruppy zakonu Puassona. *Doklady Akademii Nauk SSSR. Tom 55/2, 103-106.*

Čech, Radek & Altmann, Gabriel (2011). *Problems in Quantitative Linguistics 3.* Lüdenscheid: RAM-Verlag.

Cherubim, Dieter (2002). Hochton-Archaismen in akademischen Sprachspielen. In: *Archaismen, Archaisierungsprozesse, Sprachdynamik. Klaus-Dieter Ludwig zum 65. Geburtstag* (pp. 73-90). Ed. by Undine Kramer. Frankfurt u.a.: Lang.

Cherubim, Dieter, & Hilgendorf, Suzanne (1998). Sprachverhalten im Alter. Beobachtungen und Diskussionen zum Begriff des Altersstils. In: Fiehler, Reinhard, & Thimm, Caja (Eds.), *Sprache und Kommunikation im Alter* (S. 230-256). Opladen: Westdeutscher Verlag.

Cramer, Irene M. (2005). Das Menzerathsche Gesetz. In: Köhler, Altmann & Piotrowski (Eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch* (S. 659-688).

Crystal, David (1993). *Die Cambridge Enzyklopädie der Sprache.* Frankfurt: Campus.

Deutsches Rechtswörterbuch. Bd. 9: Mahlgericht bis Notrust (1992-1996). Herausgegeben von der Heidelberger Akademie der Wissenschaften. Bearb. v. Heino Speer. Weimar: Böhlau.

Dike, Edwin Berck (1935). Obsolete English words: Some recent views. *The Journal of English and Germanic Philology XXXIV, No. 3, July 1935, 351-365.*

Dittenberger, W. (1881). Sprachliche Kriterien für die Chronologie der Platonischen Dialoge. *Hermes. Zeitschrift für Classische Philologie 16, 321- 345.*

Dittmar, Norbert, & Bredel, Ursula (1999). *Die Sprachmauer. Die Verarbeitung der Wende und ihrer Folgen in Gesprächen mit Ost- und WestberlinerInnen.* Berlin: Weidler.

Drobisch, M.V. (1866). Ein statistischer Versuch über die Formen des lateinischen Hexameters. In: *Königlich-Sächsische Gesellschaft, Philosophisch-Historische Klasse, Berichte über die Verhandlungen 18, 73-139.*

Duden. Deutsches Universalwörterbuch. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag [2]1989/[4]2001.

***Duden. Das Große Wörterbuch der deutschen Sprache*.** 10 vols. 3rd, völlig neu bearb. u. erw. Aufl. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag [3]1999; [2]1993-95.

***Duden. Das Herkunftwörterbuch.* 3**., völlig neu bearb. u. erw. Aufl. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag [3]2001.

Dshurjuk, T.V., & Levickij, V.V. (2003). Satztypen und Satzlängen im Funktional- und Autorenstil. *Glottometrics 6, 40-51.*

Eco, Umberto (1997). *Die Suche nach der vollkommenen Sprache*. München: dtv.

Eggers, Hans (1962). Zur Syntax der deutschen Sprache der Gegenwart. *Studium Generale 15, 49-59.*

Eggers, Hans (1973). *Deutsche Sprache im 20. Jahrhundert.* München: Pieper.

Eisenberg, Peter (1998). *Grundriss der deutschen Grammatik. Bd. 1: Das Wort*. Stuttgart/ Weimar: Metzler.

Fan, Fengxiang (2006). Models for dynamic inter-textual type-token relationship. *Glottometrics 12, 1-10.*

Fenk-Oczlon, Gertraud (1997). Thesen zu einer natürlichen Typologie. *Papiere zur Linguistik 56, 107-116.*

Fenk-Oczlon, Gertraud (2001). Familiarity, information flow, and linguistic form. In: Bybee, Joan, & Hopper, Paul (eds.), *Frequency and the Emergence of Linguistic Structure* (pp. 431-448). Amsterdam/ Philadelphia: Benjamins.

Fickermann, I., Markner-Jäger, B, & Rothe, U. (1984). Wortlänge und Bedeutungskomplexität. In: Boy, J., & Köhler, R. (eds.), *Glottometrika 6* (pp. 115-126). Bochum: Brockmeyer.

Finkenstaedt, Thomas, & Wolff, Dieter (1973). *Ordered Profusion. Studies in Dictionaries and the English Lexicon with Contributions* by H. Joachim Neuhaus *and* Winfried Herget. Heidelberg: Winter.

Förstemann, Ernst (1846). Ueber die numerischen Lautverhältnisse im Deutschen. *Germania. Neues Jahrbuch der Berlinischen Gesellschaft für deutsche Sprache und Alterthumskunde. 7. Bd., 83-90.*

Förstemann, Ernst (1852). Numerische Lautverhältnisse im Griechischen, Lateinischen und Deutschen. In*: Zeitschrift für vergleichende Sprachforschung [= Kuhns Zeitschrift] 1, 163-179.*

Fucks, Wilhelm (1955). Theorie der Wortbildung. *Mathematisch-Physikalische Semesterberichte. Bd. 4, 195-212.*

Fucks, Wilhelm (1955a). Unterschied des Prosastils von Dichtern und anderen Schriftstellern. *Sprachforum 1, 233-244.*

Fucks, Wilhelm (1968). *Nach allen Regeln der Kunst*. Stuttgart: Deutsche Verlags-Anstalt.

Gabelentz, Georg von der (1894). Hypologie [= Typologie] der Sprachen, eine neue Aufgabe der Linguistik. *Indogermanische Forschungen 4, 1-7.*

Gadol, Joan (1969). *Leon Battista Alberti. Universal Man of the Early Renaissance*. Chicago/ London: The University of Chicago Press.

Gardt, Andreas (1994). *Sprachreflexion in Barock und Frühaufklärung. Entwürfe von Böhme bis Leibniz.* Berlin/ New York: de Gruyter.

Gardt, Andreas (1999). *Geschichte der Sprachwissenschaft in Deutschland. Vom Mittelalter bis ins 20. Jahrhundert.* Berlin/ New York: de Gruyter.

Gerlach, Rainer (1982). Zur Überprüfung des Menzerath'schen Gesetzes in der Morphologie. In: Lehfeldt, Werner, & Strauss, U. (Eds.), *Glottometrika 4* (pp. 95-113). Bochum: Brockmeyer.

Glück, Helmut (Hrsg.) (1993). *Metzler Lexikon Sprache.* Stuttgart/ Weimar: Metzler.

Goebl, Hans (2004). Sprache, Sprecher und Raum: Eine kurze Darstellung der Dialektometrie. Das Fallbeispiel Frankreich. *Mitteilungen der Österreichischen Geographischen Gesellschaft, 146. Jg. (Jahresband), Wien 2004, 247-286.*

Goebl, Hans (2005). Dialektometrie. In: Köhler, Altmann & Piotrowski (Eds.). *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch* (pp. 498-531). Berlin/ New York: de Gruyter

Goethe-Wörterbuch (1978ff.). Bd. I, II: Hrsg. von der Akademie der Wissenschaften der DDR, der Akademie der Wissenschaften in Göttingen und der Heidelberger Akademie der Wissenschaften. Bd. III: Hrsg. von der Berlin-Brandenburgischen Akademie der Wissenschaften, der Akademie der Wissenschaften in Göttingen und der Heidelberger Akademie der Wissenschaften. Stuttgart/ Berlin/ Köln/ Mainz: Kohlhammer 1978, 1989, 1998.

Greenberg, Joseph H. (1960). A Quantitative Approach to the Morphological Typology of Languages. *International Journal of American Linguistics 26, 178-194.*

Grimm, Jacob, & Grimm, Wilhelm (1852-1960). *Deutsches Wörterbuch*. Leipzig: Hirzel (Quellenverzeichnis 1971). Reprint: München: dtv 1984.

Groeben, Norbert (1982). *Leserpsychologie: Textverständnis – Textverständlichkeit.* Münster: Aschendorff.

Große Konkordanz zur Luther-Bibel. 2., neubearb. Aufl. Stuttgart: Calwer Verlag 1989.

Grote, Andrea, & Schütte, Daniela (2000). Entlehnung und Wortbildung im Computerwortschatz – neue Wörter für eine neue Technologie. In: Busch, Albert, & Wichter, Sigurd (Eds.), *Computerdiskurs und Wortschatz. Corpusanalysen und Auswahlbibliographie* (pp. 27-124). Frankfurt: Peter Lang.

Grote, Andrea, & Schütte, Daniela (2000a). Wortschatzverzeichnis: Wortbildung und Entlehnung im Computerdiskurs. In: Busch, Albert, & Wichter, Sigurd (Eds.), *Computerdiskurs und Wortschatz. Corpusanalysen und Auswahlbibliographie* (pp. 347-450). Frankfurt: Peter Lang.

Grotjahn, Rüdiger (1979). *Linguistische und statistische Methoden in Metrik und Textwissenschaft.* Bochum: Brockmeyer.

Grotjahn, Rüdiger, & Altmann, Gabriel (1993). Modelling the Distribution of Word Length: Some Methodological Problems. In: Köhler, Reinhard, & Rieger, Burghard B. (eds.), *Contributions to quantitative linguistics* (pp. 141-153). Dordrecht u.a.: Kluwer.

Grzybek, Peter (2001). Kultur-Ökonomie. Zur Häufigkeit textkonstitutiver Elemente. *Wiener Slawistischer Almanach, Sonderband 54, 485-509.*

Grzybek, Peter (2003). Viktor Jakovlevič Bunjakovskij. *Glottometrics 6, 103-106.*

Grzybek, Peter (ed.) (2006), *Contributions to the Science of Text and Language: Word length studies and related issues*. Dordrecht: Springer.

Grzybek, Peter (2006). History and Methodology of Word Length Studies. The State of the Art. In: Grzybek (ed.), *Contributions to the Science of Text and Language: Word length studies and related issues* (pp. 15-90).

Grzybek, Peter, & Altmann, Gabriel (2002). Oscillation in the frequency-length relationship. *Glottometrics 5, 97-107.*

Grzybek, Peter & Kelih, Emmerich (2003). Graphemhäufigkeiten (am Beispiel des Russischen). Teil I: Methodologische Vor-Bemerkungen und Anmerkungen zur Geschichte der Erforschung von Graphemhäufigkeiten im Russischen. *Anzeiger für Slavische Philologie XXXI, 131-162*.

Grzybek, Peter, Kelih, Emmerich & Altmann, Gabriel (2004). Graphemhäufigkeiten (am Beispiel des Russischen). Teil II: Modelle der Häufigkeitsverteilung. *Anzeiger für Slavische Philologie XXXII, 25-54.*

Guiraud, P. (1954). *Bibliographie critique de la statistique linguistique*. Utrecht/ Anvers: Spectrum.

Guiter, H., & Arapov, M.V. (Eds.) (1982). *Studies on Zipf's Law.* Bochum: Brockmeyer.

Hammerl, Rolf (1990). Untersuchungen zur Verteilung der Wortarten im Text. In: Hřebíček, Luděk (Ed.), *Glottometrika 11* (pp. 142-156). Bochum: Brockmeyer.

Hammerl, Rolf (ed.) (1990). *Glottometrika 12.* Bochum: Brockmeyer.

Hammerl, Rolf (1991). Definition von Grundbegriffen für die Untersuchung von Definitionsfolgen und Lexemnetzen. In: Sambor, Jadwiga, & Hammerl, Rolf (Eds.), *Definitionsfolgen und Lexemnetze. Bd. 1* (pp. 2-12). Lüdenscheid: Richter-Altmann Medienverlag.

Hammerl, Rolf (1991a). Überprüfung des Martingesetzes I an deutschem Sprach-material. In: Sambor, Jadwiga, & Hammerl, Rolf (Eds.), *Definitionsfolgen und Lexemnetze. Bd. 1* (pp. 50-64). Lüdenscheid: RAM.

Hammerl, Rolf (1991b). *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells.* Trier: Wissenschaftlicher Verlag Trier.

Haß-Zumkehr, Ulrike (2001). *Deutsche Wörterbücher.* Berlin/ New York: de Gruyter.

Hentschel, Gerd (1992). Verwendungshäufigkeit und Innovation im Flexionssystem – Beobachtungen zum morphologischen Wandel im Russischen und Polnischen. *Zeitschrift für Slavistik 37, 50-59.*

Hentschel, Gerd (1995). „Zur ‚Seuche' des deutschen Lehnwortes im Polnischen und zu den ‚Selbstheilungskräften' dagegen. In: Bochnakowa, A., & Widłak, S. (eds.), *Munus amicitiae. Studia Linguistica in honorem Witoldi Mańczak k septuagenarii* (S. 69-78). Universitas Iagellonica, Ser. Varia CCCLVI, Cracowiae.

Herdan, Gustav (1966). *The Advanced Theory of Language as Choice and Chance.* Berlin/ Heidelberg/ New York: Springer.

Höhne-Leska, Christel (1975). *Statistische Untersuchungen zur Syntax gesprochener und geschriebener deutscher Gegenwartssprache.* Berlin: Akademie-Verlag.

Hoffmann, Christiane, & Krott, Andrea (2002). Einführung in die Synergetische Linguistik. In: Köhler, R. (Ed..), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik* (pp. 1-29).

Hoffmann, Lothar (1985). *Kommunikationsmittel Fachsprache. Eine Einführung.* Zweite völlig neu bearb. Aufl. Tübingen: Narr.

Hoffmann, Lothar, & Piotrowski, Raimund G. (1979). *Beiträge zur Sprachstatistik*. Leipzig: VEB Verlag Enzyklopädie.

Hollberg, Cecilie (1997). Wortlängenhäufigkeiten in italienischen Pressetexten. In: Best, K.-H. (Ed.), *Glottometrika 16* (pp. 127-137).

Hřebíček, Luděk (1997). *Lectures on Text Theory.* Prague: Academy of Sciences of the Czech Republic, Oriental Institute.

Hřebíček, Luděk (2000). *Variation in Sequences.* Prague: Academy of Sciences of the Czech Republic, Oriental Institute.

Hug, Marc (2000). Partial Disambiguation of Very Ambiguous Grammatical Words. Vortrag am 25.8.2000, QUALICO IV, Prag.

Hug, Marc (2001). Partial Disambiguation of Very Ambiguous Grammatical Words. *Journal of Quantitative Linguistics 7, 217-226.*

Imsiepen, Ulrike (1983). Die e-Epithese bei starken Verben im Deutschen. In: Best & Kohlhase (Eds.), *Exakte Sprachwandelforschung* (pp. 119-141).

Jing, Zhuo (2001). Satzlängenhäufigkeiten in chinesischen Texten. In: Best (Ed.), *Häufigkeitsverteilungen in Texten* (pp. 202-210).

Kaeding, Friedrich Wilhelm (Ed.) (1897/98). *Häufigkeitswörterbuch der deutschen Sprache.* Bd. 1, 2. Berlin-Steglitz: Selbstverlag. Teilabdruck*: Grundlagenstudien aus Kybernetik und Geisteswissenschaften. Bd. 4/ 1963.*

Kaßel, Anja (2002). *Zur Verteilung rhythmischer Einheiten in deutschen und englischen Texten.* Staatsexamensarbeit, Göttingen.

Kegel, Gerd ([3]1987). *Sprache und Sprechen des Kindes.* 3., neubearb. u. erw. Aufl. Opladen: Westdeutscher Verlag.

Kelih, Emmerich, & Best, Karl-Heinz (2014). *Entlehnungen und Fremdwörter: Quantitative Aspekte.* Lüdenscheid: RAM-Verlag.

Kelih, Emmerich, & Grzybek, Peter (2004). Häufigkeiten von Satzlängen: Zum Faktor der Intervallgröße als Einflussvariable (am Beispiel slowenischer Texte). *Glottometrics 8, 23-41.*

Kelih, Emmerich, Grzybek, Peter, & Stadlober, Ernst (2003). Das Grazer Projekt zu Wortlängen(häufigkeiten). *Glottometrics 6, 94-102.*

Kempgen, Sebastian (1995). *Russische Sprachstatistik*. München: Sagner.

Kennedy, Garvin (1985). *Einladung zur Statistik.* Frankfurt/ New York: Campus.

King, Stephen (2000). *Das Lesen und das Schreiben*. O.O.: Ullstein.

Kirkness, Alan (Hrsg.) (1988). *Deutsches Fremdwörterbuch* (1913-1988). Begründet v. Hans Schulz, fortgeführt v. Otto Basler, weitergeführt im Institut für deutsche Sprache. Bd. 7: Quellenverzeichnis, Wortregister, Nachwort. Berlin/New York: de Gruyter.

Kirkness, Alan (1991). Die nationalpolitische Bedeutung der Germanistik im 19. Jahrhundert: Ersetzt statt erforscht – Thesen zu Lehndeutsch, Purismus und Sprach-germanistik. In: Wimmer, Rainer (Ed.), *Das 19. Jahrhundert.*

Sprachgeschichtliche Wurzeln des heutigen Deutsch (pp. 294-306). Berlin/ New York: de Gruyter.

Knauer, Karl (1936). *Ein Künstler poetischer Prosa in der französischen Vorromantik: Jean-François Marmontel. Habilitationsschrift.* Bochum-Langendreer: Druck: Heinrich Pöppinghaus.

Knauer, Karl (1965; [4]1971). Die Analyse von Feinstrukturen im sprachlichen Zeit-kunstwerk. Untersuchungen an den Sonetten Baudelaires. In: Kreuzer & Gunzenhäuser (Eds.), *Mathematik und Dichtung* (pp. 193-210).

Knaus, Marina (2008). Zur Verteilung rhythmischer Einheiten in russischer Prosa. *Glottometrics 16, 57-62.*

Knüppel, Anke (2001). *Untersuchungen zum Zipf-Mandelbrot-Gesetz an deutschen Texten.* In: Best, K.-H. (Ed.), *Häufigkeitsverteilungen in Texten* (pp. 248-280).

Köhler, Reinhard (1984). Zur Interpretation des Menzerathschen Gesetzes. In: Boy, J., & Köhler, Reinhard (eds.), *Glottometrika 6* (pp. 177-183). Bochum: Brockmeyer.

Köhler, Reinhard (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Köhler, Reinhard (1995). *Bibliography of Quantitative Linguistics*. With the Assistance of Christiane Hoffmann. Amsterdam: J. Benjamins.

Köhler, Reinhard (1999). Syntactic Structures: Properties and Interrelations. *Journal of Quantitative Linguistics 6, 46-57.*

Köhler, Reinhard (2001). The Distribution of Some Syntactic Construction Types in Text Blocks. In: Uhlířová, Ludmila, Wimmer, Gejza, Altmann, Gabriel, & Köhler, Reinhard (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in Honour of Luděk Hřebíček* (pp. 136-148). Trier: Wissenschaftlicher Verlag Trier.

Köhler, Reinhard (2005). Gegenstand und Arbeitsweise der Quantitativen Linguistik. In: Köhler, Altmann & Piotrowski (Eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch* (pp. 1-16).

Köhler, Reinhard (Ed.) (2002). *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik.* http://ubt.opus.hbz-nrw.de/ voll-texte/2004/279.

Köhler, Reinhard, & Altmann, Gabriel (o.J.). Einführung in die Quantitative Linguistik. http://www.ldv.uni-trier.de/download.php?id=223795,25,1

Köhler, Reinhard, & Altmann, Gabriel (1986). Synergetische Aspekte der Lin-guistik. *Zeitschrift für Sprachwissenschaft 5, 253-265.*

Köhler, Reinhard, & Altmann, Gabriel (2005). Aims and Methods of Quantitative Linguistics. In: Altmann, Gabriel, Levickij, Viktor, & Perebyinis, Valentina (eds.), *Problemy kvantytatyvnoï linhvistyky/ Problems of Quantitative Linguistics* (pp. 12-41). Černivci: Ruta.

Köhler, Reinhard & Altmann, Gabriel (2009). *Problems in Quantitative Linguistics 2.* Lüdenscheid: RAM-Verlag.

Köhler, Reinhard & Altmann, Gabriel (2014). *Problems in Quantitative Linguistics 4.* Lüdenscheid: RAM-Verlag.

Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Eds.) (2005). *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch.* Berlin/ New York: de Gruyter.

Köhler, Reinhard, & Martináková-Redenková, Zuzana (1998). A systems theoretical approach to language and music. In: Altmann, Gabriel, & Koch, Walter A. (eds.), *Systems. New Paradigms for the Human Sciences* (pp. 514-546). Berlin/ N.Y.: de Gruyter.

Köhler, Sybilla (1994). *Statistiker und Statistik. Zur Genese der statistischen Disziplin in Deutschland zwischen dem 18. und 20. Jahrhundert.* Diss.phil., Dresden.

König, Werner ([15]2005). *dtv-Atlas zur deutschen Sprache.* München: dtv.

Körner, Helle (2002). Der Zuwachs der Wörter im Deutschen auf *-ion*. *Glottometrics 2, 82-86.*

Körner, Helle (2004). Zur Entwicklung des deutschen (Lehn-)Wortschatzes. *Glottometrics 7, 25-49.*

Koß, Gerhard (1990). *Namenforschung.* Tübingen: Niemeyer.

Kotrasch, Brita (2004/2005). Albert Thumb – Sein Leben und sein Werk. Das ‚Handbuch der neugriechischen Volkssprache' in seinen Briefen. *Göttinger Beiträge zur Byzantinischen und Neugriechischen Philologie, H. 4/5, 121-149.*

Kreuzer, Helmut, & Gunzenhäuser, Rul (Ed.) ([4]1971). *Mathematik und Dichtung.* München: Nymphenburger

Kromer, Victor V. (2006). About Word Length Distribution. In: Grzybek, Peter (ed.), *Contributions to the Science of Text and Language: Word length studies and related issues* (S. 199-210). Dordrecht: Springer.

Krohn, Dieter (1992). *Grundwortschätze und Auswahlkriterien.* Göteborg: Acta Uni-versitatis Gothoburgensis.

Krott, Andrea (1994). *Ein funktionalanalytisches Modell der Wortbildung.* Magisterarbeit, Trier.

Krupa, Viktor (1965). On Quantification of Typology. *Linguistics 12, 31-36.*

Labov, William (1976). Kontraktion, Tilgung und inhärente Variation der Kopula im Englischen. In: William Labov, *Sprache im sozialen Kontext* (S. 90-157). Ed. by Dittmar, Norbert, & Rieck, Bert-Olaf. Kronberg: Scriptor.

Leibniz, Gottfried Wilhelm (1666/ 1962). Dissertatio de arte combinatoria. In: Leibniz, G.W. (1962), Mathematische Schriften, Hrsg. v. c. I. Gerhard. Bd. V: *Die mathematischen Abhandlungen* (pp. 1-79). Hildesheim: Olms.

Léon, Jaqueline, Loiseau, Sylvain (eds.) (2016). *History of Quantitative Linguistics in France.* Lüdenscheid: RAM-Verlag

Leopold, Edda (1995). Ein synergetisches Modell für musikalische Texte. In: Boroda, M.G. (ed.), *Musikometrika 6* (pp. 117-125). Bochum: Brockmeyer.

Leopold, Edda (1998). *Stochastische Modellierung lexikalischer Evolutionsprozesse.* Hamburg: Kovač.

Lindell, Anneli, & Piirainen, Ilpo Tapani (1980*). Untersuchungen zur Sprache des Wirtschaftsmagazins „Capital".* Vaasa: Vaasan Kauppakorkeakoulun Julkaisuja Tutkimuksia No 67, Philologie 5.

Livesey, Eleanor Anne (2001). *Satzlängen in englischen und deutschen Pressetexten.* Staatsexamensarbeit, Göttingen.

Lord, R.D. (1958). Studies in the history of probability and statistics. VIII. De Morgan and the statistical study of literary style. *Biometrika 45, 282.*

Lyon, James K., & Inglis, Craig (1971). *Konkordanz zur Lyrik Gottfried Benns.* Hildesheim/ New York: Olms.

Mačutek, J. & Altmann, G. (2007). Discrete and continuous modelling in quantitative linguistics. *Journal of Quantitative Linguistics 14(1), 81-94.*

Marbe, Karl (1904). *Über den Rhythmus der Prosa*. Giessen: J. Ricker'sche Verlagsbuchhandlung.

Marbe, Karl (1916). *Die Gleichförmigkeit in der Welt.* München: Beck.

Martin, Robert (1974). Syntaxe de la définition lexicographique: Étude quantitative des définissants dans le „Dictionnaire fondamental de la langue fran←aise". In: David, J., & Martin, R. (eds.), *Statistique et Linguistique* (pp. 60-71). Paris: Klincksieck.

Meier, Helmut (1967). *Deutsche Sprachstatistik.* Zweite erw. u. verb. Aufl. Hildes-heim: Olms.

Menzerath, Paul (1954). *Die Architektonik des deutschen Wortschatzes.* Bonn: Dümmler.

Menzerath, Paul, & de Oleza, Joseph M. (1928). *Spanische Lautdauer. Eine experimentelle Untersuchung.* Berlin/ Leipzig: de Gruyter.

Merten, Klaus (1983). *Inhaltsanalyse.* Opladen: Westdeutscher Verlag.

Meuser, Katherine, Schütte, Jana Madlen & Stremme, Sina (2008). Pluralmorpheme in den Kurzgeschichten von Wolfdietrich Schnurre**.** *Glottometrics 17, 12-17*

Meyers Enzyklopädisches Lexikon. Bd. 25. Mannheim u.a.: Bibliographisches Institut/ Lexikonverlag 1985.

Meyers grosses Taschenlexikon in 24 Bänden. 2., neu bearb. Aufl. Mannheim/ Wien/ Zürich: B.I.-Taschenbuchverlag 1987.

Mikk, Jaan (2000). *Textbook: Research and Writing.* Frankfurt: Peter Lang.

Miller, George A. (1993). *Wörter. Streifzüge durch die Psycholinguistik.* Heidelberg/ Berlin/ New York: Spektrum Akademischer Verlag.

Miyajima, Tatsuo (1992). Relationships in the Length, Age and Frequency of Classical Japanese Words. In: Rieger, Burghard (Ed.), *Glottometrika 13* (pp. 219-229). Bochum: Brockmeyer.

Montemurro, Marcelo A., & Zanette, Damián H. (2001). Entropic analysis of the role of words in literary texts. *arXiv:cond-math/0109218v1 12 Sep 2001*, 1-9. (http://xxx.lanl.gov/abs/cond-mat/0109218)

Morley, Muriel E. (²1967). *The Development and Disorders of Speech in Childhood.* Edinburgh/ London: E. & S. Livingstone Ltd.

Moser, Hugo (1971). Typen sprachlicher Ökonomie im heutigen Deutsch. In: Hugo Moser, u.a. (Ed.), *Sprache und Gesellschaft.* Jahrbuch 1970 (pp. 89-117). Düsseldorf: Schwann.

Müller-Hasemann, Wolfgang (1983). Das Eindringen englischer Wörter ins Deutsche ab 1945. In: Best & Kohlhase (Hrsg.), *Exakte Sprachwandelforschung* (pp. 143-160).

Munske, Horst Haider (1988). Ist das Deutsche eine Mischsprache? Zur Stellung der Fremdwörter im deutschen Sprachsystem. In: Munske, Horst Haider, Polenz, Peter von, Reichmann, Oskar, & Hildebrandt, Reiner (Eds.), *Deutscher Wortschatz Lexikologische Studien. Ludwig Erich Schmidt zum 80. Geburtstag von seinen Marburger Schülern* (pp. 46-74). Berlin/ New York: de Gruyter.

Nasvytis, A. (1953). *Die Gesetzmäßigkeiten kombinatorischer Technik.* Berlin/ Göttingen/ Heidelberg: Springer.

Niehaus, Brigitta (1997). Untersuchung zur Satzlängenhäufigkeit im Deutschen. In: Best (Ed.), *Glottometrika 16* (S. 213-275).

Nitsch, Olaf (1997). *Wortkomplexität in Texten der Computerfachpresse.* Staatsexamensarbeit, Göttingen.

Oksaar, Els (1972). Stilstatistik und Textanalyse. In: Backes, Herbert (Ed.), *Festschrift für Hans Eggers zum 65. Geburtstag* (pp. 630-648). Tübingen: Niemeyer.

Ord, J. K. (1972). *Families of frequency distributions*. London: Griffin.

Orlov, Ju. K. (1982). Dynamik der Häufigkeitsstrukturen. In: Guiter, H. & Arapov, M.V. (eds.), *Studies on Zipf's Law.* Bochum: Brockmeyer.

Orlov, Ju. K. (1982b). Ein Modell der Häufigkeitsstruktur des Vokabulars. In: Orlov, Ju. K., Boroda, M.G., & Nadareijšvili, I. Š., *Sprache, Text, Kunst. Quantitative Analysen* (pp. 118-192). Bochum: Brockmeyer.

Osman, Nabil (Ed.) (^{6}1992). *Kleines Lexikon untergegangener Wörter. Wortuntergang seit dem Ende des 18. Jahrhunderts.* München: Beck.

Paul, Hermann (^{4}1909). *Prinzipien der Sprachgeschichte.* Halle: Niemeyer.

Pawłowski, Adam (2008). Prolegomena to the History of Corpus and Quantitative Linguistics. Greek Antiquity. *Glottotheory 1, 48-54.*

Pearl, Raymond (1926). *The Biology of Population Growth.* London: Williams and Norgate.

Pieper, Ursula (1979). *Über die Aussagekraft statistischer Methoden für die linguistische Stilanalyse.* Tübingen: Narr.

Piotrovskaja, A.A., & Piotrovskij, R.G. (1974). Matematičeskie modeli diachronii i tekstoobrazovanija. In: *Statistika reči i avtomatičeskij analiz teksta* (pp. 361-400). Leningrad: Nauka.

Piotrowski, R.G., Bektaev, K.B., & Piotrowskaja, A.A. (1985). *Mathematische Linguistik.* Bochum: Brockmeyer.

Polikarpov, Anatolij A. (2006). Towards the Foundations of Menzerath's Law. On the Functional Dependence of Affix Length On their Positional Number Within Words. In: Grzybek, P. (ed.), *Contributions to the Science of Text and Language: Word length studies and related issues* (pp. 215-240).

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified modeling of length in language.* Lüdenscheid: RAM-Verlag

Popescu, I.-I., Lupea, M., Tatar, D., Altmann, G. (2015). *Quantitative Analysis of Poetic Texts.* Berlin/Boston: de Gruyter

Poppe, Stefanie (2007). Die Verteilung von Kompositalängen in deutschen journalistischen Texten. *Göttinger Beiträge zur Sprachwissenschaft 15, 79-85.*

Pott, A. F. (1884). Einleitung in die allgemeine Sprachwissenschaft. *Internationale Zeitschrift für allgemeine Sprachwissenschaft 1 (=Techmers Zeitschrift), 1-68.*

Prün, Claudia (1999). G.K. Zipf's Conception of Language as an Early Prototype of Synergetic Linguistics. *Journal of Quantitative Linguistics 6, 78-84.*

Prün, Claudia (2002). Die linguistischen Hypothesen von G.K. Zipf. In: Köhler (Ed.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik* (pp. 271-321).

Prün, Claudia (2005). Das Werk von G.K. Zipf. In: Köhler, Altmann & Piotrowski (Eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch* (S. 142-152).

Pustet, Regina, & Altmann, Gabriel (2005). Morpheme Length Distribution in Lakota. *Journal of Quantitative Linguistics 12, 53-63.*

Ricken, U. (1981). Linguistik und Naturwissenschaft in der Aufklärung. In: Scharf, J.-H., & u. Kämmerer, W. (Eds.), *Leopoldina-Symposion Naturwissenschaftliche Linguistik* (pp. 544-564). Halle: Deutsche Akademie der Naturforscher LEOPOLDINA. (= Nova Acta Leopoldina N.F. 54, Nr. 245).

Rosengren, Inger (1972/77). *Ein Frequenzwörterbuch der deutschen Zeitungssprache. Bd. 1, 2.* Lund: Gleerup.

Rote Liste (1989). Verzeichnis von Fertigarzneimitteln der Mitglieder des Bundesverbandes der Pharmazeutischen Industrie e.V. Hrsg. v. Bundesverband der Pharmazeutischen Industrie e.V. Aulendorf/ Württ.: Editio Cantor.

Rothe, Ursula (1991). The Diversification of the Case: Genitive. In: Rothe (Ed.), *Diversification Processes in Language: Grammar* (pp. 140-156).

Rothe, Ursula (1991a). Diversification Processes in Grammar. An Introduction. In: Rothe (Ed.), *Diversification Processes in Language: Grammar* (pp. 3-32).

Rothe, Ursula (Ed.) (1991), *Diversification Processes in Language: Grammar.* Hagen: Margit Rottmann Medienverlag.

Rothe, Ursula, Wagner, Klaus R., & Altmann, Gabriel (1992). Verteilung der Länge von Sprechakten in der Kindersprache. In: Wagner, Klaus R. (Ed.), *Kindersprachstatistik* (S. 47-56). Essen: Die Blaue Eule.

Rothweiler, Monika (2001). *Wortschatz und Störungen des lexikalischen Erwerbs bei spezifisch sprachentwicklungsgestörten Kindern.* Heidelberg: Winter.

Roukk, Maria (2001). Satzlängen in Texten von A. Tschechov. *Göttinger Beiträge zur Sprachwissenschaft 5, 113-120.*

Rundblad, Gabriella (2000). On the Correlation between Lexical Stability and Word Creation Device. *Journal of Quantitative Linguistics 7, 31-41.*

Sachs, Lothar ([5]1978). *Angewandte Statistik. Statistische Methoden und ihre Anwen-dungen.* Berlin/ Heidelberg/ New York: Springer.

Sambor, Jadwiga, & Hammerl, Rolf (Hrsg.), *Definitionsfolgen und Lexemnetze. Bd. 1.* Lüdenscheid: RAM-Verlag.

Sanada-Yogo, Haruko (1999). Analysis of Japanese Vocabulary by the Theory of Synergetic Linguistics. *Journal of Quantitative Linguistics 6, 239-251.*

Saporta, Sol (1963, ²1966). Phoneme Distribution and Language Universals. In: Greenberg, Joseph H. (ed.), *Universals of Language. Report of a Conference Held at Dobbs Ferry, New York, April 13-15, 1961* (pp. 61-72). Cambridge, Mass./ London: The M.I.T. Press.

Schierholz, Stefan J. (1991). *Lexikologische Analysen zur Abstraktheit, Häufigkeit und Polysemie deutscher Substantive.* Tübingen: Niemeyer.

Schlaefer, Michael (1999). *Ein Leitfaden für den Benutzer.* Göttingen: Deutsches Wörterbuch von Jacob u. Wilhelm Grimm, Neubearbeitung, Arbeitsstelle Göttingen.

Schmidt, Franz (1966). *Zeichen und Wirklichkeit.* Stuttgart: Kohlhammer.

Schmidt-Samoa, Carsten (2002). *Welche Rolle spielt die Homogenität deutscher Texte bei der Anpassung der hyper-Poisson-Verteilung? Eine Computersimulation.* Seminararbeit, Göttingen.

Schneemann, Okke F. (2001). *Sprachstatistische Untersuchungen zu Wort- und Silbenlängen in deutschen Musikzeitschriften.* Staatsexamensarbeit, Göttingen.

Schulte, Klaus (1979). *Phonemhäufigkeit und Artikulation* Villingen-Schwenningen: Neckar-Verlag.

Schweers, Anja, & Zhu, Jinyang (1991). Wortartenklassifizierung im Lateinischen, Deutschen und Chinesischen. In: Rothe (Ed.), *Diversification Processes in Language: Grammar* (pp. 157-165).

Seibicke, Wilfried (1982). *Die Personennamen im Deutschen.* Berlin/New York: de Gruyter.

Sherman, L.A. (1888). Some observations upon the sentence-length in English prose. *University of Nebraska Studies 1, 119-130.*

Sowinski, Bernhard (1979). *Werbeanzeigen und Werbesendungen.* München: Olden-bourg.

Sowinski, Bernhard (1991). *Stilistik.* Stuttgart: Metzler.

Sowinski, Bernhard (1998). *Werbung*. Tübingen: Niemeyer.

Stiberc, Andrea (1999). *Sauerkraut, Weltschmerz, Kindergarten und Co.* Freiburg: Herder.

Störig, Hans Joachim (1997). *Abenteuer Sprache. Ein Streifzug durch die Sprachen der Erde.* 2., überarb. Aufl. München: Humboldt-Taschenbuchverlag.

Strauss, Udo, Fan, Fengxiang & Altmann, Gabriel (2008). *Problems in Quantitative Linguistics 1.* Lüdenscheid: RAM-Verlag.

Strauss, Udo, Grzybek, Peter, & Altmann, Gabriel (2006). Word Length and Word Frequency. In: Grzybek, P. (ed.), *Contributions to the Science of Text and Language: Word length studies and related issues* (pp. 277-294).

Strobel, Heike (1996). *Wortlängen in Briefen und Erzählungen von Böll und Heming-way.* Staatsexamensarbeit, Göttingen.

Suhren, Svenja (2002). *Untersuchung zum Gesetz von Zwirner, Zwirner und Frumki-na am Beispiel des niederdeutschen „De lütte Prinz".* Staatsexamensarbeit, Göttingen.

Swoboda, Helmut (1974). *Knaurs Buch der modernen Statistik.* München: Droemer Knaur.

Ternes, Katarina (2011). Entwicklungen im deutschen Wortschatz. *Glottometrics 21, 25-53.*

Thumb, Albert (1911). Experimentelle Psychologie und Sprachwissenschaft. Ein Beitrag zur Methodenlehre der Philologie. *Germanisch-Romanische Monatsschrift 3, 1-15; 65-74.*

Thumb, Albert, & Marbe, Karl (1901). *Experimentelle Untersuchungen über die psychologischen Grundlagen der sprachlichen Analogiebildung.* Leipzig: Engelmann.

Tuldava, Juhan (1995). *Methods in Quantitative Linguistics.* Trier: Wissenschaftlicher Verlag Trier.

Tuldava, Juhan (1998). *Probleme und Methoden der quantitativ-systemischen Lexikologie.* Trier: Wissenschaftlicher Verlag Trier.

Twain, Mark (1880; 1961). The Awful German Language. In: *The Complete Humorous Sketches and Tales of Mark Twain.* Ed. by Charles Neider. New York: Doubleday. S. 439-455. (Dt.: M. Twain (1985). *Bummel durch Europa.* Anhang D: Die schreckliche deutsche Sprache. Frankfurt: Insel) (pp. 527-545.)

Uhlířová, Ludmila (1995). On the Generality of Statistical Laws and Individuality of Texts. A Case of Syllables, Word Forms, their Length and Frequencies. *Journal of Quantitative Linguistics 2, 238-247.*

Valentin, Karl (1996). *Sämtliche Werke in acht Bänden. Bd. 4.* Hrsg. v. Helmut Bachmaier u. Manfred Faust. München/ Zürich: Pieper.

Verhulst, P.-F. (1838). Notice sur la loi que la population suit dans son accroissement. c*orrespondance Mathématique et Physique, Tome X, 3-21.*

Verhulst, P.-F. (1845). Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles, Tome XVIII, 5-38.*

Vettermann, Anikó, & Best, Karl-Heinz (1997). Wortlängen im Finnischen. *Suomalais-ugrilaisen seuran aikakauskirja/ Journal de la Société Finno-Ougrienne 87, 249-262.*

Viereck, Wolfgang, Viereck, Karin, & Ramisch, Heinrich (2002). *dtv-Altas Englische Sprache.* München: dtv.

Wagner, Klaus R. (1974). *Die Sprechsprache des Kindes. Teil 1.* Düsseldorf: Schwann.

Wagner, Klaus R., Altmann, Gabriel, & Köhler, Reinhard (1987). Zum Gesamtwortschatz der Kinder. In: Wagner, Klaus R. (Ed.), *Wortschatz-Erwerb* (pp. 128-142). Bern u.a.: Peter Lang.

Wängler, Hans-Heinrich (1963). *Rangwörterbuch hochdeutscher Umgangssprache*. Marburg: Elwert.

Wellmann, Hans (1998). Die Wortbildung. In: *Duden. Grammatik der deutschen Gegenwartssprache.* 6., neu bearb. Aufl. (pp. 408-557). Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag.

Wermser, Richard (1976). *Statistische Studien zur Entwicklung des englischen Wortschatzes.* Bern: Francke Verlag.

Wichter, Sigurd (1991). *Zur Computerwortschatz-Ausbreitung in die Gemeinsprache. Elemente der vertikalen Sprachgeschichte einer Sache.* Frankfurt u.a.: Peter Lang.

Wichter, Sigurd (1998). Technische Fachsprachen im Bereich der Informatik. In: Hoffmann, Lothar, Kalverkämper, Hartwig, & Wiegand, Herbert Ernst (Eds.), in Verbindung mit Galinski, Christian, & Hüllen, Werner, *Fachsprachen – Languages for Special Purposes* (pp. 1173-1182). Berlin/ New York: de Gruyter.

Wickmann, Dieter (1981). Unbekannte Verfasserschaft – statistisch gesehen. In: Scharf, Joachim-Hermann, & Kämmerer, Wilhelm (Eds.), *Leopoldina-Symposion Naturwissenschaftliche Linguistik* (S. 277-281). Halle: Deutsche Akademie der Naturforscher LEOPOLDINA 1981. (= Nova Acta Leopoldina, N.F. 54, Nr. 245).

Wimmer, Gejza (2005). The type-token relation. In: Köhler, Altmann & Piotrowski (Eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch* (pp. 361-368).

Wimmer, Gejza, & Altmann, Gabriel (1996). The Theory of Word Length Distribution: Some Results and Generalizations. In: Schmidt, Peter (Ed.), *Glottometrika 15* (pp. 112-133), Trier: Wissenschaftlicher Verlag Trier.

Wimmer, Gejza, & Altmann, Gabriel (1999). Review Article: On Vocabulary Richness. *Journal of Quantitative Linguistics 6, 1-9.*

Wimmer, Gejza, & Altmann, Gabriel (1999a). *Thesaurus of univariate discrete probability distributions.* Essen: Stamm.

Wimmer, Gejza, & Altmann, Gabriel (2005). Unified Derivation of Some Linguistic Laws. In: Köhler, Altmann & Piotrowski (eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch* (pp. 791-807).

Wimmer, Gejza, & Altmann, Gabriel (2006). Towards a Unified Derivation of Some Linguistic Laws. In: Grzybek (ed.), *Contributions to the Science of Text and Language: Word length studies and related issues* (pp. 329-337).

Wimmer, Gejza, Köhler, Reinhard, Grotjahn, Rüdiger, & Altmann, Gabriel (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics 1, 98-106.*

Wimmer, Gejza, Witkovský, Viktor, & Altmann, Gabriel (1999). Modification of Probability Distributions Applied to Word Length Research. *Journal of Quantitative Linguistics 6, 257-268.*

Winter, Horst (1986). Benennungsmotive für chemische Stoffnamen. *Special Language/ Fachsprache 8, 155-162.*

Wittek, Martin (2001). Zur Entwicklung der Satzlänge im gegenwärtigen Deutschen. In: Best, K.-H. (Ed.) 2001, *Häufigkeitsverteilungen in Texten* (pp. 219-247).

Wörterbuch zu HEINRICH VON KLEIST. Sämtliche Erzählungen, Anekdoten und kleine Schriften (1989). 2., völlig neu bearb. Aufl. v. Helmut Schanze. Tübingen: Niemeyer.

Wolff, Dieter (1969). *Statistische Untersuchungen zum Wortschatz englischer Zeitungen.* Saarbrücken, diss. phil.

Yu, Xiaoli (2001). Zur Komplexität chinesischer Schriftzeichen. *Göttinger Beiträge zur Sprachwissenschaft 5, 121-129.*

Yu, Xiaoli (2002). *Satzlängen in Fachtexten des 19. Jahrhunderts.* Magisterarbeit, Göttingen.

Ziegler, Arne, & Altmann, Gabriel (2001). Beziehung zwischen Synonymie und Polysemie. In: Ondrejovič, Slavomír & Považaj, Matej (eds.), *Lexicographica '99. Sborník na Počest Kláry Buzássyovej* (pp. 226-229). Bratislava: Veda.

Ziegler, Arne, Best, Karl-Heinz, & Altmann, Gabriel (2002). Nominalstil. *ETC – Empirische Text- und Kulturforschung 2, 72-85.*

Zipf, George Kingsley (1932). Relative Frequency, Abbreviation, and Semantic Change. In: George Kingsley Zipf, *Selected Studies of the Principle of Relative Frequency in Language* (pp. 8-27). *C*ambridge, Mass.: Harvard University Press.

Zipf, George Kingsley (1935/ 1968). *The Psychobiology of Language: An Introduction to Dynamic Philology.* Cambridge, Mass.: The M.I.T. Press.

Zipf, George Kingsley (1949). *Human Behavior and the Principle of Least Efford.* Cambridge, Mass.: Addison-Wesley.

Zwirner, Eberhard, & Zwirner, Kurt (1935). Lauthäufigkeit und Zufallsgesetz. *Forschungen und Fortschritte 11, Nr. 4, 43-45.*

Zwirner, Eberhard, & Zwirner, Kurt (1936). Streuung sprachlicher Merkmale. *Forschungen und Fortschritte 12, Nr. 15*: 191-192. (Can also be found in: Zwirner & Ezawa (Eds.), part III, 55-59.)

Zwirner, Eberhard, & Zwirner, Kurt (1938). Lauthäufigkeit und Sprachvergleichung. *Monatsschrift für höhere Schulen 37, 246-253*. (Can also be found in: Zwirner & Ezawa (Eds.), part III, 68-74.)

Zwirner, Eberhard, Ezawa, Kennosuke (Eds.) (1969). *Phonometrie. Dritter Teil: Spezielle Anwendungen I.* Basel/ New York: Karger.

Zwirner, Eberhard, & Zwirner, Kurt (1966). *Phonometrie. Erster Teil: Grundfragen der Phonometrie*. 2nd ed. Basel/ New York: Karger.

Software

Altmann-Fitter (1997/2005). *Iterative Fitting of Probability Distributions.* Lüdenscheid: RAM-Verlag. (New Version 2005)

MAPLE V Release 4 (1996). Berlin u.a.: Springer.

NLREG*. Nonlinear Regression Analysis Program*. Ph. H. Sherrod. copyright (c) 1991 - 2001.

Author Register

Subject Register

The RAM-Verlag Publishing House edits since 2001 also the journal ***Glottometrics*** **–** up to now 37 issues – containing articles treating similar themes. The abstracts can be found in http://www.ram-verlag.eu/journals-e-journals/glottometrics/.

Open Access to Glottometrics

The contents of the last issue (37, 2017) is as follows:

Herausgeber – Editors of Glottometrics

G. Altmann	Univ. Bochum (Germany)	ram-verlag@t-online.de
K.-H. Best	Univ. Göttingen (Germany)	kbest@gwdg.de
R. Čech	Univ. Ostrava (Czech Republic)	cechradek@gmail.com
F. Fan	Univ. Dalian (China)	Fanfengxiang@yahoo.com
P. Grzybek	Univ. Graz (Austria)	peter.grzybek@uni-graz.at
E. Kelih	Univ. Vienna (Austria)	emmerich.kelih@univie.ac.at
H. Liu	Univ. Zhejiang (China)	lhtzju@gmail.com
J. Mačutek	Univ. Bratislava (Slovakia)	jmacutek@yahoo.com
G. Wimmer	Univ. Bratislava (Slovakia)	wimmer@mat.savba.sk
P. Zörnig	Univ. Brasilia (Brasilia)	peter@unb.br