GLOTTOMETRICS

To Honor G.K. Zipf

3

2002

RAM-VERLAG

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitschrift für die quantitative Erforschung von Sprache und Text
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden
Glottometrics kann aus dem **Internet** heruntergeladen, auf **CD-ROM** (in PDF Format) oder in **Buchform** bestellt werden

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors
Glottometrics can be downloaded from the **Internet**, obtained on **CD-ROM** (in PDF) or in form of **printed copies**

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an
Orders for CD-ROM´s or printed copies to

RAM-Verlag RAM-Verlag@t-online.de

Herunterladen / Downloading: http://www.ram-verlag.de

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Glottometrics. –3 (2002) –. – Lüdenscheid: RAM-Verl., 2002
Erscheint unregelmäßig. – Auch im Internet als elektronische Ressource unter der Adresse http://www.ram-verlag.de verfügbar.-
Bibliographische Deskription nach 3 (2002)
ISSN 1617-8351

Foreword

George Kingsley Zipf was born 100 years ago. We chose this way to honor a scientist who originally was a linguist, published mostly in psychological journals and is cited in about twenty scientific disciplines. He is the founder of quantitative linguistics which is sometimes called "Zipfian linguistics".

G.K. Zipf discovered the dynamic character of language and opened the door for the entry of the systemic view in linguistics. His discoveries came too early, time was not ripe, the static view of language dominated the whole 20th century. Today, in the age of self-organization, self-regulation, chaos theory, dynamic systems, self-organized criticality, etc. his laws are re-discovered again and again. Not only in human sciences - where mathematics is still avoided and even the existence of laws is sometimes denied - but especially in the natural sciences the power law (= Zipf´s law) is found behind highly variegated phenomena and seems to capture a mechanism working somewhere in the heart of the world.

G.K. Zipf is the only quantitative linguist accepted in qualitative linguistics as well, not only because he considered a number of languages but mainly because of his ideas which are more modern than they appear to be in his books. An attentive reader can find a new hypothesis on every page of his works. They can be adopted in synergetic linguistics without much modification.

G.K. Zipf cared about interdisciplinarity. He transferred his ideas to settlement geography, economics, psychology and sociology and tried to explained human behavior from common principles. Today, they are a kind of substratum used by all of us unconsciously.

In this volume a collective of admirers of G.K. Zipf, consisting of physicists, mathematicians, informetricians, biologists, psychologists, musicologists and linguists, tried to express its reverence for a scientist whose seminal ideas have become the basis of our daily research. The volumes dedicated to Zipf appear as serial issues of a linguistic journal, but contain matters of very general interest. This was the way we have chosen in order to honor a scientist who revolutionized linguistics and linked it to other sciences. Linguistics is not an isolated science any more, the dissemination of knowledge is not so unidirectional as it was in the last century. Linguistics has became a science, an honorable member of a great family.

Looking back to linguistics in Zipf´s century from our present, safe observation point, we can easily understand the causes of the long stagnation: dynamics was no subject matter of linguistics, structuralism and generativism were too static. Statics was a dogma and G.K. Zipf put an end to it.

The contributions are ordered neither alphabetically nor as to content, the only criterion was their arrival at the editorial office. This policy will, as we hope, stimulate linguists to read non-linguistic articles and non-linguists to have a look at linguistic problems and help us solve them.

We are especially obliged to Robert Zipf, the son of G.K. Zipf, who collected inestimable biographic data leading us through G.K. Zipf´s life down to his roots. Future biographers will find the most authentic sources here.

We hope that we acted in compliance with the ideas and intentions of G.K. Zipf, the wanderer between sciences.

Editorial Board

Contents

Foreword

Prün, Claudia, Zipf, Robert
Biographical notes on G.K. Zipf 1

Rousseau, Ronald
George Kingsley Zipf: life, ideas, his law and informetrics 11

Altmann, Gabriel
Zipfian linguistics 19

Hřebíček, Luděk
Zipf´s law and text 27

Uhlířová, Ludmila
Zipf´s notion of „economy“ on the text level 39

Gumenjuk, A., Kostyshin, A., Simonova, S.
An approach to the analysis of text structure 61

Andersen, Simone
Speakers´s information content: length-frequency correlation as partial correlation 90

Majerník, Vladimir
A conceptualization of the configurational and functional organization 110

Best, Karl-Heinz
The distribution of rhythmic units in German short prose 136

Adamic, Lada, A., Huberman, Bernardo
Zipf´s law and the Internet 143

George Kingsley Zipf (1902-1950)

Glottometrics 3, 2002, 1-10 *To honor G.K. Zipf*

Biographical notes on G. K. Zipf

Claudia Prün, Robert Zipf [1]

Abstract. Harvard philologist George Kingsley Zipf has been underestimated by mainstream linguistics for the past half century. After short consideration of the significance of Zipf's work, this paper presents a personal account of Zipf's family background, career and private life by one of his sons. An extensive Bibliography is added.

Keywords: G.K. Zipf

Significance

G. K. Zipf, a Harvard philologist in the second quarter of the 20th century, has not yet received much recognition in the historiography of linguistics, though his significance for the development of a linguistic theory cannot be overestimated. Because this significance is just beginning to be recognized more widely, the present paper presents only a start in the consideration of Zipf in a historical respect. As it presents a unique, and quite personal, view of this scholar, we will not treat the development of his ideas in connection with and emerging from the intellectual currents prevailing in his time. Rather, we focus on family background, biographical cornerstones, and some personal reminiscences of one of the authors (and his family), together with a bibliography of Zipf's works. Also added are the Reviews which were accessible, though they have not been appreciated in this article. Owing to the cooperation of Zipf's eldest son, this paper contains invaluable sources for the historical investigation and a biography of G. K. Zipf.

Zipf´s name is best known from the rank-frequency and frequency distribution laws which are named after him ("Zipf´s laws", Prün, to appear). Though he was not the first to detect certain regularities in the frequency structure of texts, he certainly was the one to work most on the subject, explaining the phenomena, and extending his hypotheses to other fields of science, mainly sociology. This is also a field where he is still extensively referenced, rather than in linguistics - for the time being - as a look into the Social Sciences Citation Index (SSCI) will convince those who look. An obituary (Lundberg and Dodd, 1950) appeared in the American Sociological Review but not in a philological or linguistic periodical.

Zipf tried to establish the science of language similar to natural sciences (1935, 1; 1949, 1). His method was therefore quantitative, and his explanations would today be called functional-systemic (Prün, 1999). Functional, in the sense used here, means serving requirements of production economy, transmission effectiveness, inventory optimization and many more requirements which are imposed upon the language system to maintain the communication function of the system. Functional explanation in linguistics as an explicit paradigm (Altmann, 1981; Köhler since 1986), which has been implicitly utilized by Zipf, has not caught firm hold in the linguistic community until recently. But its close relation to explanatory approaches in other sciences, as in sociology; its power to explain not only language generation but language

1 Address correspondence to: Claudia Prün, Anton-Caspary-Str. 14, D-54295 Trier.
E-mail: PRUEN@t-online.de

constitution by self-organization and self-regulative mechanisms, as in biology; and the close correspondence of its results to those of neurolinguistic and psycholinguistic investigation convince us that Zipf's significance as a linguist is only beginning to be recognized, and his achievements will be treated more thoroughly by linguistic historiography in the future.

Family background

George Kingsley Zipf was born on January 7, 1902, in the family house at 33 North Whistler, in Freeport, Illinois. His father was Oscar Robert Zipf, and his mother was Maria Louisa Bogardus Zipf.

George Kingsley Zipf's grandfather was Frederick Sebastian Zipf, who was born in Tauberbischofsheim, Germany. He was a Roman Catholic, and, according to a family tradition, sang in his church choir. The Zipf family relatives owned and ran the local brewery, Brauerei Zipf. Frederick Sebastian Zipf came to America from Tauberbischofsheim about 1850. At one point, he is believed to have sold shoes in his new home country to support himself and his family. He married Ottilia, a girl of German background. R. Zipf supposes that they probably married in the U.S. They had many children. According to one family tradition, Frederick Sebastian is buried in the Roman Catholic cemetery in Kankakee.

One son, Oscar Robert Zipf, G. K. Zipf's father, was born in Kankakee, IL, on February 14, 1866. He went to Michigan Law School (part of the University of Michigan), graduated with the class of 1889, and immediately practiced in Salt Lake City, UT. On April 28, 1892, he married Maria Louisa Bogardus, in Paxton, IL. Five years later, in 1897, they moved to Freeport, Stephenson County, IL, where O. R. Zipf continued to practice law. He also served a two-year term as county judge. O. R. Zipf died on March 28, 1942.

Maria Louisa Bogardus was the daughter of Charles Bogardus and Hannah Whittaker Pells. Both the Bogardus and the Pells families arrived in New Amsterdam. Everardus Bogardus arrived as Dominie in 1632, the second Dominie in New Amsterdam, and the Pells family arrived around 1650. Maria Louisa Bogardus was born on September 10, 1865, in Ridgway, Orleans County, New York. When she was five, her family moved to Paxton, IL. She died on December 31, 1936.

Maria Louisa Bogardus's father, (and G. K. Zipf's grandfather) was Charles Bogardus, from Orleans County, New York. He served in the Union Army from 1862 to1865. He was elected a First Lieutenant in August, 1862, received subsequent promotions to Captain and Major, and was wounded at the battle of Monocacy in July, 1864. He was discharged for wounds in December, 1864, but, in January, 1865, he was mustered back in with rank of Lieutenant Colonel, and given command of his regiment, the 151st New York Volunteers. In April, 1865, he was made Colonel by Brevet for "gallant and meritorious service before Petersburg, Va." He was discharged, with his regiment, in June, 1865. The family believes that he was, at one time, the youngest colonel in the Union Army.

After his discharge, Charles Bogardus pursued various business interests in Paxton, and in upper Michigan. He also served in the Illinois State Legislature.

Zipf's parents, Oscar Robert Zipf and Maria Louisa Bogardus Zipf had four sons: Oscar Robert, Charles Bogardus, George Kingsley, and Theodore. Oscar Robert served in the Army Air Corps in Europe during the First World War, and continued in the Air Corps as a test pilot for several years. He left the service, but joined again and served during the Second World War with the 14th Air Force in China, where he managed an airfield. He left the Air Corps after the Second World War, but later rejoined and served in the United States Air Force during the Korean War. His nephew remembers his Uncle Bob teaching him how to play gin rummy, a game at which the nephew rarely loses, thanks to the good instruction. Oscar Robert married Nance late in his life;

they had no children. His widow lived until 1991. Charles Bogardus was a medical doctor who practiced in Freeport for his whole life. He lived in the family home there. Theodore, G. K. Zipf's younger brother, died young during the great flu epidemic.

Biography

G. K. Zipf was born on January 7, 1902 and attended Freeport High School, where he won the gold F (for Freeport) as a prize for excellence in studies. He attended Harvard College, in the class of 1923, but he graduated with the class of 1924, summa cum laude. After graduation, he studied at the University of Berlin with W. Schulze, and Bonn (F. Sommer), where he first got the idea of examining language as a natural phenomenon. After returning to the United States, he earned a PhD in comparative philology at Harvard. His dissertation was *Relative Frequency as a Determinant of Phonetic Change* (Zipf, 1929). He then joined the Harvard faculty as an instructor in German. He also continued his linguistic researches, and published his first important book, *The Psycho-Biology of Language* (Zipf, 1965), in 1935. An extended edition in German had been planned with C. Winter publishers in Heidelberg, Germany (Zipf: Erwiderung, 1938; Zipf and Rogers, 1939:112). It was even published, according to the accounts of Zipf's widow, but seems to have been lost during the war. A French translation was published in 1974.

During the next several years, G. K. Zipf broadened his interests to include various social phenomena, and published *National Unity and Disunity* (Zipf, 1941). During Word War II, he was asked to move to Washington to work on the war effort in some capacity, but he declined and stayed at Harvard. At that time he ran first year German at Harvard, and changed the emphasis to learning vocabulary, so that at the end of one year of college German, students could read German with the aid of a dictionary. He presented this pedagogical concept and the linguistic reasons justifying the heavily lexical approach in "On the problem of grammatical rules and the study of 'General Language'" (1938). His achievements as a teacher are also highly praised in the Harvard Gazette obituary (Crozier, Rogers, Walsh, 1950: 81f.). The idea was that students would probably not have several years at Harvard to study a language, but would join the armed services after the current year of school, or their draft boards would draft them after the school year was over. He continued his social relations researches as well. His son, R. Zipf, helped him with some of the statistical calculations, which required considerable hand calculations in that pre-computer time.

Zipf´s German family background probably had very little part in Zipf's interest in German. The family tradition is that his grandfather had said, "We were Germans, we spoke German. We are now Americans, we will speak English". G. K. Zipf probably learned very little, if any, German at home, since his mother did not speak German, and German was not spoken in the home.

Zipf had an interest in George Meredith, which resulted in one publication (New facts in the early life of George Meredith, 1938), but did no further work in this field. The end of the text suggests that he had in his mind how the lives of individuals were embedded into greater processes and systems.

On June 20, 1931, Zipf married Joyce Waters Brown, of Webster, Massachusetts. She was born on April 5, 1902. Joyce descends from four governors of colonial Massachusetts. She grew up in Webster, Massachusetts and graduated from Wykeham Rise School in Connecticut, where she received a prize „to the best", and attended Smith College. She was working in the Harvard Alumni Association Office in Cambridge at the time she met and married G. K. Zipf. They had four children, Robert (born September 19, 1932), Katharine Slater (born October 1, 1934), Joyce Bogardus (born February 17, 1938), and Henry, born August 6, 1939. All four children are still

living, and have a total of ten children and seven grandchildren among them. After marriage, the Zipfs lived on Grozier Road in Cambridge, Mass., but shortly moved to Walker Street, also in Cambridge. In 1936, they moved to Duxbury, Mass., in a house on Tremont Street. In September 1943, they moved to Newton, Mass., where their children attended the public schools.

In 1939, G. K. Zipf was appointed University Lecturer, accounting for his interdisciplinary work. His work was not only in language, but also increasingly in the social sciences. He continued in this appointment until his death. This death was particularly tragic, since he had received a Guggenheim fellowship earlier in the year. He had also given a series of lectures at the University of Chicago in the later winter or early spring, as well as lectures at other universities, so he was achieving increased recognition throughout the academic community. R. Zipf, attending the College of the University of Chicago at the time, remembers the notices on the bulletin boards around campus.

Zipf planned to use the freedom provided by his Guggenheim Fellowship to pursue his ideas researching American business, and its enterprises and operations. The Guggenheim appointment was "for quantitative study of certain marketing phenomena for the purpose of disclosing underlying statistical regularities." This direction is also shown in his last few papers, published in 1949 and 1950. These papers analyzed various aspects of American business.

The Guggenheim fellowship at that time required a physical examination, and it was during this examination that Zipf's cancer was discovered. He had an operation in June, 1950, but the cancer was too far advanced to do anything, and so he returned home after about a month in the hospital, and died several months later on September 25, 1950. His ashes are buried in the Mayflower Cemetery in Duxbury, Massachusetts. In August, 1961, his widow remarried, to David W. Bailey, and moved back to Cambridge, Mass.

Family life

When he lived in Duxbury, G. K. Zipf worked at home, but commuted to Cambridge by train two or three times a week to work there and to teach his classes. The family had seven and one-half acres of house, lawn, meadow, garden, and woodland. During the war, they raised chickens, ducks, and pigs. At one time, they raised a pig jointly with a neighbor across the street, Harry Bradley, then Chairman of the Plymouth Cordage Company. Mr. Bradley bought the piglet, the Zipfs raised the pig, and the families shared the pig equally when it was slaughtered.

Zipf was an avid gardener, and they also had a large vegetable garden. All family members worked on the garden though not everybody shared "father's" enthusiasm for gardening. They grew vegetables and fruits in considerable variety and quantity, even had a vineyard, and there was a large flower garden close to the house. However, more than half the land was pine woods, and other large areas of woodland abutted the property, so the children could explore the woods easily. At Christmas, Zipf and his eldest son would go out into the woods, select a Christmas tree, cut it down and bring it into the house. The entire family decorated the tree.

Several years after they moved to Newton, they sold the Tremont Street house and bought a house on Upland Road, on Powder Point (also in Duxbury) as a summer house. This house was within an easy walk to Duxbury Bay and Duxbury beach. The house had beach rights and the beach was easily reached, though the house was not on the water. The property was much smaller, but still it had some woods, and Zipf planted a garden. However, he only enjoyed the house for two or three summers before he died. The Zipfs had a small sailboat, and R. Zipf recalls sailing in Duxbury bay, when the tide was in. Occasionally, he would take his father sailing in the boat, but generally, Zipf worked on his main book, *Human Behavior and the Principle of Least Effort* (Zipf, 1972), which came out in 1949.

G. K. Zipf worked late at night, into the early hours of the morning, but would sleep until late morning, unless he had an early class. When the family moved to Newton, their house had a third floor, and he made one of the third floor rooms into an office. He would work there very late at night, and the neighbors would frequently comment on seeing the light on in his third floor office when they were going to bed. R. Zipf remembers the office, with a desk, with various papers on it, reflecting the work his father was working on at that particular time. It did not have many books; Zipf believed in using libraries, and did not believe in owning books. The books he had were mostly language books - books in German on various German philological subjects, including the Gothic Bible, and various medieval German studies. He also had a Sanskrit reader and grammar, an Anglo-Saxon dictionary, and other similar books. Years later, his widow gave all of these to Widener library.

Zipf is reported as "approachable, enthusiastic, original, and inspiring" (Crozier, Rogers, Walsh, 1950: 82). His natural warmth and sympathetic ways made him popular with his students. Still, he was well known for making his students assist in his counting and calculating work (Birkhan, 1979: 50). He frequently mentions in his publications the names of those who helped with the data acquisition and statistics, but Birkhan (ibid.) says that a former student of Zipf reports that they had not taken his ideas very seriously.

George Kingsley Zipf ´s work was controversial, but he believed in its value, and devoted himself fully to it. Owing to this, we have 6 monographs and 36 journal articles by Zipf (including 3 with coauthors), which are listed in the bibliography below.

G. K. Zipf: Bibliography

Zipf, George Kingsley: *Relative frequency as a determinant of phonetic change*. Harvard studies in classical philology 40, 1929.

Zipf, George Kingsley: *Selected studies of the principle of relative frequency in language*. Cambridge/Mass., Harvard Univ.Press, 1932.

Zipf, George Kingsley: *The psycho-biology of language. An introduction to dynamic philology*. Cambridge/Mass., M.I.T. Press, 2nd ed. 1968 [First Edition: Boston, Houghton-Mifflin, 1935].

Zipf, George Kingsley: Observations on the possible effect of mental age upon the frequency-distribution of words from the viewpoint of dynamic philology. In: *Journal of psychology* 4 (1937), 239-244.

Zipf, George Kingsley: Statistical methods in dynamic philology (Reply to M. Joos). In: *Language* 132 (1937), 60-70.

Zipf, George Kingsley: Erwiderung. In: *Indogermanische Forschungen* 56 (1938), 75-77. [ad Jost (1937)]

Zipf, George Kingsley: Homogeneity and heterogeneity in language. In answer to Edward L. Thorndike. In: *Psychological record* 2 (1938), 347-367.

Zipf, George Kingsley: On the problem of grammatical rules and the study of 'General Language'. In: *Modern language journal* 22/4 (1938), 243-249.

Zipf, George Kingsley: Phonometry, phonology, and dynamic philology. An attempted synthesis. In: *American speech* 13 (1938), 275-285.

Zipf, George Kingsley: *New facts in the early life of George Meredith*. Harvard studies and notes in philology and literature 20 (1938).

Zipf, George Kingsley and Rogers, Francis Millet: Phonemes and variphones in four present-day Romance languages and Classical Latin from the viewpoint of dynamic philology. In: *Archives néerlandaises de phonétique expérimentale* 15 (1939), 111-147.

Zipf, George Kingsley: The generalized harmonic series as a fundamental principle of social organization. In: *Psychological record* 4 (1940), S. 43.

Zipf, George Kingsley: On the economical arrangement of tools, the harmonic series and the properties of space. In: *Psychological record* 4 (1940), 147-159.

Zipf, George Kingsley: *National unity and disunity; The nation as a bio-social organism*. Bloomington, Ind., Principia Press, Inc., 1941.

Zipf, George Kingsley: Children's speech. In: *Science* 96 (1942), 344-345.

Zipf, George Kingsley: The unity of nature, least-action, and natural social science. In: *Sociometry* 5 (1942), 48-62.

Whitehorn, J. C. and Zipf, George Kingsley: Schizophrenic language. In: *Archive of neurology and psychiatry* 49 (1943), 831-851.

Zipf, George Kingsle: Letter to the editor. In: *American journal of socology* 48 (1943), 503-504.

Zipf, George Kingsley: The repetition of words, time-perspective and semantic balance. In: *Journal of general psychology* 32 (1945), 127-148.

Zipf, George Kingsley: The meaning-frequency relationship of words. In: *Journal of general psychology* 33 (1945), 251-256.

Zipf, George Kingsley: Some psychological determinants of the structure of publications. In: *American journal of psychology* 58 (1945), 425-442.

Zipf, George Kingsley: The P_1P_2 / D hypothesis: The case of railway express. In: *Journal of psychology* 22 (1946), 3-8.

Zipf, George Kingsley: On the dynamic structure of concert programs. In: *Journal of abnormal and social psychology* 41 (1946), 25-36.

Zipf, George Kingsley: Cultural-chronological strata in speech. In: *Journal of abnormal and social psychology* 41 (1946), 351-355.

Zipf, George Kingsley: The P_1P_2 / D hypothesis: On the intercity movement of persons. In: *American sociological review* 11 (1946), 677-686.

Zipf, George Kingsley: Some determinants of the circulation of information. In: *American journal of psychology* 59 (1946), 401-421.

Zipf, George Kingsley: The psychology of language. In: Hariman, Philip Lawrence (ed.): *Encyclopedia of psychology*, New York, 2nd printing 1951, 332-341 [first printing 1946].

Zipf, George Kingsley: On Dr. Miller's contribution to the P_1P_2 / D hypothesis. In: *American journal of psychology* 60 (1947), 284-287. [ad Miller (1947)]

Zipf, George Kingsley: The frequency and diversity of business establishments and personal occupations. A study of social stereotypes and cultural roles. In: *Journal of psychology* 24 (1947), 139-148.

Zipf, George Kingsley: Prehistoric ‘cultural strata' in the evolution of Germanic: The case of Gothic. In: *Modern language notes* 62 (1947), 522-530.

Zipf, George Kingsley: The hypothesis of the ‘Minimum Equation' as a unifying social principle. With attempted synthesis. In: *American sociological review* 12 (1947), 627-650.

Zipf, George Kingsley: On the number, circulation-sizes and the probable purchasers of newspapers. In: *American journal of psychology* 61 (1948), 79-89.

Zipf, George Kingsley: *Human behavior and the principle of least effort*. An introduction to human ecology. New York: Hafner reprint, 1972 [First printing: Cambridge/Mass., Addison-Wesley, 1949].

Zipf, George Kingsley: Relative frequency and dynamic equilibrium in phonology and morphology. In: *Proceedings of the 6th international congress of linguists*. Paris, 1949, 391-408.

Zipf, George Kingsley and Rucker, Allen W.: How to set salary brackets that spur achievement. In: *Modern management* 9 (1949), 4-7.

Zipf, George Kingsley: The frequency-distribution of wages and the problem of labor unrest. In: *American journal of psychology* 29 (1950), 315-324.

Zipf, George Kingsley: Empiric regularities in the frequency-distribution of chemical manufacturers and chemical distributors by product-diversity in the U.S.A. In: *American journal of psychology* 30 (1950), 195-198.

Zipf, George Kingsley: Brand names and related social phenomena. In: *American journal of psychology* 63 (1950), 342-366.

Zipf, George Kingsley: Wage distribution and the problem of labor harmony. In: Sorokin, Pitirim Aleksandrovich (ed.): *Explorations in altruistic love and behavior*. Boston, 1950, 333-346.

Zipf, George Kingsley: A note on brand-names and related economic phenomena. In: *Econometrica*, 18, (1950), 260-263

Zipf, George Kingsley: Quantitative analysis of Sears, Roebuck and Company’s catalogue. In: *The journal of marketing*, July, 1950, pages 1 – 13.

Zipf, George Kingsley: Empiric regularities in the frequency-distribution of directorships in American corporations. In: *American psychologist* 5 (1950), 245 [= Announcing a speech at Sep 7th,1950].

Reviews

Relative frequency as a determinant of phonetic change (Zipf 1929)

Hermann, Eduard: George Kingsley Zipf: Relative frequency as a determinant of phonetic change. In: *Philologische Wochenschrift* 51 (1931), 598-603.

Kent, Roland G.: Relative frequency as a determinant of phonetic change. By George Kingsley Zipf. In: *Language* 6 (1930), 86-88.

Meillet, A.: G. Kingsley Zipf. Relative frequency as a determinant of phonetic change. In: *Bulletin de la société de linguistique de Paris* 31 (1931), 17.

Meriggi, Piero: Zipf, George Kingsley: Relative frequency as a determinant of phonetic change. In: *Indogermanische Forschungen* 50 (1932), 246-247.

Twaddell, W. Freeman: New light on phonetic change. In: *Monatshefte für deutschen Unterricht* 21 (1929), 230-237.

Richter, Elise: Zipf, George Kingsley: Relative frequency as a determinant of phonetic change. In: *Archiv für das Studium der neueren Sprachen* 157 (1930), 291-296.

Sütterlin, L.: George Kingsley Zipf: Relative frequency as a determinant of phonetic change. In: *Literaturblatt für germanische und romanische Philologie* 52 (1931), Sp. 241-243.

Selected studies of the principle of relative frequency in language (Zipf 1932)

Cohen, Marcel: Besprechung von Zipf: Selected studies of the principle of relative frequency in language. In: *Bulletin de la société de linguistique de Paris* 33/2 (1932), 10-11.

Malone, Kemp: Some linguistic studies of 1931 and 1932. In: *Modern Language Notes* 48 (1933), 378-396.

Prokosch, E.: George Kingsley Zipf: Selected studies of the principle of relative frequency in language. In: *Language* 9 (1933), 89-92.

The psycho-biology of language (Zipf 1935)

Cohen, Marcel : George Kingsley Zipf.- The psycho-biology of language. In: *Bulletin de la société de linguistique de Paris* 36 (1935), 8-11.

Empson, William: He lisped in numbers. In: *Spectator*, Feb. 14, 1936, 270.

García, Erica: Zipf, George K. The psycho-biology of language: An introduction to dynamic philology. Introduction by George A. Miller. In: *Romance philology* 22 (1968), 39-42.

Joos, Martin: Review of G. K. Zipf. The psycho-biology of language. In: *Language* 12 (1936), 196-210.

Jost, Karl: Besprechung von Zipf's Psycho-Biology. In: *Indogermanische Forschungen* 55 (1937), 139-142.

Jost, Karl: Entgegnung zu Zipfs Erwiderungen. In: *Indogermanische Forschungen* 56 (1938), 78-80.

Kühlwein, H. A. Wolfgang: George Kingsley Zipf, The psycho-biology of language: An introduction to dynamic philology. In: *Linguistics* 44 (1968), 98-99.

Průcha, Jan: Psychobiologická teorie jazyka. In: *Slovo a slovesnost*, 30 (1969), 96-98.

Thorndike, E. L.: George Kingsley Zipf. The psycho-biology of language. In: *Journal of educational psychology* 27 (1936), 391.

Trnka, Bohumil: George Kingsley Zipf: The psycho-biology of language. An introduction to dynamic philology.- Human behavior and the principle of least effort. An introduction to human ecology. In: *Philologica* 5 (1950), 3-5.

National Unity and Disunity; The Nation as a Bio-Social Organism (Zipf 1941)

Riemer, Svend; George Kingsley Zipf: National United and Disunity; The Nation as a Bio-Social Organism. In: *The American Journal of Sociology*, 48 (1942), 285-287.

[See also reply by Zipf („Letter to the editor“, 1943).]

Human behavior and the principle of least effort (Zipf 1949)

Bentley, M.: Human behavior and the principle of least effort. An introduction to human ecology. By George Kingsley Zipf. In: *American journal of psychology* 64 (1951), 149-150.

Chao, Yuen Ren: Human behavior and the principle of least effort. An introduction to human ecology. By George Kingsley Zipf. In: *Language* 26 (1959), 394-401.

Classe, A.: G. K. Zipf: Human behavior and the principle of least effort. An introduction to human ecology. In: *Archivum linguisticum* 2 (1950), 76-78.

Cohen, Marcel : George Kingsley Zipf.- Human behavior and the principle of least effort. An introduction to human ecology. In: *Bulletin de la société de linguistique de Paris* 46 (1950), 12-13.

Hudgins, Clarence V.: An integrating principle for human behavior. In: *American speech* 24 (1949), 293-295.

Martinet, André: George Kingsley Zipf, Human behavior and the principle of least effort. In: *Word* 5 (1949), 280-282.

Průcha, Jan: Psychobiologická teorie jazyka. In: *Slovo a slovesnost*, 30 (1969), 96-98.

Stewart, John Q.: Human behavior and the principle of least effort: An introduction to human ecology. In *Science*, 110 (1949), 669.

Trnka, Bohumil: George Kingsley Zipf: The psycho-biology of language. An introduction to dynamic philology.- Human behavior and the principle of least effort. An introduction to human ecology. In: *Philologica* 5 (1950), 3-5.

Walsh, Joseph L.: Human behavior and the principle of least effort. An introduction to human ecology, George Kingsley Zipf. In: *Scientific American*, August 1949, 56 – 58.

Others

Miller, George A.: Population, distance, and the circulation of information. In: *American journal of psychology*, 60 (1947), 276-284.

References

Altmann, Gabriel (1981). Zur Funktionalanalyse in der Linguistik. In: Esser, Jürgen; Hübler, Axel (eds.): *Forms and functions*. Tübingen: Narr, 25-32.

Birkhan, Helmut (1979). *Das „Zipfsche Gesetz", das schwache Präteritum und die germanische Lautverschiebung*. Sitzungsberichte der österreichischen Akademie der Wissenschaften, philosophisch-historische Klasse 348.

Crozier, William J., Rogers, Francis M. and Walsh, Joseph L. (1950). George Kingsley Zipf. *Harvard University Gazette XLVI/13, 81-82* [Obituary note].

Köhler, Reinhard (1986). *Struktur und Dynamik der Lexik* (= Quantitative linguistics 31). Bochum: Brockmeyer.

Lundberg, G. A. and Dodd, S. C. (1950). Obituary. In: *American sociological review* 15, p. 104.

Mandelbrot, Benoît (1987). *Die fraktale Geometrie der Natur*. Basel: Birkhäuser [Engl.: The fractal geometry of nature, 1977].

Miller, George A. (1968). Introduction. In: Zipf, G. K.: *The psycho-biology of language. An introduction to dynamic philology.* Cambridge/Mass.: M.I.T. Press. 2nd printing, iii-x.

Prün, Claudia (1999). G. K. Zipf's conception of language as an early prototype of synergetic linguistics. In: *Journal of quantitative linguistics* 6, 78-84.

Prün, Claudia (to appear). The work of G. K. Zipf. In: *Encyclopedia of quantitative linguistics*.

Van der Vlis, J. H., Heemstra, E. R. (1988). George Kingsley Zipf (1902-1950). In: dto.: *Geschiedenis van Kansrekening en Statistiek: 175-188*. Rijswijk.

Glottometrics 3, 2002, 11-18 *To honor G.K. Zipf*

George Kingsley Zipf: life, ideas, his law and informetrics

Ronald Rousseau[1]

Abstract. In this article we present a short biography of the linguist George Kingsley Zipf. We recall his work on the frequency of words in Chinese language and briefly discuss Zipf's principle of least effort. We mention his influence in the field of informetrics and end this contribution by highlighting some recent applications of Zipf's law in Internet research, geography and economics.

Keywords: G.K. Zipf, Zipf's law, word frequencies, informetrics

1. Introduction

We first recall the formulation of Zipf's law (Egghe & Rousseau, 1990; Rousseau & Rousseau, 1993). Zipf's law is a relation between the frequency of occurrence of an event and its rank when the events are ranked with respect to the frequency of occurrence (the most frequent one first). More precisely, when r denotes the rank, and $f(r)$ denotes the frequency of occurrence of the event at rank r, Zipf's equation states that:

(1) $f(r).r = C$

where C is a constant. The equation

(2) $f(r) = C / r^{\beta}$

where β is a positive parameter, is usually called the generalised form of Zipf's law.

The following expression, relating the frequency of occurrence of an event (y) and the number of different events occurring with that frequency ($g(y)$), is – at least in the field of informetrics - known as Lotka's law (Lotka, 1926):

(3) $g(y) = A / y^{\alpha}$

where A and α are parameters. The special case of $\alpha = 2$ has a special historical interest. The relation is then known as Lotka's inverse square law. Lotka's law clearly is a power law.

Perhaps this is the right place to introduce readers working in linguistics and other fields to the term *informetrics*. According to Wilson (1999), informetrics is the quantitative study of collections of potentially informative text, directed to the scientific understanding of in-forming processes at the social level. It includes the study of bibliographies, reference lists, journal use (as in citation studies), and forms of scientific collaboration (as in co-authorship analyses). Its basic techniques are mathematical modelling and statistical analysis (Egghe & Rousseau, 1990). An important subfield of informetrics is the study of the so-called in-

[1] Address correspondence to: Ronald Rousseau, KHBO – Faculty of Industrial Sciences and Technology, Zeedijk 101, B-8400 Oostende, Belgium. E-mail: Ronald.Rousseau@kh.khbo.be

formetric (or bibliometric) laws. These regularities describe e.g. the frequency distribution of author production (Lotka's law), or journal publication patterns per subfield (Bradford's law). All informetric laws have been shown to be closely related (in mathematical form) to Zipf's law (Egghe & Rousseau, 1990; Wilson, 1999). Hence, it is no surprise to find out that the term 'Zipf's law" is well-known in the field of informetrics. We will return to this in Section 5.

2. A short biography of George Kingsley Zipf (Miller, 1965; Prün & Zipf, 2002)

George Kingsley Zipf was born in Freeport, Illinois on January 7, 1902. He graduated summa cum laude from Harvard College in 1924 and spent the following year in Germany, studying at Bonn and Berlin. He returned to Harvard and received his Ph.D. in comparative philology in 1930 (Zipf, 1929). He became instructor in German until 1936, assistant professor until 1939, and university lecturer until his death in 1950.

His Ph.D. dissertation was concerned with relative frequency of use as a determinant of phonetic change in the evolution of language. His book *The Psycho-Biology of Language*, published in 1935 by the Houghton Mifflin Company, was his first attempt to relate linguistic ideas to the real-life experiences of men (Zipf, 1935). In 1941 he published *National Unity and Disunity,* which applied his statistical methods to the study of the sizes of cities and movements of population (Zipf, 1941). His most ambitious work, *Human Behavior and the Principle of Least Effort* (Zipf, 1949) appeared one year before his premature death. This book was a further study of semantics, psychology, sociology and geography. It abounds with illustrations of the probability distributions he first noticed in his statistical studies of vocabulary (Miller, 1965).

In the words of Miller (1965) Zipf was ' that kind of man who would take roses apart to count their petals'. Indeed, he analysed masterpieces of the world literature by breaking them down to words, and sometimes even to syllables and phonemes.

Sometimes the term 'word' is used as a symbol of freedom, yet, as Zipf showed, word usage is not free, but determined by a strong statistical law. His scientific contribution is honoured, not by a Nobel prize or a similar scientific award, but – as is surely appropriate – by language itself. The terms 'Zipf's law', 'Zipf plots' and 'Zipf curves' are eponyms (things called after their inventor or, in this case, after the person who made this phenomenon known). Indeed, he devoted most of his intellectual life to exploring and explaining the regularities we now refer to as Zipf's law. I must add a word of caution here. Some authors mean by the term 'Zipf's law' what we have referred to as Lotka's law. Other authors are aware of a possible confusion and use terms as 'Zipf's first law' and 'Zipf's second law', or 'the dual law of Zipf'. How this confusion has arisen will be explained shortly. Anyway we prefer to restrict the term 'Zipf's law' to the rank-frequency form.

In 1932 Harvard University Press published *'Selected Studies of the Principle of Relative Frequency in Language',* a report written by George K. Zipf. In this report he investigated the occurrences of words in Latin (Plautus) and in Chinese texts. To perform the analysis of the Chinese texts Zipf needed the help of two Chinese collaborators: Mr. Kan Yu Wang and Mr. H.Y. Chang (Zipf, 1932). We may say that Mr. Wang and Mr. Chang were to Zipf, as was Mr. Lancaster Jones to Samuel Bradford, the famous British librarian and discoverer of a law which is closely related to Lotka's and Zipf's (Bradford, 1934). As these investigations are among the first studies in quantitative linguistics, this shows that, through these men, China played a pioneering role in the history of quantitative linguistics, hence also in informetrics.

Zipf and his collaborators analysed twenty fragments of Chinese text, each containing about thousand syllables, taken from twenty different sources. This yielded a corpus of 20,000 syllables. In collaboration with Qiaoqiao Zhang, then a Ph.D. student at the City University of

London (UK), and now working for CAB International, I re-examined Zipf's data on Chinese word frequencies (Rousseau & Zhang, 1992). We found that, as is usually the case, Zipf's law did not fit the Chinese language data (in a statistical sense). The following cumulative rank-frequency function with a, b and c as parameters, gave, however, an excellent fit:

(4) $R(r) = c + a \ln(1 + br)$

The occurrence of the parameter *c* in this Bradford-like function (4) is due to the fact that in Chinese the most frequently used word, which is the word ‘de’ (referring to a genitive), is used much more frequently than expected, based on Zipfian (or Bradfordian) statistics.

3. The origin of 'Zipf's law'

In his 1932 book Zipf made use only of Lotka's inverse square law. He did this without referring to Lotka who had published his work six years earlier. However, this is the reason why some people refer to Lotka's law as Zipf's (first) law. Furthermore, he did not perform any statistical test and did not include the most frequently occurring words. The main point is that he did not use the rank-frequency form (what we nowadays call 'Zipf's law'). It is only three years later, in 1935, when publishing *'The psycho-biology of language'* that he wrote, after first having presented Lotka's inverse square distribution:

'There is, however, another method of viewing and plotting these frequency distributions which is less dependent upon the size of the bulk and which reveals an additional feature. As suggested by a friend, one can consider the words of a vocabulary as ranked in the order of their frequency, e.g. the first most frequent word, the second most frequent, the third most frequent, the five-hundreth most frequent, the thousandth most frequent, etc. We can indicate on the abscissa of a double logarithmic chart the number of the word in the series and on the ordinate its frequency.'

So it is clear that Zipf did not invent 'Zipf's law’, but that a friend showed him the mathematical relation. We have, however, no clue who this mysterious friend might have been. In the works of Zipf we have examined we found only one person whom he addresses as 'my friend', namely R.Y. Chao. Professor Chao is also mentioned in his 1932 book, so the conjecture that this friend is R. Y. Chao is not too far-fetched. Note that in *'Selected Studies'* and *'The Psycho-biology of language'* there is no reference to Estoup or to Pareto, although Estoup is mentioned in Zipf's very first publication on the relative frequency of words. Later, in *'Human behavior and the principle of least effort'* Zipf did refer to Lotka, Estoup and Pareto. It is also interesting that besides to Estoup, Zipf, correctly, refers in his later work also to Godfrey Dewey (1923) and to E.V. Condon (1928) as scientists that have noted the hyperbolic nature of the frequency of word usage. Perhaps it was to one of these works that the mysterious friend drew Zipf's attention. In this context we should mention that Petruszewycz (1973) thinks that the mysterious friend could have been Alan N. Holden. The reason for this conjecture is that Holden is mentioned in an earlier work of Zipf, and that he worked for the same company as Condon, namely the Bell Telephone Company. Our suggestion as well as Petruszewycz' is a pure speculation that may be completely wrong.

4. The principle of least effort in the context of words

The previous paragraph certainly shows that Zipf was not the first to notice power law relations in the study of use frequencies of words. His many publications, however, and hypotheses to explain this ubiquitous phenomenon, brought the matter to the attention of everyone with an interest in science. Zipf found these curves to have a uniform shape under a remarkable variety of circumstances. Noticing the same regularities in all languages and even beyond, he thought that it followed from a universal property of the human mind. He searched for a principle of least effort that would explain the equilibrium between uniformity and diversity in the use of words. Indeed, too much uniformity makes a sentence incomprehensible, like in 'I saw a thing this morning when I was doing my thing in the thing'. On the other hand, also too much diversity makes a sentence unclear. An historian telling about ancient times and using the proper word for every object, will soon lose his audience as laymen do not know all the words for objects that for long have ceased to exist. So a vocabulary should not be too poor or too rich: it must be in some kind of equi-librium state.

Now, many years later, we know that Zipf's explanation was probably wrong. The regularities he observed do not follow from a universal property of the human mind, but they are rather a consequence of the laws of probability. Zipf curves are merely expressing a consequence of regarding a message source as a stochastic process. This was first demonstrated by Benoit Mandelbrot, the scientist best known for the invention of fractals (1954).

Zipf's ideas about the principle of least effort were developed as follows. Begin with the assumption that a concept is a bundle of semantic features or 'genes of meaning', in the words of Zipf. If a particular bundle occurs frequently enough in a certain community they will assign to it a phonological representation, i.e. a word. If it occurs infrequently, no specific word will be available, so when this concept arises people will have to describe it using a string of words or a phrase. For instance, we have no specific word to describe 'an elegant woman doing research in informetrics'. As language evolves, a relation evolves between the lengths of words and the frequency of occurrences of the phenomena they describe. The more a word is used the shorter it becomes. Examples abound, like for instance:

moving pictures – movies
telefacsimile – telefax – fax
acquired immuno deficiency syndrome – AIDS

These examples show that the equilibrium hypothesis and the principle of least effort have some intuitive value. So, although the purely mathematical, say stochastic approach leads to the correct distribution function, and some of Zipf's vague ideas were incorporated in the mathematical derivations (the exponent β in equation (2) turned out to be related to the fractal dimension of the text), I do not think that mathematics can give the ultimate explanation about the use of words in texts or in spoken language.

5. A relation between Zipf's law and Lotka's

We recall that Lotka's law relates the frequency of occurrence of an event (y) and the number of different events occurring with that frequency, denoted as $g(y)$ (Lotka, 1926). It is expressed as:

(3) $g(y) = A / y^{\alpha}$

where A and α are parameters. Here we will consider equation (3) as an expression of absolute frequencies: $g(y)$ denotes the number of sources with production y. It is not considered as a statistical discrete probability distribution (which would sum to one). In our context $g(y)$ may denote the number of words that occur y times (linguistic interpretation), or the number of authors that have written y articles over a certain period of time (informetric interpretation). Moreover, we assume that y takes values between 1 and a maximum value, denoted as y_m.

The next step in this mathematical modelling exercise is to consider y as a continuous variable, not anymore as a discrete one. This makes it possible to introduce ranks. Indeed, given equation (3) we obtain the rank of the source with production y (denoted as $r(y)$), as the total number (i.e. the sum) of sources with production larger than y. In a continuous setting this can be expressed (writing an integral instead of a sum) as:

$$(5) \qquad r(y) = \int_{y}^{y_m} g(t)dt + 1$$

Here the term "+1" follows from the fact that the source with the highest production must have rank 1, and not rank 0. Taking the derivative of equation (5) yields the following relation:

$$(6) \qquad \frac{d\,r(y)}{dy} = r'(y) = -\,g(y)$$

Now, Zipf's law (in its general form) states that

$$y = f(r) = \frac{C}{r^{\beta}}$$

From this expression we derive that

$$(7) \qquad r(y) = \left(\frac{C}{y}\right)^{1/\beta}$$

and taking derivatives with respect to y leads to:

$$(8) \qquad r'(y) = \frac{1}{\beta}\left(\frac{C}{y}\right)^{\frac{1}{\beta}-1}\left(-\frac{C}{y^2}\right) = -\frac{C_0}{y^{\frac{1}{\beta}+1}}$$

with

$$C_0 = \frac{1}{\beta}C^{\frac{1}{\beta}}.$$

Comparing equation (6) with equation (8) shows that Zipf's law corresponds to Lotka's law with $\alpha = \frac{1}{\beta}+1$. In particular, the original form of Zipf's law ($\beta = 1$) corresponds to Lotka's square law ($\alpha = 2$). Note that this relation only goes in one direction: starting with 'Zipf' leads

to ‘Lotka’. Starting with ‘Lotka’ does not in general lead to ‘Zipf’ but rather to a generalisation due to Mandelbrot (Rousseau, 1990).

6. Recent developments related to Zipf's law

In this section we would like to mention some recent developments related to the application of Zipf's law. Zipf showed that, by and large, his law held for words, syllables and morphemes. Consequently, it is natural to ask if the law also holds for pairs of words. Empirical evidence seems to suggest that it does (although never in a purely statistical way). My colleague Prof. Egghe devised a mathematical argument that it, in fact, does not, but that the exact relation can be approximated by a power law (Egghe, 1999). He extended his investigations to parts of words, namely to the study of N-grams (Egghe, 2000). Observe that such behaviour reflects the real Zipfian spirit! (Recall the quote about roses and petals.)

Note also that as early as 1962 Herdan (1962) wrote a severe criticism on the Zipf-Mandelbrot approach to linguistics. Moreover, statistical arguments and model-theoretical considerations should not be confused.

In recent times Zipf's law has been tested on the Internet. It turned out that popularity of Internet pages is described according to Zipf's law. This fact can be used to design better address cache tables (Aida et al., 1998). Zipf’s law was also the source of a lively debate related to the structure of DNA. It was claimed (Mantegna et al., 1994) that Zipf's law shows the difference between coding and non-coding DNA as non-coding (so-called junk) DNA fits Zipf's law much better than coding DNA. This would mean, according to the authors, that non-coding regions of DNA may carry new biological information. Yet, this does not mean that junk DNA is a kind of language. Other scientists (Chatzidimitriou-Dreismann et al., 1996), however, have shown that this distinction is not universal and lacks all biological basis.

Zipf's studies on city sizes still lead to new developments in geographical and economical studies (Brakman et al., 1998; Gabaix, 1999; Ioannides & Overman, 2002). Note also that, according to Marsili and Zhang (1998), cities in countries such as China and the former USSR do not follow Zipf's law.

7. Conclusion

We have shown that although the facts observed by Zipf were accurately reported, it seems that his explanations were not quite correct. Yet, in the words of George A. Miller

> *'Zipf belongs among those rare but stimulating men whose failures are more profitable than most men's successes'.*

It can not be denied that Zipf has had and still has a large influence on developments in many scientific fields such as linguistics, geography, informetrics, sociology, economics, physics, biology and many more.

Note. This article is a thoroughly revised version of (Rousseau, 2000), published in Chinese.

References

Aida, M., Takahashi, N. and Abe, T. (1998). A proposal of dual Zipfian model for describing HTTP access trends and its application to address cache design. *IEICE Transactions on Communications,* 81(7), 1475-1485.

Bradford, S. (1934). Sources of information on specific subjects. *Engineering*. 137, 85-86.

Brakman, S., Garretsen, H. and Van Marrewijk, C. (1998). Uneven spatial distribution in modern location-trade theory (in Dutch). *Economisch en Sociaal Tijdschrift,* 52(4), 479-507.

Chatzidimitriou-Dreismann, C.A., Streffer, R.M.F. and Larhamar, D. (1996). Lack of biological significance in the 'lingusitic' features' of noncoding DNA – a quantitative analysis. *Nucleic Acid Research*, 24(9), 1676-1681.

Condon, E.V. (1928). Statistics of vocabulary. *Science*, 67, p. 300.

Dewey, G. (1923). *Relative Frequency of English Speech Sounds*. Cambridge (MA): Harvard University Press.

Egghe, L. (1999). On the law of Zipf-Mandelbrot for multi-word phrases. *Journal of the American Society of Information Sciences,* 50, 233-241.

Egghe, L. (2000). The distribution of N-grams. *Scientometrics*, 47, 237-252.

Egghe, L. and Rousseau, R. (1990). *Introduction to Informetrics*. Amsterdam: Elsevier.

Gabaix, X. (1999). Zipf's law and the growth of cities. *American Economic Review*, 89, 129-132.

Herdan, G. (1962). *The calculus of linguistic observations*. 's Gravenhage: Mouton & Co.

Ioannides, Y.M. and Overman, H.G. (2002). Zipf's law for cities: an empirical examination. *Regional Science and Urban Economics* (to appear).

Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16, 317-323.

Mandelbrot, B. (1954). Structure formelle des textes et communications: deux études. *Word*, 10, 1-27.

Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S. , Peng C.-K., Simons, M. and Stanley, H.E. (1994). Linguistic features of noncoding sequences. *Physical Review Letters,* 73 (23), 3169-3172.

Marsili, M. and Zhang, Y.-C. (1998). Interacting individuals leading to Zipf's law. *Physical Review Letters*, 80(12), 2741-2744.

Miller, G. (1965). Introduction to *The Psycho-biology of Language: An Introduction to Dynamic Philology, by G.K. Zipf*. Cambridge (MA): MIT Press.

Petruszewycz, M. (1973). L'histoire de la loi d'Estoup-Zipf: documents. *Mathématiques et Sciences Humaines,* 11(44), 41-56.

Prün, C. and Zipf, R. (2002). Biographical notes on G.K. Zipf. *Glottometrics*, 3, 11-20.

Rousseau, R. (1990). Relations between continuous versions of bibliometric laws. *Journal of the American Society for Information Science*, 41(3), 197-203.

Rousseau, R. (2000). George Kingsley Zipf: life, ideas and recent developments of his theories. In: *R &D Evaluation and Indicators* (Jiang Guo-Hua, ed.), Red Flag Publishing House, p. 458-469.

Rousseau, R. and Rousseau, S. (1993). Informetric distributions: a tutorial review. *Canadian Journal of Information and Library Science*, 18(2), 51-63.

Rousseau, R. and Zhang, Q. (1992). Zipf's data on the frequency of Chinese words revisited. *Scientometrics*, 24(2), 201-220.

Wilson, C.S. (1999). Informetrics. *Annual Review of Information Science and Technology*, 34, 107-247.

Zipf, G.K. (1929). *Relative frequency as a determinant of phonetic change*. Harvard Studies in Classical Philology 40. Cambridge (MA): Harvard University Press.

Zipf, G.K. (1932). *Selected studies of the principle of relative frequency in language.* Cambridge (MA): Harvard University Press.

Zipf, G.K. (1935)*, The psycho-biology of language: An introduction to dynamic philology*. Houghton Mifflin. Reprinted: (1968). Cambridge,Mass.: The M.I.T. Press.

Zipf, G.K. (1941). *National unity and disunity. The nation as a bio-social organism.* Bloomington (IN): Princeton Press.

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Cambridge (Mass.): Addison-Wesley Press.

Glottometrics 3, 2002, 19-26 *To honor G.K. Zipf*

Zipfian linguistics

Gabriel Altmann[1]

> *Zipf belongs among those rare but stimulating men whose failures are more profitable than most men´s successes.* G.A. Miller (1968: vii)

Abstract. Zipf´s ideas are the foundation stones of modern quantitative linguistics. Many of them have been developed both mathematically and in content, but his work seems to be inexhaustible. His influence is not restricted to linguistics but incessantly penetrates other sciences.

Keywords: Zipf, language laws, language synergetics, ranking, frequency, relations

In the two last centuries linguistics have experienced a colorful variety of different paradigms, research programmes and local schools and, especially in the second half of the 20th century, a stormy rise of new linguistic disciplines. All of them have extended our look at language and converted linguists to specialists in narrow domains. They all have a common feature: they came into existence like an explosion and soon began to die, even if some of them conquered the whole globus for a time.

One could simply reduce the "why?"-question of this movement to the fact that this is the normal way of science, but this generalization cannot be considered an explanation. There is also no single cause to which their decay could be reduced. Usually there are some internal controversies or anomalies, inappropriate methods or narrow methodologies which, taken together, contribute to the fall of a doctrine. In the evolution of linguistics there is a complex of reasons, the main ones of which are as follows. The majority of disciplines followed a very practical aim: it wanted to *describe* special parts of language. It was either the development of language, its phonological, morphological, syntactical structure or its spatial or social dimension, etc. The conceptual means chosen were adequate only for the capturing of the surface of language. The deeper dynamic relations remained outside of interest as well as outside of reach of the methods used. One was not ready to deal with entities having no direct connection to grammar or semantics, e.g. frequency of occurrence or length of language units. They were simply ignored, no necessity of their systematization was or is seen. However, just as one cannot explain the functioning of a body taking only one organ into consideration, one cannot explain the functioning of language without considering several of its aspects simultaneously. Though e.g. length and frequency seen superficially have nothing to do with surface phenomena like grammar, they are missing chain links without which no theory building in linguistics is possible. But without theory construction every discipline remains at a protoscientific level.

Surely, many trials to explain several linguistic phenomena were made, but the chosen means were not sufficient, they could not satisfy the claims of natural sciences. Changes from

[1] Address correspondence to: Gabriel Altmann, Stüttinghauser Ringstr. 44, D – 58515 Lüdenscheid, Germany. E-mail: RAM-Verlag@t-online.de

A to B were described, so were rules of succession and cooccurrence, variation and dispersion, but no scientist was ready to search for the dynamics behind these phenomena.

In the period of flourishing structuralism in the twenties of the 20th century G.K. Zipf ventured to take a look at the internal dynamics of language and discovered mechanisms which were responsible for different language phenomena. Of course, already before him it was assumed that "there is something" caring for motion in language, but at that time the aversion towards figures and counting was in many linguistic circles as intense as it is today. Zipf was noticed, he and his errors were criticized, but nobody drew consequences from his research.

While the particular linguistic "schools" reached a deadlock, some slowlier, some quicklier, and even objections to the known "Zipf´s law" were raised, a thorn got stuck in the flesh of linguistics which could not be removed any more. No ignoring and no gestures of dismissal because of irrelevance helped. Zipf discovered something new that did not conform to the general trend. As soon as natural scientists discovered Zipf, the battle was definitively brought to end. Zipf "won" and 100 years after his birth his popularity increases continuously. Mathematicians strive for sounder footing of his laws, physicists and other natural scientists find more and more new analogies in their disciplines (cf. Schoeder 1991; Gell-Mann 1994; Bak 1999; Naranan, Balasubrahmanyan 1998, 2000), quantitative linguists try hard to systematize his ideas (cf. Köhler 1986; Prün 1995) and find with their help a connection to the general trend in sciences. The common heritage spread especially among natural scientists encompasses merely the Zipfian idea that the frequencies of entities ranked according to their size are functionally joined with their ranks. The fact that this function was a power function has proved to be a lucky coincidence since Zipf did not care either for convergence or norming or the necessary truncation on the right side, he merely sought an agreement with data. Such data can be found frequently, but still more frequently data that do not follow the simple power curve. Thus Zipf confronted the scientists with two challenges:

(i) Find a curve or a probability distribution capturing better the given ranking. This problem assumed gigantic proportions. It first involved mathematicians interested in linguistics. Since these mathematicians still had other interests (the most famous case is Benoit Mandelbrot), it was slowly but continuously transferred to other sciences. Since rank distributions (especially those with long tails) can easily be tranformed in a frequency spectrum, it was noticed – Rapoport (1982) assumed it, but still did not know – that ranking and frequency distribution are only two faces of the same coin and the circle was closed. A Zipfian frequency spectrum says that "great" events are seldom, "small" ones on the contrary more often, and the rest lying between them can be captured with the aid of a power curve. Often it is satisfactory, often it is not. It depends on the kind of data, its genesis, the way of counting and measuring, etc. whether a "good" fit is obtained. However, the fact that it was found in self-organization, with growth problems, in economy, with WEB-phenomena, etc. granted it immortality. It would be very troublesome for anybody to present the history of this problem in greater detail. Wentian Li divided his bibliography of Zipf´s law [http://linkage.rockefeller.edu/wli/zipf/] into 17 chapters, but some sciences like sociology, psychology, musicology are still missing. Fermat´s last theorem whose adventurous history was a good theme for novels merely kept mathematicians busy, but Zipf´s laws dispersed like an avalanche into all domains of science. There are many variants of the original forms of these laws because scientists are persuaded that innumerable boundary conditions create an extensive landscape of attractors.

(ii) The second problem to be solved was the foundation of this regularity. Why do we use elements of a class of language entities in well proportioned doses? Why is it the case even in nature? The answers are so numerous that they create a separate discipline. We are astonished that physicists often derive a Zipf Law from thermodynamics, the researchers in self-oganized

criticality find it in sand-piles, in earthquakes, in the extinction of species and in other domains. A lot of work in this domain has been done by Russian scientists but unfortunately their works are published in Russian and are only available to a small part of the scientific community. In linguistics it is believed that a class of entities is "correctly" established if the ranked frequencies of the elements abide by a ranking law. A preliminary study brought to light almost 30 formulas: some are special forms of a more general formula, but there are several ad hoc empirical formulas, too. In linguistics it is partially assumed that ranking arises automatically in the communication process and that it also controls the perception and decoding of speech. It possibly finds its way even in the storing of words in the brain. What is good enough for one discipline need not be good for another one. Natural scientists must reach for other interpretations, but in general it is to be assumed that at this point Zipf opened a door which must be passed by all of us. Viribus unitis.

The extent of Zipf´s fame in natural sciences would not be smaller if he had discovered the frequency laws only. But for linguistics his importance is definitively much greater. He did not stop after finding that frequencies and ranking abide by regularities but joined the frequency with other properties of language. Frequency and length, frequency and polysemy, frequency and conspicuity, frequency and age, etc. are only some of the connections known and examined today in synergetic linguistics. Zipf was no mathematician and in his time systems theory had little importance in linguistics. As a lonely fighter he fell between two stools and elaborated a conception of relationships and links which today are the bases of the well developed synergetic linguistics. There will be a time in the future when it is rightly called "Zipfian linguistics", even if today it is presented "in different clothes" as we are accustomed to Zipf.

In all these connections he sought a common, always effective factor and found his famous "principle of least effort". If e.g. a word/phrase is used more frequently, then the associated effort can be reduced by shortening the word/phrase. One can observe it easily in greetings. However, thereby merely the speaker's effort gets reduced but the process must not result in a total disappearance of the word/phrase. This is, of course, the concern of the hearer who gets decoding problems because of reduced redundancy. Thus he forces the speaker to a compromise. This is the source of self-regulation in all domains, controlled by the requirement of least effort of *both* participants in communication. Owing to research by Köhler (1986, 1987, 1988, 1990a,b,c,d, 1991, 1992a,b, 1999) and his pupils (Krott 1994; Prün 1995) we know today that this principle is merely one of many that must be fulfilled for a smooth course of communication. They play the role of order parameters, they are used as parameters in the curves but their measurability is a very complex problem. Preliminarily they are being estimated from data but their interpretation is satisfactory only in a few cases.

If frequency is embedded in control cycles in which a negative feedback is necessary on equilibrium grounds, it is merely one of many influences having a simultaneous effect on other properties. Zipf´s seminal discovery of the role of frequency which was put by him in the center of our attention (and still more by G. Herdan 1960 or the modern typologists, see Bybee, Hopper 2001, Fenk-Oczlon 1989) obtained an appropriate position in synergetic linguistics, often as primus inter pares, but not as a unique moving force. Herdan´s assumption that words do not only have fixed formal and semantic properties but also fixed population probabilities which can be approximated by relative frequencies in the samples turned out to be illusory: Orlov (1982) showed that there are no basic populations in language, every text is an actual population in which the flow of information depends on the "planned size", the so called "Zipf size". The sample frequencies of elements from great inventories, e.g. that of words, are no good approximations since frequencies change daily and the great mass of frequencies, namely that of spoken language, can never be captured. If all words used by Shakespeare are counted, the result is not a population that can be called "Shakespeare" but a

mixture of smaller, very heterogeneous populations that does not allow reliable inference. Any text is generated with many boundary conditions for which there must be a place in language laws. However, if we blurred the boundary conditions, we would have the same case as if we tried to compute the average rate of fall from a fall of a small iron ball and that of a big sheet of paper having the same weight.

What is said above enables us to draw at least two conclusions: (i) there are no isolated properties in language, each property is linked with at least one other property; (ii) language laws hold only for homogeneous data.

From the first conclusion it can be deduced hypothetically that all properties of language are linked with one another but a direct link cannot always be proven. However, the connection can be reconstructed by means of properties lying between them and with correct application of the ceteris paribus condition which does not always allow everything.

From the second conclusion it can be deduced that mixtures of texts, e.g. frequency dictionaries of a whole language, are adequate for practical but not for theoretical purposes. The mixing of texts fatally violates the ceteris paribus principle which is dealt with very carefully in synergetic linguistics (cf. e.g. Prün 1995: 56-57). Mixing of texts blurs the boundary conditions which is false from the methodological point of view.

Zipf mostly considered properties which did not play any relevant role in orthodox linguistics and tried to examine their connections. The list presented below is surely an incomplete one of properties, requirements and processes for which he tried to set up hypotheses. He mixed some properties, i.e. he used different terms for the same thing:

Frequency, rank, diversification, unification, inventory minimization, intervals between occurrences of identical entities, abbreviation, equilibrium, least effort, formation of configurations, conspicuity of sounds, articulation effort, intensity of sounds, complexity of sounds, magnitude of sounds, size of the phoneme inventory, phoneme frequency, information value, comprehensibility, thresholds, Morpheme length, size of morpheme inventory, morpheme frequency, crystallization of configurations, specificity, morpheme accent, predictability, independence of usage, word length, word frequency, word accent, number of meanings, etymological strata, emotional intensity, semantic specificity, crystallisation degree, word repetition with intervals, geographic and social diversification, distinctness of meaning, degree of inflection, word classes, sentence accent, sentence core, word order.

Most of them do not even appear in current introductions to linguistics. For the majority of linguists it is a foreign world. But in Köhler´s synergetic linguistics the old concepts have been specified and new ones have been coined. An excellent survey of models and criticism can be found in Prün (1995) but we are far from capturing and systematizing all of Zipf´s hypotheses.

Zipf´s discoveries have two fascinating aspects. First, he was able to show that many properties considered irrelevant in classical linguistics are firmly embedded in the language self-regulation, that a theory can be constructed only if all of them are considered and if phenomena are looked at that do not lie merely on the surface of language. As a matter of fact, he is the first language theoretician, because he saw language in its dynamics. Statics, firmly established by de Saussure´s influence until the end of the century, did not interest him.

The second aspect which is still more important is his "discovery" of the role of the power curve or the zeta distribution frequently called Zipf distribution (cf. Wimmer, Altmann 1999). Whatever connection he examined, at the end he mostly came to the power curve. One could say that his mathematics was poor but this fact is irrelevant today. One hundred years after his birth we understand the importance of his discovery if we look at other sciences and ascertain that "Zipf´s law" is established in about twenty disciplines. The power curve is the central

concept of self-organized criticality, the most actual branch in the domain of self-organization. The power curve and its modifications are the result of the operation of a mechanism controlling different processes in nature and driving forward the evolution. This is the door through which linguistics can join the family of better developed sciences[2]. Today, Zipf´s problems are – necessarily – treated in a more complex way than 50 years ago but in language theory there is no way getting round Zipf.

Zipf´s work is a paradigm, a research program, a mountain of scientific problems[3] – whatever the name of this enterprise in the philosophy of science. Let us try to draw a general, summarizing image. In the "dynamic philology" or "Zipfian" or "synergetic" linguistics we are always concerned with relations between different entities. They are the heart of Zipf´s doctrine. Until now the following kinds of relations have been found:

(i) *Hierarchic* relationships between entities of a higher and those of a lower level, e.g. between the length of the construct and that of its components known as Menzerath´s law. The law enables us to find and establish text or language levels (cf. Altmann, Schwibbe 1989; Hřebíček 1995, 1997).

(ii) *Collateral* relations holding either between the properties of one and the same entity or between the properties of different ones. The first kind encompasses e.g. the relations between the length of the word, its frequency, its number of meanings, emotionality etc. The second kind encompasses e.g. the relation between the size of the phoneme inventory and mean word length in language or between sentence complexity and text readability. The relations are usually mutual but in models the circumstances are simplified. This is the central part of language synergetics (cf. Köhler 1986; Krott 1994; Prün 1995).

(iii) *Sequential* relations between identical entities in text resulting from the linear ordering of the text. There are many different models and examinations in this domain but merely a few substantial hypotheses relating the kind of repetition to another aspects of text (cf. Altmann 1988, Pawlowski 2001).

(iv) *Historical* relations capturing not merely the course of a change in time but bringing it also in relation with different factors and collateral and hierarchic influences. That is, not merely answers to "how"-questions but also to "why"-questions which are, of course more seldom (cf. Brainerd 1983).

(v) *Diversificational* relations capturing the synchroneous dialectal, idiolectal, socio-lectal etc. variability of language phenomena. Since here the "causes" are very much hidden, they can just abstractly be reduced to the principle of least effort and are usually modelled by stochastic processes (cf. Rothe 1991; Altmann 2003).

(vi) *Frequency distributions* of entities. Each property of a language entity in text or dictionary follows a probability distribution – which is neither uniform nor normal – controlled internally by a kind of self-regulation and in turn leveling out excesses in other control cycles. It is a loop-like relationship. Usually it is examined at the beginning of research. The research is especially advanced in the domain of the length of language entities (cf. Best 2001). This is the source of the famous Zipf laws also encompassing the ranking of frequencies (cf. Chitashvili, Baayen 1993; Baayen 2001). The frequency is assumed to be the strongest impulse of self-organization but wide-ranging examinations are still missing.

The complete picture of these relations is displayed in Fig. 1. Of course, any entity is simultaneously involved in all these relations, but it is not possible to examine them all simultaneously. Nevertheless, the more relations are taken simultaneously into account the better the fit of the pertinent curves.

---

[2] "Dynamic Philology has the ultimate goal of bringing the study of language more into line with the exact sciences." (Zipf 1968: 3).

[3] Köhler´s Bibliography (1995) contains 6332 items, today there is about the double number of publications.

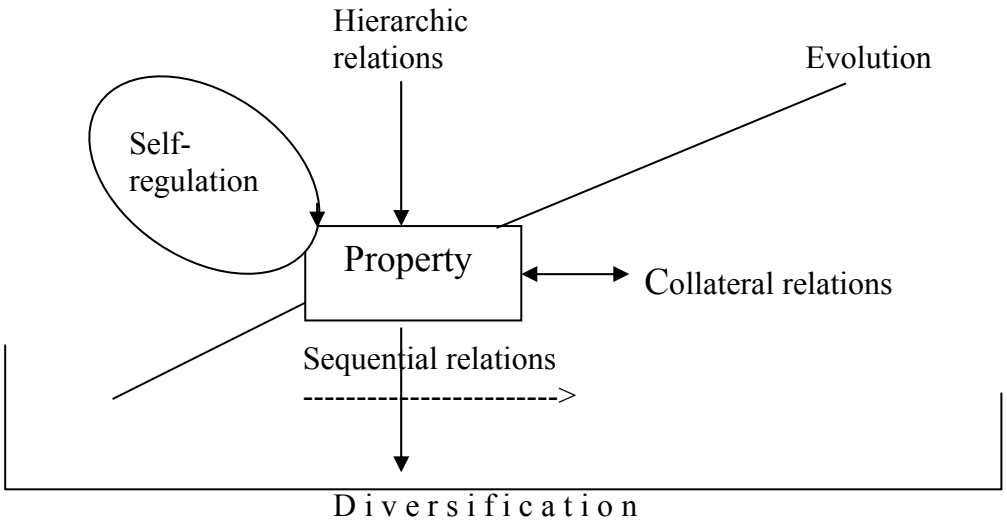


Figure 1. The relation between language entities

Besides the hierarchical relations Zipf´s works comprise fundamental ideas on all of these aspects. However, the majority of them still waits for their modern treatment, several new relations can be found in investigations on typology, universals, psycholinguistics and quantitative linguistics.

Though the number of publications in Zipfian linguistics increases exponentially, progress will be very slow for the following reasons:

(i) New relations must be discovered first and then embedded in the existing network of interrelations. Every step will be more complex, more difficult not only on mathematical grounds but especially on empirical ones. The model of any new relation must be tested on many texts or languages which requires teams of collaborators. This shows the necessity of publishing even raw data in order to avoid unnecessary work.

(ii) Every new discovered relation – if it is not merely an input – radically changes the whole system expressed by formulas. Even with simple systems treated in synergetic linguistics up to now it has been found that the simplification is sometimes very drastic: great parts of systems are considered constant (ceteris paribus) in order to obtain simple formulas. At last, as a matter of fact, single input-output systems are obtained whose results are not very satisfactory.

(iii) Still worse, in language one does not operate with "blind" natural forces. Even human requirements are introduced, e.g. minimization of production or memory effort and many others, whose meaning is intuitively clear, but whose operationalization and measurement must always be postponed. If they are used as parameters, they are estimated from data, but do not have a unique a posteriori interpretation; or they are pooled and considered constant, a procedure of drastic simplification distorting the results and forcing us to data manipulation.

(iv) Mathematically, both multiple input-output systems and partial differential equations as well as fuzzy reasoning will be necessary to cope with the complex language reality. Thus the mathematical perspective of Zipfian linguists is not very rosy. While mathematician must argue "merely" with their own constructs and physicists tackle "merely" with physical realities, Zipfian linguists will be forced to grapple with biological, psychological, neurological, even aesthetical realities. They will be forced to leave the safe ground of linguistics and play the role of wanderers between worlds.

(v) If language is merely a self-regulating system, what is the meaning of self-organization? Where does it intervene? Practically everywhere. Of course, language itself does not do anything, we are the actors and daily bring it out of its equilibrium leaving the negative

feedback to smooth the waves next day. But when we speak, we disseminate memes (cf. Blackmore 1999; Aunger 2000) of different kind and many of them find positive feedbacks disturbing the equilibria. All those who set up and test models of language phenomena observe that even if a model is well corroborated there are always exceptions. These are the points where evolution begins, either spontaneously or by deliberate action. From the methodological point of view this is a very disagreeable situation since we seldom can ascertain all boundary conditions responsible for the deviation. We are, perhaps, in the worst situation among all sciences.

Thus Zipf´s heritage is rather a threshold than a house and linguists will never be able to cross it alone. It is a program whose parts have been just touched in the 20th century but its most stubborn part is still locked up. Without Zipf we would not even have a research program.

Zipf made, according to his critics and reviewers, many mistakes. Fortunately, he never cared for his critics, he continuated his research unmoved and constantly conveyed linguistics out of its comfortable equilibrium. "Far from equilibrium" his research embodied in a strange attractor attracting not only its neighbouring disciplines, but turning up in very distant ones and leaving its traces everywhere. His importance for linguistics can without hesitation be compared with that of Newton for physics.

References

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

Altmann, G. (2003). The diversification process. In: Altmann, G., Köhler, R., Piotrowski, R.G. (eds.), *Handbook of Quantitative Linguistics*. Berlin: de Gruyter (to appear).

Altmann, G., Schwibbe, M. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim: Olms.

Aunger, R. (ed.) (2000). *Darwinizing culture. The status of memetics as a science*. New York: Oxford University Press.

Baayen, R.H. (2001). *Word frequency distributions*. Dordrecht: Kluwer.

Bak, P. (1999). *How nature works: the science of self-organized criticality.* New York: Copernicus-Springer.

Balasubrahmanyan, V.K., Naranan, S. (2000). Information theory and algorithmic complexity: applications to linguistic discourses and DNA sequences as complex systems. Part II. Complexity of DNA sequences, analogy with linguistic discourses. *J. of Quantitative Linguistics 7, 153-183.*

Best, K.-H. (ed.) (2001). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.

Blackmore, S. (1999). *The meme machine.* Oxford: OUP.

Brainerd, B. (ed.) (1983). *Historical Linguistics*. Bochum: Brockmeyer.

Bybee, J., Hopper, P. (eds.) (2001). *Frequency and the emergence of linguistics structure*. Alsterdam: Benjamins

Chitashvili, R.J., Baayen, R.H. (1993). Word frequency distributions. In: Hřebíček, L., Altmann, G. (eds.), *Quantitative text analysis: 54-135*. Trier: WVT.

Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. Linguistics 17, 517-556.

Gell-Mann, M. (1994). *The quark and the jaguar.* New York: Freeman.

Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells.* Trier: WVT.

Hřebíček, L. (1995). *Text levels. Language constructs, constituents and the Menzerath-Altmann law.* Trier: WVT.

Hřebíček, L. (1997). *Lectures on text theory.* Prague: Oriental Institute**.**

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Köhler, R. (1987a). Systems theoretical linguistics. *Theoretical Linguistics 14, 241-257.*

Köhler, R. (1987b). Sprachliche Selbstregulation als Mechanismus des Sprachwandels. In: Boretzky, N., Enninger, W., Stolz, Th. (eds.), *Beiträge zum 3. Essener Kolloquium über Sprachwandel und seine bestimmenden Faktoren: 185-200*. Bochum: Brockmeyer.

Köhler, R. (1988). Die Selbstregulation der Lexik. In: Bluhme, H. (ed.), *Beiträge zur quantitativenb Linguistik: 156-166.* Tübingen: Narr.

Köhler, R. (1989). Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsmechanismus. In: Altmann, G., Schwibbe, M. (eds.), *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen: 108-112.* Hildesheim: Olms.

Köhler, R. (1990a). Linguistische Analyseebenen, Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation. *Glottometrika 11, 1-18.*

Köhler, R. (1990b). Zur Charakteristik dynamischer Modelle. *Glottometrika 11, 39-46.*

Köhler, R. (1990c). Elemente der synergetischen Linguistik. *Glottometrika 12, 179-187***.**

Köhler, R. (1990d). Synergetik und sprachliche Dynamik. In: Koch, W.A. (Hrsg.), *Natürlichkeit der Sprache und der Kultur: 96-112*. Bochum: Brockmeyer.

Köhler, R. (1991). Diversification of coding methods in grammar. In: Rothe, U. (Hrsg.), *Diversification processes in language: grammar: 47-55*. Hagen: Rottmann.

Köhler, R. (1992a). Self-regulation and self-organization in language. In: Saukkonen, P. (ed.), *What is language synergetics?: 14-16.* Oulu: Acta Universi-tatis Ouluensis, Series B, Humaniora 16.

Köhler, R. (1992b). Methoden und Modelle in der quantitativen Linguistik. In: Faulbaum, F. (ed.), *SoftStat ´91. Advances in Statistical Software 3: 489-495.* Stuttgart: Fischer.

Köhler, R. (ed.) (1995). *Bibliography of Quantitative Linguistics*. Amsterdam: Bemjamins.

Köhler, R. (1999). Syntactic structures: Properties and interrelations. *J. of Quan^-titative Linguistics 6, 46-57.*

Krott, A. (1994). *Ein funktionalanalytisches Modell der Wortbildung.* Trier: Magisterarbeit.

Miller, G.A. (1968) Introduction. In: Zipf, G.K., *The psycho-biology of language. An introduction to dynamic philology.* Cambridge, Mass.: The M.I.T. Press.

Naranan, S., Balasubrahmanyan, V.K. (1998). Models for powert law erelations in linguistics and information science. *J. of Quantitative Linguistics 5, 35-61***.**

Naranan, S., Balasubrahmanyan, V.K. (2000). Information theory and algorithmic complexity: applications to linguistic discourses and DNA sequences as complex systems. Part I. Efficiency of the genetic code of DNA. *J. of Quantitative Linguistics 7, 129-151.*

Orlov, Ju.K., Boroda, M.G., Nadarejšvili, I.Š. (1982). *Sprache, Text, Kunst. Quantitative Analysen*. Bochum: Brockmeyer.

Pawlowski, A. (2001): *Metody kwantytatywne v sekwencyjnej analize tekstu*. Warszawa: Uniwersytet Warszawski.

Prün, C. (1995). *Die linguistischen Hypothesen von G.K. Zipf aus system-theoretischer Sicht.* Trier: Diss.

Rapoport, A. (1982). Zipf´s law re-visited. In: Guiter, H., Arapov, M.V. (eds.), *Studies on Zipf´s law: 1-28.* Bochum: Brockmeyer.

Rothe, U. (ed.) (1991). *Diversification processes in language: grammar*. Hagen: Rottmann.

Schroeder, M. (1991). *Fractals, chaos, power laws.* New York: Freeman.

Zipf, G.K. (1968[2]). *Thy psycho-biology of language. An introduction to dynamic philology*. Cambridge, Mass.: The M.I.T. Press.

Glottometrics 3, 2002, 27-38 *To honor G.K. Zipf*

Zipf's Law and Text

Luděk Hřebíček[1]

Abstract. Two questions are to be solved: (1) Which is the sense of the immense corpuses? Are they necessary for the deeper investigation of the rank-size relation? The idea of *text* (instead of the predominating idea of *sentence* and the lower units) as a new linguistic paradigm cannot exterminate the importance of the Zipf law. (2) Certainly, rank can be immediately derived from size. However, seeking the linguistic reasons of ranks could be set against the possible objection of tautology in connection with the Zipf law.

Keywords: Text, context, levels, construct, corpus, rank, Zipf-Mandelbrot law, Zipf-Alekseev distribution, Menzerath-Altmann law

Usual approaches to the analysis of the rank-size relation, once discovered by G. K. Zipf and called Zipf's law, mainly concern two linguistic objects directly: lexical units and a corpus. As a mere exception, phones/letters are taken into consideration, see especially B.B. Mandelbrot (1953) and several other papers quoted and elucidated by A. Rapoport (1982). Zipf's law can be understood as a general description of lexical structures proper to natural languages. Zipf himself offered very ingenious explanation of the law in the more generalized *Principle of Least Effort*, see Zipf (1949), which can be conceived as a law characterizing human behavior. This means that the mental operations with meanings, in a linguistic treatment, are reduced to the relation between rank numbers and the ordered descending frequencies of lexical units in language corpuses.

From the viewpoint of mathematical statistics no reasonable objections (except those presented in the quoted paper by Anatol Rapoport) can be expressed against the mentioned reduction to words and corpuses. From the linguistic point of view, however, there is one fact leaping to the eye: the gap between lexical units and corpuses indicates a hardly admissible absence of phenomena relevant for the lexical structures actually occurring in texts. Any *lexical* unit of a natural language is an intuitive abstraction of its meanings deduced from its usage in texts. The real semantic properties of words can be more adequately observed when they are tested as units of individual texts. Two questions can be asked in the connection with these facts:

(1) Is the mentioned approach basing on corpuses able to express the real lexical structure of a language?
(2) When individual text structures are taken into account, are the theoretical assertions anyhow connected with Zipf's law sufficiently strong to hold the shock caused by the shift of testing from corpuses to texts?

[1] Address correspondence to: Luděk Hřebíček, Junácká 17, CZ – 16900 Praha 6, Czech Republic. E-mail: hrebicek@orient.cas.cz

Text structure

The skepticism evoked by such vague terms like "text structure" is quite comprehensible, but since the times of Zipf's discoveries linguistics has made certain progress by clarifying some of the puzzles of language structures. Speaking about text structure we have in mind especially the mutual relations of the language units. Let us stress two of their properties:

(a) Any observed continuous sound sequence of a language functions as the carrier of discontinuous sequence of code symbols on different language levels.

(b) The units of different levels can be described as sets distinguished by self-similarity, see Hřebíček (1995; 1997: 48ff.; 2000; to appear).

The final aim of this reasoning is the image of a text as a scaled language unit, in which language inventory (or code) is projected into the complicated lattice obtained through the scaling (or segmentation) of the continuous language process into units and their levels. This also concerns lexical units and their meanings connected with their position in the lattice. The way of generation of this image can be characterized in a short form as follows:

Both mentioned properties are consequences of the principle called the Menzerath-Altmann law, see, for example, Altmann (1980), Köhler (1982, 1986), Altmann & Schwibbe et al. (1989). This principle is based on the notions of language *construct* and their *constituent(s)* introduced into consideration by Gabriel Altmann as a generalization of the basic idea formulated by Paul Menzerath. Let us recall its verbal formulation: *The longer a language construct the shorter its components (constituents).*

This assertion has been tested empirically in different languages at different language levels. Thereupon it was observed that the same principle is also valid for the supra-sentence structures. Naturally, the supra-sentence structures cannot be assumed anywhere else but in individual texts. This means that text is a linguistic unit like word or sentence. The enlargement of the Menzerath-Altmann law to texts is discussed, for example, in Hřebíček (1989, 1992, 1995, 1996, 1997, 2000).

The supra-sentence unit can be characterized as *larger context* (*LC*).[2] Each LC is a language construct having as its constituent(s) those sentences of a text in which a given lexical unit occurs. *Narrower context* (*NC*) is then a sentence containing such a lexical unit. In several languages of different typological properties it was corroborated that

> *the larger an LC in the number of sentences, the shorter the mean lengths of the sentences forming its NC.*

This means that lexical units in a text bear meanings, the properties of which can be characterized quantitatively with the help of the Menzerath-Altmann law or with the help of the related theoretical propositions. This theory was tested on texts in several languages of different typological and stylistic properties.

One property of texts deserves to be mentioned in connection with LC and their testing: Their contextual qualities can be observed if the analyzed text is of optimal size. The reason is found in the character of the instrument used for observation, i.e. statistics. The extremely short texts (consisting, for example, of one word or several words, one sentence or a small number of sentences) are not suitable for such observations. For understandable reasons, however, exceedingly large texts also cannot be used for the same purpose; in excessively large texts the contextual details of lexical units are wiped off and cannot be reflected by statistical data. If a text increases over some reasonable limits, the distribution of lexical units substantially changes its course. The same holds when texts are accumulated into immense

[2] Note of the editor: it is usually called "hreb".

corpuses. Thus texts suitable for analyses of their lexical structures are, for example, longer articles from a newspaper, short stories or chapters of a novel (but not the whole novel). In other words, very large texts are not homogeneous; they are semantically segmented and comprised of "sub-texts". This property is reflected by the observed data. The concept of *data homogeneity* was coined and applied to this problem by Altmann (1992); in his work one can find its theoretical description and explanation.

Corpus vs. text

It is an established linguistic tradition to create the mostly abstract lexical meanings of the word units in languages by augmentation of the sources, from which the lexicographic data are achieved. One cannot imagine this approach directed to some general lexical meaning otherwise than as a constriction of the semantic field of the individual lexical units originally presented in texts. As a rule, language facts are thus skewed. In the above quoted paper, Anatol Rapoport discusses an *ad hoc* assumption that the frequency of occurrence of a meaning associated with a lexical unit is proportional to the number of meanings associated with that lexical unit. Different meanings, however, as well as the nuances of lexical meanings are observable only in texts where the new shades of meanings of the old words are constructed. One cannot reliably distinguish between new meanings and shades of old meanings. Both those categories are respected by the position of lexical units in the lattices of texts. These properties disappear when a corpus increases in a boundless manner.

In the sufficiently rich bibliography of the studies concerning Zipf's law there are items connecting analyses with texts instead of corpuses. The study by Callan (1997), for example, presents statistic data forming the so-called TIME collection. It consists of 423 short Time magazine articles comprising word occurrences of the total text (= corpus) length 245,412 words. Consequently, such a study is directed to the characterization of a certain stylistic type. The linguistic and semantic status of the units, however, is not completely clear; in this paper instead of lexical units "terms" are watched. The reason is evident: the highest frequency in such English corpuses is ascribed to definite article *the*. One can hardly expect this unit to have a lexical meaning. The English articles *the* and *a* actually are morphological constituents of word forms. This correction must be stressed when we decided to take language levels into account. Similar standpoints are not principally against corpuses in general, because they can be useful for practical purposes; however, they cannot be used as the basis of theoretical argumentations.

In the mentioned collection, Zipf's relation between rank (r) and size (s), which should be approximately rs = constant, is documented on 50 "terms" with the highest sizes (= frequencies). Among these 50 items one can find pairs with equaling frequencies, but – surprisingly – differing ranks. This approach is acceptable when we operate on an approximative level of thinking.

Ranks in a text

The present paper submits a discussion of the variables significant for Zipf's law in connection with the respective data obtained from texts. The text from which observed data were obtained is a Turkish short story (text **3**). Its length equals 721 word forms consisting of 353 lexical units (= vocabulary of the text). In order to analyze the actual rank-size relation in a clear way, the lexical units were arranged into a sequence according to the descending relative frequencies p; rank number r was ascribed to each value of p. Figure 1 presents the

curve of the values p with related r. It is evident that the trend of p with rising r is not constant. Figure 2 testifies that the rank-size product represented by the values rp is far from being constant as it should be according to the prescription following the rough and not sufficiently general formulation of Zipf's law. The problem of saving the law when texts become the source of statistical data does not seem to be simple.

The rank-size relation was later expressed in a more general form rs^{γ} = constant by Zipf himself. Rapoport writes (o.c., p. 2) that this general formulation

"will hold for any definition of size (if it holds for one of them), as long as all definitions are related by a power relation, $s_i = as_j^k$, with a, k constants, $a > 0$, where s_i and s_j are any two measures of size, and only then."

If in this power relation k is a negative quantity observed in a text, it has no other meaning than the respective parameter of the Menzerath-Altmann law. Thus we can see that the two laws are connected with each other. Here we bring some empirical evidence of this fact. Besides, we seek for the possibility to observe Zipf's law in individual texts.

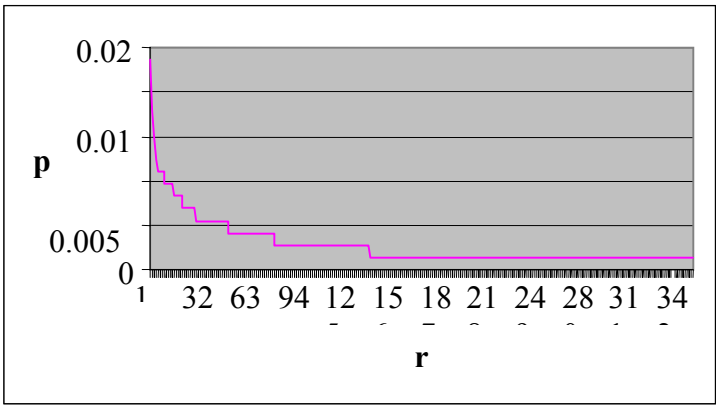


Figure 1. The relative word frequency p with increasing rank number r.

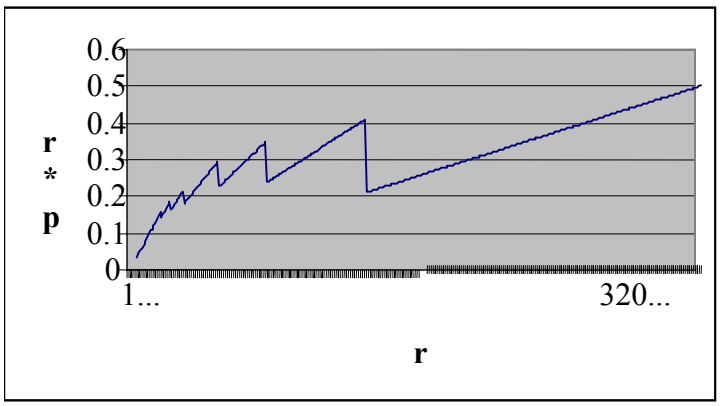


Figure 2. The sequence of products rp.

A more adequate generalization of Zipf's idea was found in the formulation by Mandelbrot who identified the effort in producing words with the number of letters occurring in a respective word. The assumption used for derivation of a new general expression is the average amount of information per signal in any message. He wrote the formula:

$$P_r = A(r+m)^{-B} \quad , \tag{1}$$

where *A, m, B* are constants, *r* is rank and P_r a probability connected with a rank.

When testing the presence of the relations discovered by Zipf and Mandelbrot, and newly explicated and clarified by Rapoport, we used the Zipf-Mandelbrot distribution as it is exposed in Altmann-FITTER (1994: 92) in the following tests, also see Wimmer & Altmann (1999: 666):

$$P_x = \frac{(b+x)^{-a}}{F(n)} \quad , \qquad x = 1,2,...,n \, , \quad a, b > 0, \quad n \in \mathrm{N}, \quad F(n) = \sum_{i=1}^{n} (b+i)^{-a} \, . \tag{2}$$

Let us test the total range of ranks from $r = 1$ to 353 and the corresponding word frequencies presented in Fig. 1. The absolute values were analyzed with the help of Altmann-FITTER. For the Zipf-Mandelbrot distribution, see formula (2), a very good correspondence between the observed and computed values was obtained: the chi^2-test of goodness-of-fit results in $X^2 = 15.93$ with the degrees of freedom $df = 304$ and probability $P \approx 1.00$.

Thus we see that the relations on which Zipf's law is grounded can be traced in the frequencies of lexical units in the text analyzed.

What is the matter with the parallel Menzerath-Altmann's law which is the fundamental text law? In the Zipf-Mandelbrot distribution which is derived from the Principle of Least Effort (i.e. Zipf's law) the variable *frequency of lexical units* is explained as the function of *r*, see formula (1). In the Menzerath-Altmann law the same variable (this time, however, labeled *x*) is the function of the supra-sentence language construct *LC*. Now we must turn our attention to the so-called Zipf-Alekseev (or Zipf-Dolinskiy) theoretical distribution and explicate how this distribution is connected with the Menzerath-Altmann law. The following short explication refers to Hřebíček (1996; 1997: 41-44).

The starting assumption of the Zipf-Alekseev distribution can be formulated in the following way:

$$f_x \approx f_1 \frac{1}{x} \, .$$

It means that in a text the number of lexical units f_x having frequency *x* is directly proportional to the number of lexical units with frequency $x = 1$ and indirectly proportional to *x*. When this relation is supplemented by a coefficient *c*, a completely unrealistic equation (in a logarithmic transformation) is obtained:

$$\ln f_1 - \ln f_x = (\ln c)(\ln x) \quad ,$$

because from it follows that

$$\left(\frac{f_1}{f_x} \right)^{1/\ln x} = c \, .$$

However, we already know that the Menzerath-Altmann law $y = Ax^b$ (where *y* represents a constituent, *x* construct, and *A* and *b* are parameters) is an important text law corroborated by testing. When this law is substituted for *c*, we obtain the equation

$$\left(\frac{f_1}{f_x}\right)^{1/\ln x} = Ax^b \quad .$$

With $e^a = A$ we have the Zipf-Alekseev equation:

$$f_x = f_1 x^{-(a+b\ln x)} \quad . \tag{3}$$

This result means that the Zipf-Alekseev distribution is a special formulation of the Menzerath-Altmann law.

Our task is now to interpret the mutual relation of the independent variables of the two discussed laws, r and x. Let us try to formulate identity

$$r \equiv x$$

and understand it as a statistical hypothesis to be tested in observed texts. For this testing we use the transformed and modified Zipf-Alekseev distribution, see Altmann-FITTER (1994: 79) and Wimmer & Altmann (1999: 665):

$$P_x = \begin{cases} \alpha, & x = 1 \qquad 0 < \alpha < 1 \\ \dfrac{(1-\alpha)x^{-(a+b\ln x)}}{T}, & x = 2,3,4,...,n \qquad T = \sum_{j=2}^{n} j^{-(a+b\ln j)} \end{cases} \quad . \tag{4}$$

Parameters $\hat{\alpha} = f_1/N$ and $\hat{n} = x_{max}$ are fixed (do not change by iteration); a, b are real numbers and n is a natural number.

Let us compare the above mentioned results of the fitting with the help of the Zipf-Mandelbrot theoretical distribution with that of right truncated modified Zipf-Alekseev: there is no significant difference between the observed and expected values of frequency: $X^2 = 15.04$ with $df = 300$ and $P \approx 1.00$. It can be deduced that there is no reason to reject the hypothesis of the identity $r \equiv x$. This hypothesis is also corroborated by the analyses of texts presented below. Further we write f_r instead of f_x.

What sort of identity, however, is it? These variables belong to two functions which remain different. For this reason we specify their identity as "functional identity" or as "functional similarity". This facilitates the understanding of r with the help of x.

Now we want to make a digression to the preceding topic of "text vs. corpus" in obtaining an evidence of Zipf's ideas. The same analyses as above were accomplished with the sequences of f_r from the corpuses published by Callan (1997). We have in mind the above mentioned TIME collection and also the WSJ87 collection; the most frequent 50 "terms" obtained from each of those data were tested by Zipf-Mandelbrot distribution and right truncated modified Zipf-Alekseev distribution. The results are as follows:

Zipf-Mandelbrot: $X^2 = 2685.86$ with $df = 46$ and P = 0.0000 (C = 0.0293).
Zipf-Alekseev: $X^2 = 2374.90$ with $df = 45$ and P = 0.0000 (C = 0.0258).

Consequently, no good fitting was obtained. It is not clear whether we have a convincing argument against too large corpuses, but the preference of texts with their contextual abilities of lexical units should be further investigated. Here we must repeat Altmann's opinion that

the inhomogeneity of the data obtained from corpora (and also from too large texts) is the reason for refusing such sets of data for theoretical analyses.

The plain frequencies

The Principle of Least Effort is revealed through variables r and s – or better: r and f_r. Variable r derives its values from the set of f_r. One should, however, exactly know what frequency actually means in linguistics and what the possible linguistic interpretations of size are in connection with languages and texts.

An acceptable interpretation of word frequencies within the frame of the Menzerath-Altmann text model is based on contextual abilities of lexical units. This property has been quite extens-ively documented in the above quoted works. There are two contextual "sizes" corresponding to the categories of *language constructs* and their *constituents*, featured on the supra-sentence level by *LC* and *NC*, respectively. This can be demonstrated by the following example: When the frequency of a lexical unit in a text becomes by one higher (and when the occurrences higher than one of any lexical unit in one and the same sentence is neglected), the number of *NC*s increases by one. This means that the size of the respective *LC* (= construct) rises up. Such dynamism evidently states the meaning of the supposed lexical unit in a more precise way on the contextual basis. This is the real sense of the law. Generally, each text thus specifies the actual meanings of its lexical units.

Therefore we can say that variable f_r is a way of description of meanings and their structures in texts. Variable r is then deduced from f_r and therefore can be understood as another expression of size proper to each lexical unit in a text. It makes no difference whether f_r is treated in the relative or absolute values; when both expressions – absolute or relative – are ordered to the sequence r, the curves obtained are practically the same.

If a subset of lexical units has some equal values of frequencies, the units cannot be characterized by different values of r and their actual sequence thus remains unknown. In order to eliminate the repetition of equal frequencies ascribed to different lexical units we tried to examine the distributions of word frequencies in an abbreviated appearance.

As an object of analysis, let us take the above analyzed text again. It was already proved that the descending sequence of the values of f_r, in which the values of r are ascribed to the equal frequencies at random, corresponds to two theoretical distributions: Zipf-Mandelbrot and Zipf-Alekseev. The way of arrangement of the equal frequencies, however, represents too great an approximation and one can hesitate whether such an approximation is acceptable. The computer program and, generally, the computing of any theoretical distribution, requires the sequence of natural numbers 1, 2, 3, ..., therefore the ranks cannot be given in that way which is usual in statistics (the rank number of the equaling values of a variable equates with the average obtained from their ranks in case they would be different).

At first we tried to apply the following transformational steps to the observed data of frequencies:

(1) A curve has been put through the observed data; it was proved that the curve was the closest approximation to the observed distribution.

(2) Unequal theoretical values obtained thus were substituted for the observed values.

The theoretical values were defined as $f_{r,exp} = Ar^{-b}$ with $A = 17$ and b estimated as 0,4617. We expected to obtain a distribution corresponding to the two mentioned theoretical distributions, i.e. Zipf–Mandelbrot and Zipf–Alekseev. Against our expectations, the fitter program unmasked our machination and offered the Zeta distribution (= Zipf distribution, or so-called *power law*) with parameter $b = 0.4617$ as the only one theoretical distribution applicable to the transformed data. This result, however, exhibits the deep functional meaning

contained in quantitative data coming from texts and their distributional properties.

Zipf's original idea, as one can understand it, does not concern lexical units, but their frequencies. In the text discussed we observe totally 353 lexical units, but only 12 different frequencies. See Table 1 for the distribution of the absolute frequencies observed in the same Turkish text as above (text **3**).

The distribution of f_r with r taken from this table was analyzed and the following results were obtained:

Zipf–Mandelbrot distribution: Parameters: $a = 3.56$, $b = 13.20$, $n = 12$.
Chi^2-test: $X^2 = 1.55$, $df = 8$, $P = 0.99$
Right truncated modified Zipf–Alekseev: $a = 0.17$, $b = 0.25$, $n = 12$, $\alpha = 0.1977$.
Chi^2-test: $X^2 = 1.90$, $df = 7$, $P = 0.97$.

The probabilities P of the observed chi^2 are high and both theoretical distributions can be taken as the basis for the data observed in the text. We already tried to indicate that the observed quantitative data together with their distributions, as well as the respective theoretical distributions, have a reliable linguistic groundwork. The theoretical distributions can be deduced as a consequence of the two discussed linguistic laws.

Table 1. The observed and expected values of f_r and their ranks r

r	The original ranks	f_r	Expected: Zipf–Mandelbrot	Expected: Zipf–Alekseev
1	1	17	17.71	17.00
2	2	13	13.90	14.20
3	3	11	11.08	11.06
4	4 - 5	9	8.96	8.82
5	6 - 9	8	7.32	7.19
6	10 - 15	7	6.06	5.97
7	16 - 21	6	5.06	5.04
8	22 - 29	5	4.26	4.31
9	30 - 51	4	3.61	3.73
10	52 - 81	3	3.09	3.26
11	82 - 143	2	2.66	2.87
12	144 - 353	1	2.30	2.54

The presented analysis concerns one individual Turkish text. Similar results were obtained from a set of Turkish texts, see Table 2.

Table 2. The results comparable with Table 1 from ten Turkish texts
(v – total of lexical units, n – text length in word forms,
Z–M: Zipf Mandelbrot; *Z–A*: Zipf–Alekseev).

TEXT **1**: $v = 302$, $n = 669$. TEXT **2**: $v = 345$, $n = 624$.

r	f_r	*Z–M*	*Z–A*	f_r	*Z–M*	*Z–A*
1	19	22.79	19.00	17	18.50	17.00
2	18	16.72	18.85	13	13.97	14.93
3	13	12.54	13.61	12	10.74	10.92
4	11	9.57	10.03	9	8.40	8.32

5	8	7.43	7.59	7	6.66	6.55
6	7	5.85	5.88	6	5.35	5.28
7	5	4.66	4.64	5	4.34	4.35
8	4	3.76	3.73	4	3.56	3.64
9	3	3.07	3.05	3	2.95	3.08
10	2	2.52	2.52	2	2.46	2.64
11	1	2.09	2.10	1	2.07	2.29

Text **1**: *Z–M*: $X^2 = 1.95$, $df = 7$, $P = 0.96$. *Z–A*: $X^2 = 1.13$, $df = 6$, $P = 0.98$.
Text **2**: *Z–M*: $X^2 = 1.27$, $df = 7$, $P = 0.99$. *Z–A*: $X^2 = 1.56$, $df = 6$, $P = 0.956$.

TEXT **3.** See Table 1.

TEXT **4**: $v = 368$, $n = 756$. TEXT **5**: $v = 389$, $n = 873$.

r	f_r	***Z–M***	***Z–A***	f_r	***Z–M***	***Z–A***
1	18	18.46	18.00	31	25.89	31.00
2	14	14.17	14.15	13	18.87	15.28
3	10	11.13	11.25	12	14.43	12.57
4	9	8.92	9.00	11	11.43	10.46
5	8	7.26	7.31	10	9.31	8.85
6	7	6.00	6.03	9	7.74	7.59
7	6	5.02	5.04	8	6.55	6.59
8	5	4.24	4.27	7	5.62	5.78
9	4	3.62	3.65	6	4.88	5.12
10	3	3.12	3.15	5	4.29	4.56
11	2	2.70	2.75	4	3.80	4.09
12	1	2.36	2.41	3	3.39	3.70
13	–	–	–	2	3.05	3.36
14	–	–	–	1	2.75	3.06

Text **4**: *Z–M*: $X^2 = 1.71$, $df = 8$, $P = 0.99$. *Z–A*: $X^2 = 1.74$, $df = 7$, $P = 0.97$.
Text **5**: *Z–M*: $X^2 = 6.09$, $df = 10$, $P = 0.81$. *Z–A*: $X^2 = 3.63$, $df = 9$, $P = 0.93$.

TEXT **6**: $v = 405$, $n = 810$. TEXT **7**: $v = 460$, $n = 880$.

r	f_r	***Z–M***	***Z–A***	f_r	***Z–M***	***Z–A***
1	21	23.00	21.00	20	19.39	20.00
2	17	17.95	18.50	13	15.62	15.36
3	14	14.34	14.80	12	12.70	12.10
4	12	11.70	11.98	11	10.41	9.79
5	11	9.70	9.87	9	8.60	8.10
6	10	8.15	8.27	8	7.16	6.84
7	8	6.94	7.02	7	6.00	5.86
8	7	5.97	6.03	6	5.05	5.08
9	6	5.18	5.23	5	4.28	4.45
10	5	4.53	4.58	4	3.65	3.94
11	4	4.00	4.04	3	3.13	3.51
12	3	3.55	3.59	2	2.69	3.15
13	2	3.17	3.21	1	2.33	2.84
14	1	2.84	2.88	–	–	–

Text **6**: *Z–M*: $X^2 = 3.06$, $df = 10$, $P = 0.98$. *Z–A*: $X^2 = 2.88$, $df = 9$, $P = 0.97$.
Text **7**: *Z–M*: $X^2 = 2.09$, $df = 9$, $P = 0.99$. *Z–A*: $X^2 = 2.95$, $df = 8$, $P = 0.94$.

TEXT **8**: $v = 623$, $n = 1572$. TEXT **9**: $v = 627$, $n = 1296$.

r	f_r	*Z–M*	*Z–A*	f_r	*Z–M*	*Z–A*
1	30	33.07	30.00	21	19.91	21.00
2	28	28.53	30.59	18	17.55	21.25
3	25	24.69	25.56	17	15.48	16.82
4	23	21.43	21.44	14	13.68	13.76
5	18	18.66	18.19	13	12.10	11.54
6	15	16.29	15.63	11	10.72	9.87
7	14	14.26	13.57	10	9.51	8.56
8	13	12.52	11.90	9	8.44	7.52
9	12	11.01	10.52	8	7.51	6.58
10	11	9.72	9.37	7	6.68	5.97
11	10	8.59	8.40	6	5.95	5.39
12	9	7.61	7.57	5	5.31	4.88
13	8	6.56	6.86	4	4.74	4.46
14	7	6.01	6.24	3	4.24	4.08
15	6	5.36	5.71	2	3.79	3.76
16	5	4.79	5.23	1	3.40	3.47
17	4	4.29	4.82	–	–	–
18	3	3.85	4.45	–	–	–
19	2	3.45	4.12	–	–	–
20	1	3.11	3.82	–	–	–

Text **8**: *Z–M*: $X^2 = 4.03$, $df = 16$, $P = 0.999$. *Z–A*: $X^2 = 5.65$, $df = 15$, $P = 0.99$.
Text **9**: *Z–M*: $X^2 = 3.44$, $df = 12$, $P = 0.99$. *Z–A*: $X^2 = 4.78$, $df = 11$, $P = 0.94$.

TEXT **10**: $v = 775$, $n = 2033$.

r	f_r	*Z–M*	*Z–A*	r	f_r	*Z–M*	*Z–A*
1	39	46.67	39.00	**16**	14	11.32	11.49
2	37	39.51	40.22	**17**	13	10.72	10.85
3	30	34.13	35.12	**18**	12	10.17	10.27
4	29	29.95	30.89	**19**	11	9.67	9.74
5	27	26.61	27.46	**20**	10	9.22	9.25
6	26	23.90	24.65	**21**	9	8.80	8.81
7	23	21.65	22.32	**22**	8	8.42	8.39
8	22	19.75	20.35	**23**	7	8.06	8.01
9	21	18.14	18.67	**24**	6	7.73	7.66
10	20	16.75	17.22	**25**	5	7.43	7.33
11	19	15.55	15.95	**26**	4	7.15	7.03
12	18	14.49	14.84	**27**	3	6.88	6.74
13	17	13.56	13.86	**28**	2	6.63	6.48
14	16	12.73	12.98	**29**	1	6.40	6.23
15	15	11.99	12.20	–	–	–	–

Text **10**: *Z–M*: $X^2 = 22.06$, $df = 25$, $P = 0.63$. *Z–A*: $X^2 = 19.01$, $df = 24$, $P = 0.75$.

From these results it is evident that the properties ascribed to the plain word frequencies f_r and their ranks r can be traced in all texts analyzed. If not a proof, then an indication of the presence of this quality not only in Turkish texts can be found in Table 3, where the results of an analogous analysis of a Czech text are presented.

Table 3. The results obtained from a Czech text.

TEXT **11**: $v = 446$, $n = 1003$.

r	f_r	$Z–M$	$Z–A$	r	f_r	$Z–M$	$Z–A$
1	30	30.39	30.00	**9**	8	7.08	7.02
2	21	24.50	26.25	**10**	7	6.10	6.35
3	18	19.97	18.80	**11**	6	5.29	5.79
4	17	16.45	14.69	**12**	5	4.61	5.32
5	16	13.68	12.07	**13**	4	4.03	4.92
6	14	11.48	10.24	**14**	3	3.55	4.57
7	10	9.70	8.89	**15**	2	3.13	4.27
8	9	8.26	7.85	**16**	1	2.78	4.00

Text **11**: *Z–M*: $X^2 = 3.76$, $df = 12$, $P = 0.99$. *Z–A*: $X^2 = 8.82$, $df = 11$, $P = 0.64$.

Conclusions

Any linguistic corroboration of the validity of Zipf´s law does not need immense corpuses or texts, because they are always statistically inhomogeneous. Their essential inhomogeneity wipes off the traces of the characteristics expressed by this law. They, however, can be found in individual texts that offer homogeneous data. In texts Zipf's law (especially in the formulation by B.B. Mandelbrot) appears to be linguistically more convincing than in large corpuses or texts. This result has been obtained when the distributions of the frequencies of lexical units in texts were tested.

If the Zipf–Alekseev theoretical distribution is applied to the same observed frequencies, one need not reject the hypothesis of no significant difference between the observed and theoretical values. The reason for this is the way, in which this theoretical distribution is deduced with the help of the Menzerath–Altmann law. The results of the parallel testing with the help of the two theoretical distributions can be understood as a statistical corroboration of the functional similarity (or identity) between the rank and size of the larger context *LC* of an analyzed text. Thus we obtained information on what variable "rank" really means.

The present paper brings arguments for that understanding of Zipf´s law which does not concern lexical items, but directly the distribution of *the values* of their frequencies and their ranks. The observed ranks are full of linguistic sense.

References

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika 2, 1-10.*

Altmann, G., Schwibbe, M. H. et al. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim-Zürich-New York: Olms.

Altmann, G. (1992). Das Problem der Datenhomogenität. *Glottometrika 13, 287-298.*

Altmann-FITTER (1994). *Begleitbuch zu* Altmann-FITTER. *Iterative Anpassung diskreter Wahrscheinlichkeitsverteilungen*. Lüdenscheid: RAM–Verlag.

Callan, J. (1997). Characteristics of text. [Internet address: http://hobart.cs.umass.edu/~allan/cs646-f97/char_of_text.html].

Hřebíček, L. (1989). The Menzerath-Altmann law on the semantic level. *Glottometrika 11*, *47-56.*

Hřebíček, L. (1992). *Text in communication: supra-sentence structures*. Bochum: Brockmeyer.

Hřebíček, L. (1995). *Text levels. Language conctructs, constituents and Menzerath–Altmann's law.* Wissenschaftlicher Verlag Trier.
Hřebíček, L. (1996).Word associations and text. *Glottometrika 15, 96-101.*
Hřebíček, L. (1997). *Lectures on text theory.* Prague: Oriental Institute.
Hřebíček, L. (2000). *Variation in sequences.* Prague: Oriental Institute.
Hřebíček, L. (to appear), Text laws. In: Köhler, R. et al. (eds), *Handbook of quantitative linguistics*. Berlin: de Gruyter.
Köhler, R. (1982). Das Menzerathsche Gesetz auf Satzebene. *Glottometrika 4*, *103-113.*
Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.
Mandelbrot, B. B. (1953). An information theory of the statistical structure of language. In: Jackson, W. (ed.), *Communication theory*: *486-502*. London: Butterworth.
Rapoport, A. (1982). Zipf's law re-visited. In: Guiter, H., Arapov, M. V. (eds), *Studies on Zipf's law: 1-28.* Bochum: Brockmeyer.
Wimmer, G., Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions.* Essen, Stamm.
Zipf, G.K. (1935/1968). *The psycho-biology of language. An introduction to dynamic philology.* Cambridge, Mass.: The M.I.T. Press, 2n ed.
Zipf, G. K. (1949), *Human behavior and the principle of least effort*. Cambridge (Mass.): Addison-Wesley Press.

Texts analyzed

Turkish texts:
1 – M. Kaplan, Dil ve Kültür. *Türk Dili*, 1993/11, Sayı 500, Ağustos 1993, 108-110.
2 – Y. Kadri Karaosmanoğlu, Muradiye. In: Muzaffer Reşit (ed.), *Türk Nesir Antolojisi.* İstanbul, Varlık, 1969, 67-71.
3 – Necati Cumalı, Rossana. In: N. C., *Revizyonist*. İstanbul, Tekin, 1979, 207-211.
4 – Necati Cumalı, İnsanlar Kardeştir. In: N. C., *Revizyonist*. İstanbul, Tekin, 42-46.
5 – Tarık Dursun K., Bıçak İşi. In: T. D. K., *Güzel Avrat Otu*. İstanbul, Düşün, 1960, 64-68.
6 – Mahmut H. Şakiroğlu, Alessandro Manzoni, I promessi sposi. [A review.] *Erdem*, Cilt 3, Sayı 7, Ocak 1987, 275-277.
7 – Tarık Dursun K., Musikili Oda. In: T. D. K., Güz*el Avrat Otu.* İstanbul, Düşün, 1960, 10-14.
8 – Necati Cumalı, *Yağmurlar ve Topraklar.* İstanbul, Cem, 1973, 326-340.
9 – Demir Özlü, *Bir Yaz Mevsimi Romansı*. İstanbul, Ada, 1990, 7-13.
10 – Necati Cumalı, *Yağmurlar ve Topraklar*. İstanbul, Cem, 1973, 5-15.

A Czech text:
11 – Ivan Kraus, Klášterní ulice. In: I. K., *Číslo do nebe*. Marsyas, Praha 1993, 14-19.

Glottometrics 3, 2002, 39-60 *To honor G.K. Zipf*

Zipf's notion of "economy" on the text level
(A case study in Czech)

Ludmila Uhlířová[1]

Abstract. The role of nominal phrases with the determiner 'this' in text structure is examined and interpreted in terms of Zipf´s opposition of the Force of Unification and the Force of Diversification. Their balance is demonstrated on Czech data.

Keywords: Zipf, economy, rank-frequency, unification, diversification, Czech, text, anaphor

1. George Kingsley Zipf was a natural scientist. He saw speech as a natural phenomenon and analyzed it in a broad framework of human ecology, i. e. of that branch of science which deals with human habits, modes of life and relations to the environment. In his famous book (1949) he claimed that, like any human action, "the entire phenomenon of speech is presumably subject to the Principle of Least Effort" (1949:55). His contribution to the science of language is so important and long-standing that today, more than half century after his book was published and one hundred years after he was born, some people believe he was a linguist.

Zipf´s well-known statistical laws are based on the idea that two "economies", or "opposing forces" are in constant operation in the stream of speech: the Force of Unification and the Force of Diversification. Any speech (any spoken or written text) is a result of a "balance" between them. The first law manifesting that balance, sounds: If all running words in a text are ranked in the decreasing order of their frequency of occurrence, then the product of rank r and frequency f is constant: $r.f = const$. Zipf illustrated the law on data from James Joyce's novel Ulysses.

Since 1949, the law has been permanently commented on, discussed, criticized, corrected, and empirically tested so many times that we cannot afford the space for referring even to most important studies (see Köhler, 1995, for bibliography). Instead, let us make a very brief historical digression and mention an interesting fact which maybe is not generally known. Roughly at the same time when Zipf began to write his works in America, functionally oriented quantitative linguistics was born in Prague. In Czechoslovakia, Zipf's studies were soon reviewed by the Czech anglicist and quantitative linguist Bohumil Trnka (1950). (There are reasons to assume as probable that both scholars knew each other personally.) Trnka, who pointed out the independence of and the differences between the starting points and aims of the Prague School's and Zipf's quantitative researches, was certainly not wrong when he highly praised Zipf's work in his English review on the one hand, expressing his conviction that 'it will not fail to influence the linguistic thought of today' (1950:5), nor was he wrong when he, on the other, stressed that Zipf's laws were only partially applicable and demanded them to be revised. He accurately described Zipf's contribution such that he as a statistician showed the advantages of a statistical method compared with qualitative analysis in the sense that statistical analysis "is able to afford to neglect the narrow limits of one language and

[1] Address correspondence to: Ludmila Uhlířová, Ústav pro jazyk český, ČAV, Letenská 4, CZ-11851 Praha. E-mail: uhlirova@ujc.cas.cz

concentrate on linguistic problems of a general character". Trnka believed in the existence of general quantitative laws which govern the structure of all languages, and saw the attempts to formulate them as a major task of future quantitative linguistics. He himself, however, avoided their formulation, considering it premature at that time.

Trnka was proved right by the subsequent development of quantitative linguistics. During several last decades, substantial progress has been achieved in the knowledge of universal quantitative laws. This holds true not only of a relatively large number of distributional laws including the above-mentioned Zipf´s law, but also other laws of crucial importance which were found, e.g. the Menzerath-Altmann law (see Altmann, 1980, 1988, Hřebíček, 1995, 1997, 2000), the Sherman-Altmann laws (Best, 2001a), diversification laws, Martin's law, Frumkina's law etc. (see Best, 2001b, for the survey of the most important ones). Furthermore, it is much more comfortable to apply and test the laws now when linguists have very large computerized corpora of data at their disposal and a special software is available (Altmann-Fitter, 1997; Ziegler, Altmann, 2002).

The Zipf's law mentioned above is a means to a very general end. If we succeed in proving that a language phenomenon abides by it, we give further empirical evidence in favour of Zipf's underlying idea that the principle of "economy", the principle of two competing Forces, really IS of a universal nature. Below, we shall try to prove it on a concrete phenomenon at the text level. We shall deal with the nominal group *tento* N ('this N') in Czech: *tento člověk* 'this man', *tato kniha* ´this book´, *toto rozhodnutí* 'this decision' etc.

2. *Tento* is a demonstrative pronoun; it is flexible and its forms vary according to gender, case and number. Semantically, *tento* expresses a "near proximity" (to borrow a brief characteristics from Halliday, Hasan, 1976:38), or - in almost the same Czech wording (Mluvnice češtiny II, 1986:92) - "vztahy prostorové vzdálenosti od mluvčího, které jsou mluvčímu buď bližší, nebo vzdálenější", i.e. 'relations of a spatial distance from the speaker, which can be either nearer, or farther'. *Tento* stands in opposition to the demonstrative *ten* 'that', which expresses a "far proximity" (Halliday, Hasan, 1976:38). Grammatically, it is a left modifier to its head noun N.[2]

The modifier *tento* is a linguistic means with a text-forming role. It provides its nominal group with a "phoric" function, and it does so irrespective of whether there are other premodifiers or postmodifiers to the N or not. Below, properties of the distribution of *tento* N in texts will be examined.

2.1. Firstly, the following question should be asked: What is the rank-frequency distribution of *tento*N in texts?

Source data. For present-day Czech, we have the Czech National Corpus (http://ucnk.ff.cuni.cz) at our disposal, which consists of several very large corpora. The core corpus, called Syn2000, is composed of 120 million word forms together with a detailed part-of-speech and morphological tagging. Syn consists of texts from three broad communicative fields, mainly representing language for special purposes (science, technology, and humanities), language of media and language of narratives. For the purpose of the present study we chose:

(a) a smaller corpus Synek, which is available on CD as a separate corpus, consisting of 20 million words, in which the proportionality of texts from different communicative fields sources is the same as in Syn (the texts in Synek were chosen from Syn);

[2] In Czech, *tento* may also be used independently of N, e.g. *výsledek je tento* lit. 'the result is this'; such cases, which are not frequent, are outside the scope of the present study.

(b) a very big corpus Mladá fronta, consisting of 152.8 million words, available on the web and representing texts from the most popular Czech daily Mladá fronta.

The computer was given an instruction [lemma = 'tento'][tag = 'N.*'] to choose all occurrences of *tento*N, with the following results:

corpus	length	Number of occurrences
Mladá fronta	152.8 mil.	308 839
Synek	20 mil.	28 476

As the number of occurrences is too high to handle, we took a sample of one thousand occurrences from the beginning, then a sample of one thousand of occurrences from the middle, then from the end, and finally one thousand random occurrences from each corpus. In this way we obtained eight samples consisting of the same number of occurrences of *tento*N. The data are shown in Tables 1-8, respectively. The first column of each table comprises lexemes N together with the English translations,. the second column ranks. The third column indicates the empirical frequencies of *tento*N and the last the calculated values showing the fitting of the Zipf's model in the well-known Zipf-Mandelbrot modification,

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

where $x = 1,2,3,...,n$ and b and a are parameters, $F(n)$ is the normalizing constant (see Wimmer, Altmann 1999, for a more detailed definition and mathematical comment).

Tables 1-8: Rank-frequency distribution of *tento*N in eight samples.[3]

Table 1
Rank-frequency distribution in the sample of 1000 *tento*N
in the Synek corpus (taken from the beginning)

lexeme	***r***	***f***	***NP***
chvíle while	1	65	41,91
místo place	2	23	32,67
svět world	3	19	26,86
muž man	4	15	22,87
dům house	5	15	19,94
okolnost circumstance	6	14	17,70
člověk man	7	14	15,93
případ case	8	14	14,49
způsob manner	9	13	13,30
slovo word	10	13	12,30
doba time	11	12	11,44
země country	12	12	10,70
věc thing	13	11	10,05
okamžik moment	14	10	9,48
skutečnost reality	15	10	8,98
žena woman	16	9	8,52

[3] To save space, we do not present the tables in full here; we cut them at $f > 5$.

noc night	17	9	8,12
práce work	18	9	7,75
směr direction	19	8	7,41
situace situation	20	8	7,11
kraj region	21	7	6,82
pán gentleman	22	7	6,57
stránka aspect	23	6	6,33
myšlenka idea	24	6	6,10
město city	25	6	5,90
kniha book	26	6	5,71
příležitost occasion	27	6	5,53
vlastnost property	28	6	5,36
druh kind	29	6	5,20
část part	30	6	5,05
loď ship	31	6	4,91
a = 0,9165, b = 2,2028, X^2 = 24,52 , df = 27, P = 0,60			

Table 2
Rank-frequency distribution in the sample of 1000 *tento*N in the Synek corpus (taken from the middle)

lexeme	***r***	***f***	***NP***
případ case	1	29	23,63
země country	2	19	19,86
zákon law	3	15	17,44
týden week	4	15	15,73
souvislost connection	5	13	14,44
otázka question	6	12	13,42
den day	7	12	12,58
strana side	8	11	11,89
oblast region	9	11	11,29
situace situation	10	11	10,78
stát state	11	10	10,33
směr direction	12	10	9,93
rok year	13	10	9,57
člověk man	14	9	9,25
firma company	15	9	8,96
problém problem	16	9	8,70
slovo word	17	9	8,45
projekt project	18	8	8,23
částka amount	19	8	8,02
chvíle while	20	8	7,83
doba time	21	8	7,65
možnost possessivity	22	8	7,48
rozhodnutí decision	23	8	7,32
důvod reason	24	7	7,18
město city	25	7	7,04
krok step	26	7	6,90
způsob manner	27	7	6,78
století century	28	7	6,66
záležitost matter	29	7	6,55
typ type	30	7	6,44
smysl sense	31	6	6,34

informace information	32	6	6,24
změna change	33	6	6,15
muž man	34	6	6,06
film film	35	6	5,97
organizace organization	36	6	5,89
a = 0,5100, b = 1,4592, X^2 = 2,39, df = 196, P ~ 1,00			

Table 3
Rank-frequency distribution in the sample of 1000 *tento*N in the Synek corpus (taken from the end)

lexeme	***r***	***f***	***NP***
vůz vehicle	1	54	42,21
motor motor	2	27	34,42
rok year	3	26	29,03
model model	4	23	25,09
soutěž competition	5	20	22,08
případ case	6	20	19,71
systém system	7	18	17,79
kategorie category	8	16	16,21
sezóna season	9	15	14,88
typ type	10	14	13,76
verze version	11	12	12,79
víkend weekend	12	11	11,94
značka mark	13	11	11,20
den day	14	11	10,55
disciplína discipline	15	10	9,96
směr direction	16	10	9,44
týden week	17	10	8,97
země country	18	9	8,54
auto car	19	8	8,15
pořad agenda	20	8	7,80
trend trend	21	8	7,47
skutečnost reality	22	8	7,17
sport sport	23	7	6,89
změna change	24	7	6,64
chvíle while	25	6	6,40
závod factory	26	6	6,18
jezdec driver	27	6	5,97
období period	28	6	5,77
a = 1,0269, b = 3,5456, X^2 = 6,11, df = 24, P = 0,9999			

Table 4
Rank-frequency distribution in the sample of 1000 *tento*N in the Synek corpus (a random sample)

lexeme	***r***	***f***	***NP***
případ case	1	33	23,25
den day	2	16	20,22
země country	3	16	17,90
oblast region	4	15	16,07
otázka question	5	12	14,57
způsob manner	6	12	13,34

týden week	7	12	12,30
chvíle while	8	11	11,41
věc thing	9	10	10,65
souvislost connection	10	10	9,98
člověk man	11	9	9,39
ohled respect	12	9	8,87
smysl sense	13	8	8,40
strana side	14	8	7,98
doba time	15	8	7,60
muž man	16	7	7,26
krok step	17	7	6,95
skupina group	18	7	6,66
vůz vehicle	19	7	6,39
cena price	20	6	6,15
částka amount	21	6	5,92
akce activity	22	6	5,71
zákon law	23	6	5,52
směr direction	24	6	5,34
období period	25	6	5,17
a = 0,9713, b = 5,4813, X^2 = 6,32, df = 21, P = 0,9991			

Table 5
Rank-frequency distribution in the sample of 1000 *tento*N in the Mladá fronta corpus (taken from the beginning)

lexeme	***r***	***f***	***NP***
souvislost connection	1	26	30,94
rok year	2	22	26,33
týden week	3	22	23,01
případ case	4	19	20,50
země country	5	17	18,53
den day	6	16	16,93
krok step	7	16	15,61
otázka question	8	15	14,50
oblast region	9	14	13,55
chvíle while	10	12	12,73
strana side	11	12	12,01
doba time	12	11	11,38
společnost society	13	11	10,82
problém problem	14	10	10,31
skutečnost reality	15	10	9,85
údaj datum	16	10	9,44
účel purpose	17	10	9,07
typ type	18	9	8,72
ohled respect	19	9	8,40
způsob manner	20	9	8,11
fond fond	21	9	7,84
firma company	22	9	7,59
situace situation	23	9	7,35
opatření measure	24	8	7,14
směr direction	25	8	6,93
problematika problem	26	7	6,74
smysl sense	27	7	6,56

člověk man	28	7	6,39
bod point	29	6	6,23
téma topic	30	6	6,08
částka amount	31	6	5,93
suma sum	32	6	5,80
zákon law	33	6	5,67
smlouva treaty	34	6	5,54
organizace organization	35	6	5,42
místo place	36	6	5,31
služba service	37	6	5,20
cesta way	38	6	5,10
druh kind	39	6	5,00
podnik enterprise	40	6	4,91
postup procedure	41	6	4,82
a = 0,8173, b = 3,5830, X^2 = 21,89, df = 347,.			P ~ 1,00

Table 6
Rank-frequency distribution in the sample of 1000 *tento*N
in the Mladá fronta corpus (taken from the middle)

lexeme	***r***	***f***	***NP***
případ case	1	25	27,05
rok year	2	23	23,14
země country	3	23	20,32
týden week	4	19	18,19
den day	5	15	16,51
chvíle while	6	15	15,15
krok step	7	14	14,02
člověk man	8	12	13,07
oblast region	9	11	12,25
sezóna season	10	11	11,55
věc thing	11	11	10,93
způsob manner	12	9	10,38
otázka question	13	9	9,89
strana party	14	9	9,45
změna change	15	8	9,06
problém problem	16	8	8,70
doba time	17	8	8,37
měsíc month	18	8	8,07
instituce institution	19	8	7,79
možnost possibility	20	8	7,53
víkend weekend	21	7	7,30
výsledek result	22	7	7,08
souvislost connection	23	7	6,87
firma company	24	7	6,68
místo palce	25	7	6,50
druh kind	26	7	6,33
století century	27	6	6,17
práce work	28	6	6,02
část part	29	6	5,87
vláda government	30	6	5,74
vůz vehicle	31	6	5,61
záležitost matter	32	6	5,49

údaj datum	33	6	5,37
trend trend	34	6	5,26
a = 0,7666, b = 3,4211, df = 374, X^2 = 20,26, P ~ 1,00			

Table 7
Rank-frequency distribution in the sample of 1000 *tento*N in the Mladá fronta corpus (taken from the end)

lexeme	***r***	***f***	***NP***
rok year	1	45	42,81
den day	2	38	31,67
týden week	3	19	25,56
století century	4	17	21,64
případ case	5	16	18,89
místo place	6	14	16,83
doba time	7	13	15,23
chvíle while	8	13	13,95
země country	9	12	12,89
směr direction	10	11	12,00
město city	11	10	11,24
problém problem	12	10	10,59
období period	13	10	10,02
sezóna season	14	10	9,51
krok step	15	10	9,06
rubrika column	16	10	8,66
člověk man	17	10	8,30
částka amount	18	9	7,97
otázka question	19	9	7,67
souvislost connection	20	9	7,39
typ type	21	8	7,14
oblast region	22	8	6,91
věc thing	23	8	6,69
možnost possibility	24	8	6,49
druh kind	25	8	6,30
podnik enterprise	26	8	6,12
událost event	27	7	5,96
stav state	28	7	5,80
účel purpose	29	7	5,66
změna change	30	6	5,52
soutěž competition	31	6	5,39
akce activity	32	6	5,27
informace information	33	6	5,15
situace situation	34	6	5,04
a = 0,75, b = 1,0251, X^2 = 27,34, df = 362, P ~ 1,00			

Table 8
Rank-frequency distribution in the sample of 1000 *tento*N in the Mladá fronta corpus (a random sample)

lexeme	***r***	***f***	***NP***
týden week	1	23	22,32
případ case	2	20	19,75
rok year	3	19	17,77

země country	4	17	16,19
souvislost connection	5	17	14,90
den day	6	15	13,82
problém problem	7	13	12,91
oblast region	8	11	12,13
informace information	9	11	11,44
století century	10	10	10,84
období period	11	10	10,31
chvíle while	12	10	9,83
krok step	13	9	9,41
komentář comment	14	8	9,02
částka amount	15	8	8,66
možnost possibility	16	8	8,34
rozhodnutí decision	17	8	8,05
věc thing	18	7	7,77
místo place	19	7	7,52
skutečnost reality	20	7	7,28
situace situation	21	7	7,07
víkend weekend	22	6	6,86
doba time	23	6	6,67
tvrzení statement	24	6	6,49
způsob manner	25	6	6,32
důvod reason	26	6	6,16
sezóna season	27	6	6,01
akce activity	28	6	5,87
otázka question	29	6	5,74
rubrika column	30	6	5,61
člověk man	31	6	5,49
zpráva news	32	6	5,37
cesta way	33	6	5,26
strana side	34	6	5,16
$a = 0{,}7704$, $b = 4{,}7943$, $X^2 = 23{,}01$, $df = 393$, $P \sim 1{,}00$			

Fitting results, calculated with the help of Altmann-Fitter (1997), are per-suasive: The distribution of *tento*N abides bythe Zipf-Mandelbrot law with very good prob-abilities, as shown in all tables.

2.2. Having obtained good results, let us remember once again Zipf's claim that "hyperbolic distributions of the type observed quite often suggest a governing consideration of economy..." (1949:54), and let us ask the following questions: What kind of economy is involved in this case? What is the main carrier of the Force of Unification, and what is the main carrier of the Force of Diversification with the nominal group *tento*N, and how does a balance between the two Forces manifest itself?

In principle, any demonstrative *tento*, either endophoric, exophoric or cataphoric, is inherently, by its nature, a carrier of the Force of Unification. It belongs to the closed class of pro-nominalizers (Komárek 1978; Mluvnice češtiny II, 1986). It does not refer to reality directly (in contrast to nouns or nominal phrases), but indirectly, "instead of" naming units, and the speaker/hearer should know which noun is substituted in each case to make his communication be unambiguous. Logically, the other constitutive element of *tento*N, the noun, should carry the other Force, the Force of Diversification, so as to preserve the balance of Forces.

Let us compare the following examples for illustration.

(i) *Turisté obdivovali **Karlův most**. Tento skvost gotické architektury s barokními sochami byl postaven ve 14. stol.*

'Turists admired **Charles bridge**. This masterpiece of gothic architecture with baroque statues was built in the 14th century.'

(ii) *Nenechávejte **zavazadla a kabáty** bez dozoru. Za (tyto) ztracené věci neručíme.*

'Do not leave your **luggage and coats** unattended. We take no responsibility for lost things**.'**

(III) *Vezmeme **mouku**. Tuto mouku smícháme s kypřicím práškem a pak do této mouky přidáme vejce...*

'Take some **flour**. Add raising powder to the flour, then add eggs to the flour...'

The nominal group in the first sentence of each example is referentially identical with its respective postcedent in the second sentence. We can say that it is through this identity that the Force of Unification is manifested. However, in (i) and (ii), the second instances differ, both lexically and grammatically, from their antecedents. In (i) the relation between the two is that between a proper name (*Karlův most*) and its metaphor together with right modifiers (*tento skvost gotické architektury s barokními sochami* 'this masterpiece of gothic architecture with baroque statues'). In (ii) the postcedent of the nominal group *zavazadla a kabáty* 'luggage and coats' is a summarizing, abstract nominal group (*tyto) ztracené věci* 'lost things' with a generalizing meaning. Thus, in (i) and (ii) the information carried by the postcedent is diversified from that expressed by its antecedent. There is an evident flow of information forward: a new piece of information, non-recoverable by the listener from the preceding verbal context is added. In contrast to it, in (iii) the same lexeme *mouka* 'flour' is repeated twice. The second instance, *tuto mouku,* is justified by the fact that the rheme of the first sentence *mouku* has become the theme of the second clause. Its communicative status has changed, which, undoubtedly, is also a kind of semantic diversification. In the third clause, the nominal group *do této mouky* is repeated once more, now as a prepositional phrase. No new information is added; the communication status is the same, no Force of Diversification is observed. The repetition is semantically redundant, and, stylistically it looks a little awkward (still, the clause is grammatically correct). To sum up: Whereas in (i), (ii) and in the second clause of (iii) the Forces may be considered well balanced, it is not the case in the last clause of the example (iii), where a Force of Diversification is actually lacking.

Let us have a look at Tables 1-8 now and let us ask how a desirable balance of Forces manifests itself in the samples: What are the most frequent nouns co-occurring with *tento* as its heads?

We can see that the nouns given below are among the most frequent ones. They express:

(a) Orientation in time, current periods of time: *rok* 'year', *týden* 'week', *den* 'day', *chvíle* 'while' (*v této chvíli* 'at this moment'), *doba* 'time', *sezóna* 'season', *měsíc* 'month', *víkend* 'weekend', *století* 'century', *období* 'period', *okamžik* 'moment', *noc* 'night';

(b) Orientation in space, current places of action: *země* 'country', *oblast* 'field', *místo* 'place', *město* 'city, town', *svět* 'world';

(c) Communicative situation as such: *situace* 'situation';

(d) Logical structure, interdependence and interconnection between actions: *souvislost* 'connection', *okolnost* 'circumstance', *stránka* 'aspect', *důvod* 'reason', *účel* 'purpose', *ohled* 'respect', *způsob* 'way', *smysl* 'sense' (*v tomto smyslu* 'in this sense'), *změna* 'change', *směr* 'direction', *krok* 'step';

(e) Content in general, including "metatext": *otázka* 'question', *případ* 'case' (*v tomto případě* in this case'), *problém* 'problem', *skutečnost* 'fact', *téma* 'topic', *věc* 'thing', *možnost* 'possibility', *záležitost* 'matter', *informace* 'information', *slovo* 'word', *myšlenka* 'idea';

(f) Participants: *muž* 'man', *žena* 'woman', *pán* 'gentleman', *stát* 'state'.[4]

[4] Most lexemes in Tables 1-8 are highly polysemic, and their English translations given in this article represent the briefest possible approximations of their common meanings.

Most nouns mentioned above sub (a) to (f) occur across texts and genres. Some of them have the status of "general nouns" as characterized by Halliday-Hasan (1976:275): They are "the superordinate members of major lexical sets" (such as *thing* or *man* in English). In texts they are used as hyperonyms which may co-occur with many (practically: almost any) antecedent noun. Others refer to the context of situation; they name time or space (*tento týden* 'this week', *na tomto světě* 'on this world'). Still others refer not to a noun (or to a nominal group), but a shorter or longer text passage, summarizing its content. Or, finally, some connect two passages of texts; thelatter are near to conjunctives (e. g., *v tomto případě* 'in this case', *v této souvislosti* 'in this connection').

The nouns of broad, abstract, categorial, even most general meanings occur across samples and their overall frequencies are high, as the data show. Consequently, if we were e.g. given the task to find key words in a text (or, in a group of similar texts) then nouns co-occurring with *tento* would be far from beingsuitable for the purpose. In noun phrases *tento*N the respective key terms are hidden under other nouns - be they metaphors, hyperterms, general terms, and various other "cover" terms. Concrete key words, which would directly point to the proper content of individual texts in the samples, only rarely reach beyond the frequency level of $f \geq 5$. Examples of key words with $f \geq 5$ which reveal a text topic are, e.g. *částka* 'amount', *lod*' 'ship', *dům* 'house', *kniha* 'book', *zákon* 'law', *film* 'film', *motor* 'motor', *auto* 'car', *jezdec* 'driver', *cena* 'price' (see Tables 1-8 for more examples).

2.3. If the lexical semantics of the most frequent *tento*N groups does not typically reveal much information about the text content as such, as was demonstrated above, we should ask: What textual functions are typical of *tento*N?

Let us inspect textual functions of *tento*N in more detail. Let us classify five hundred instances in the Synek sample 1 (cf. the data in Table 1) from the beginning of the sample into three basic text-semantic groups: endophoric, exophoric, and cataphoric. The three notions are understood here in the usual sense in which they have been coined and generally accepted in text linguistics, see e.g. Halliday-Hasan (1976), Mluvnice češtiny (1987) etc.

In Table 9 below, all *tento*N were classified according to whether an antecedent was present in the verbal context (a backward, or - much less often - a forward anaphor, i. e. a cataphor), or in the immediate or broader "scene" (in the context of situation or in the context of culture). Then, endophoric *tento*N was classified according to the relevant syntactic, lexical and communicative links to its antecedent. Syntactically, a *tento*N may refer to another nominal phrase, a verbal phrase, a whole sentence or a text passage which precedes it at reasonable distance. Lexically, *tento*N may contain the same noun as the antecedent, or its (contextual) synonym or its metaphor, a hyperonym, or even a very general noun. Communicatively, the communicative status of a postcedent may be the same as that of its antecedent (which is a rare case, however - see example iii above), or it may undergo a change: A rheme (= antecedent) may be thematized in the following text, or, vice versa, a theme may be rhematized again (and as such, it expresses a piece of information which is "irretrievable" by the hearer - in the sense of Firbas, 1997); a thematized *tento*N may express a contrastive, parallel, alternative theme etc.

The classification presented in Table 9 could not be performed by an automatic procedure because there is no tagging of this kind available in the Czech National Corpus. Still we hope that our manually made classification is sufficiently objective and consistent, even though we could not avoid classifying borderline cases and linguists would probably divide on the details. Thus, e. g., in the following example we counted two nominal groups as correferen-tial, although in the first clause the person is speaking about a singular noun *procházkou* 'walk', whereas in the second one the nominal group *na těchto procházkách* 'these walks' has a generalized meaning and the noun is in the plural. Cf.: *Tak, skorem docela mlčky, jsme se ubírali procházkou přes město. Nejkrásnější na těchto procházkách bylo, že* [...] 'So, almost

not talking, we went for a **walk** through the town. Such **walks** (lit.: these walks) were beautiful because…' Thus, we accept a broad, rather pragmatically oriented framework of correference.

Let us notice that cases of *tento* with a non-correferential or modal function occur exceptionally in texts; their frequency of occurrence is not significant. From the systemic point of view, they are quite peripheral (see the last two lines of the table for number of occurrences in our samples as well as the Supplement for examples.).[5]

Table 9
Text-forming functions of *tento*N in the Synek corpus, the first 500 occurrences.
(Cf. the data in Table 1.)

function	***absolute frequency***	***relative frequency***	***number of example*** (see Supplement)
I. endophoric = second instance			
topicalization: rheme→theme	56	11.2	1
contrastive, stressed or unexpected theme	15	3	2
Parallel (enumerative) theme	8	1.6	4
rhematization: irretrievable information	16	3.2	5
repetition of distant (hyper)theme	14	2.8	6
the same/very similar information status	1	0.2	3
alternation of two distinct referents	1	0.2	7
synonym, nearly-synonym metaphor (rarely: hyponym)	31	6.2	8
more general term, hyperonym	50	10	9
general, categorial term	25	5	10
verbal phrase as antecedent	19	3.8	11
complex/compound sentence or text stretch as antecedent	156	31.2	12
II. exophoric			
somebody/something in the nearest environment, visible or evident	21	4.2	13, 14
general context of situation/culture	76	15.2	15, 16
III. **cataphoric= forward reference**	8	1.6	17

[5] In Czech, *tento* is regularly an initial constituent of *tento*N group, with the exception of some quantifiers, which may either precede, or follow the demonstrative, e. g. *tyto tři případy* 'these three cases' or *tři tyto případy, tento celý týden* 'this whole week' or *celý tento týden* and/or with the exception of some particles, e. g. *právě* 'just', *jenom* 'only', *dokonce* 'even'. An adjective before *tento/tato* (*veliká tato sláva* lit. 'big this glory') is strongly marked as archaic; such order has not occurred in our sample at all.

IV. **non-coreferential**	1	0.2	18
V. **modal**	2	0.4	19
Total	500	100	

The statistical data counted in Table 9 clearly confirm once again that *tento*N is a means of expression composed of two constituents with opposite text-forming Forces: *Tento* expresses the Unifying Force, whereas N is primarily a carrier of the Diversifying Force in the sense that it does not just cohere the text, but, at the same time, it adds new pieces/aspects of information and, therefore, it pushes communication forward. In this way, *tento*N is distinguished from other text-forming means, e.g. from personal or possessive pronouns as well as from ellipsis (including the implicit sentence subject - "nevyjádřený podmět") which, primarily, have Unifying roles. Cases of *tento*N, in which the noun simply re-iterates as a second instance without any change in its communicative status, are rare. The reason is that such *tento*N cannot realize the hearer's expectation for a new relevant piece of information at a given moment of communication. It is informationally poor, and it lacks a communicative "economy".

If, then, all subtypes of *tento*N, listed in Table 9, are then arranged in the decreasing order of their frequencies, as it is done in Table 10 below, we can see quite clearly that thones at the top of the frequency list are those, whose contextual links are most complex, semantically most mediated (complex/compound sentence or even a longer text passage as antecedent: 156 occur-rences, i. e. 31.2%), as well as those which are only non-verbally bound, anchored in a broad-er non-verbal context (general context of situation/culture 76 occurrences, i. e. 15.2%), i.e. those with a considerable amount of information import. A change in communicative function is expressed frequently by a postcedent *tento*N, the topicatization being the most frequent case (56 occurrences, i. e. 11.2%).

Let us note that our data concerning the topicalization are in accordance with what František Štícha (2001) showed recently. Štícha investigated the concurrence of *ten/tento/0* in anaphoric nominal groups in cases in which the **same** lexeme N (=antecedent N) is repeated; he came to the conclusion that *tento* is used if N is to be pointed out as the topicalized element. - However, let us remember that it only occurs in 22.2% that the antecedent and postcedent N are lexically identical, as the data in our sample show (see the values in first seven lines of Table 9 taken together; also see examples 1-7 in the Supplement).

Now the following question may be asked: Does the rank-frequency distribution of *tento*N subtypes abide by the Zipf-Mandelbrot model? The data in Table 10, second and third columns, offer a positive answer: the Zipf-Mandelbrot law is a general distributional law of text-forming functions (and subfunctions) of *tento*N in the sense described above. This is actually what we expected, because it is nothing more but just another evidence of the Zipfean economy dealt with above.

Table 10
Rank-frequency distribution of text-forming functions.
Empirical frequencies and Zipf-Mandelbrot model

function	***r***	***f***	***NP***
complex/compound sentence or text stretch as antecedent	1	156	142.55
general context of situation/culture	2	76	89.82
topicalization: rheme→theme	3	56	61.03
more general term, hyperonym	4	50	43.78
synonym, nearly-synonym metaphor (rarely: hyponym)	5	31	32.71

general, categorial term	6	25	25.23
somebody/something in the nearest environment, visible or evident	7	21	19.97
verbal phrase as antecedent	8	19	16.14
rhematization: irretrievable information	9	16	13.27
contrastive, stressed or unexpected theme	10	15	11.08
repetition of distant (hyper)theme	11	14	9.37
parallel (enumerative) theme	12	8	8.01
cataphoric=forward reference	13	8	6.92
modal	14	2	6.02
the same/very similar information status	15	1	5.28
alternation of two distinct referents	16	1	4.67
non-coreferential	17	1	4.15
a = 2,37, b = 3,66, X^2 = 21,18, df = 13, P = 0,07			

In this connection it is worth mentioning that - generally - anaphoric nominal phrases „demonstrative + second instance noun“ are used in Czech less and less, when observed from the developmental point of view. We can refer here to Josef Hrbáček's (1987) detailed comparison of two Czech translations of a Russian short story by Turgenev, a translation made in 1892, and another one made almost a century later, in 1977. Hrbáček showed quite persuasively that pronominalization as a means of cohesion became significantly weakened in the latter translation. Clearly, his conclusion may be considered a support of our statistical results: means of text cohesion, which are more economic, are preferred more and more. A functional difference between *tento*N and an independent personal or possessive pronoun in text structure has also become even more pronounced during the last decades in Czech.

2.4. In addition to *tento*N we examined nominal groups of the type *tento*AdjectiveN, in which there is **one** adjective inserted between the demonstrative and the head noun. Now, the computer was given an instruction to find all occurrences of the type [lemma = "ten-to"][tag="A.*"][tag="N.*"]. This three-member nominal group is considerably less frequent: there are 3696 occurrences in the Synek corpus. [6]

Let us discuss the lexical semantics of the adjective. What do the data show? Without giving a detailed empirical account, we can say that irrespective of a more general or more concrete lexical meaning of noun, evaluative adjectives highly prevail (about 90% of all occurrences in the corpus) over differentiating ones. Examples *starý chrám* 'old cathedral', *divoký cirkus* 'wild circus', *prolhaný dům* 'dishonest house', *temná hrozba* 'dark menace', *oblíbený list* 'favourite newspaper', *roztrpčující období* 'embittering period', *prapodivná osobnost* 'strange personality', *pověstný prapor* 'notoriously known battalion', *spodničková záležitost* 'underskirt affair' are much more frequent than *vodní oblast* 'water region', *odborný název* 'technical term', *kapesní mikroskop* 'pocket microscope' and the like. An obvious reason is that evaluative adjectives are more capable of adding new (= evaluative, pragmatic) meaning to the noun, and thus to modifiy/shift the meaning of the nominal group (of the antecedent) on a very general scale "good - bad". On the other hand, differentiating adjectives, especially in a two-word naming unit, do not possess such a capability at all or such a capability is con-textually limited to them.

2.5. Now let us approach *tento*N from a slightly different, but complementary point of view. In the terminology of corpus linguistics, we can say that *tento*N is a kind of a very special **collocation pattern** which consists of a functional lexeme + full-meaning lexeme. Using František Čermák's classification of collocations (2001), we can - in our opinion -

[6] Naturally, cases where the distance between *tento* and the head N is > 1, do exist in Czech, but they are still less frequent, and an automatic search in Synek would require revision.

classify *tento*N as a kind of non-fixed, non-idiomatic, but regular grammatic-semantic combination of lexemes, which belong to our common, everyday means of expression.

2.5.1. Let us return to the class of hyperonyms and general nouns once again. As we already know from Tables 1-8, such nouns occupy the lowest ranks in the frequency lists of *tento*N. If we now accept a corpus linguistical framework, we may say that we expect such words will have high *t*-scores. The *t*-score is a statistical characteristic defined in corpus linguistical studies quite a long time ago (see e.g. Church, 1993, for a detailed discussion; Čermák, 2000), and widely recommended for use (e.g. Český národní korpus. Úvod:61). The formula (Český národní korpus: Úvod...:61) is as follows:

$$t = \frac{\left(f(x,y) - \frac{f(x).f(y)}{N} \right)}{\sqrt{f(x,y)}}, \qquad (2)$$

where *f(x)* and *f(y)* are frequencies of words *x* and *y*, respectively, and *N* is the length of the sample. The formula is based on a statistical method of testing hypotheses with the help of a *t*-test. It says that the higher the *t*-score, the higher the probability of two words to co-occur with each other. If the *t*-score is significant, they make a collocation together. The results of the calculation of *t*-scores are given in Table 11 for the Mladá fronta corpus, and in Table 13 for the Synek corpus. The calculation has been performed according to the instructions available on the web site of the Czech National Corpus. The data in the fourth column show the absolute frequency of a collocation of *tento* with a given N. The relative frequency in the third column shows the relative frequency of N (= the percentage calculated from the total frequency of N in the corpus) collocated with *tento*. For example: the col-location *tento týden* 'this week' occurs 8214 times in the Synek corpus. Its relative frequency in the context of *tento* is 9.43%, which means that almost each tenth occurrence of the lexeme *týden* is collocated with *tento*. Taking into account this value of *t*-score, which is high, we can say that the co-occurrence is significant and we may speak of a collocation. As expected, the highest *t*-scores are found with nouns such as *případ* 'case', *oblast* 'region', *souvislost* 'connection', *chvíle* 'while', *týden* 'week', *směr* 'direction', *způsob* 'manner', *otázka* 'question' etc. in Synek as a whole (i. e. in 20 millions of words), and, rather similarly, with nouns such as *týden* 'week', *případ* , 'case', *oblast* 'region', *krok* 'step', *souvislost* 'connection', *způsob* 'manner', *chvíle* 'while', *částka* 'amount' etc. in Mladá fronta as a whole (i.e. in 152,8 millions of words). *t*-scores values are correlated with absolute fre-quencies of the collocated nouns - see the last columns. Now, if we compare the data in our eight samples (see tables 1-8) with *t*-scores in the Synek corpus as a whole and with Mladá fronta as a whole (Tables 11 ans 13), we can see that the highest *t*-scores are characteristic of lexemes with rather general lexical semantics; the lists of lexemes partly intersect and, in fact, both approaches confirm each other.

In contrast to it, the *mi*-score (mutual information score) goes hand in hand with relative frequencies of the collocated words. The formula (Český národní korpus, Úvod:60)

$$mi(x,y) = \log_2 \frac{N.f(x,y)}{f(x).f(y)}, \qquad (3)$$

leads to the results in Table 12, and Table 14 respectively, where the data are arranged in the decreasing order of *mi*-score, and, consequently, according to the decreasing order of the relative frequency of collocations. The highest values are achieved with nouns such as *rubrika*

'column', *komentář* 'comment', *resort* 'field', *eventualita* 'possibility', *disproporce* 'disproportion', *nepoměr* 'imbalance', *bohulibý* 'pleasing to God', *končina* 'corners', *nešvar* 'bad habit', *logik* 'logician', *mezidobí* 'interim', and others with relatively low absolute frequencies in the corpus. E. g. the word *eventualita* 'possibility' was used 41times in the collocation *tato eventualita* (i.e.: 41 is the absolute frequency of this lexeme in the context of *tento*), which makes 28.87% of all occurrences of the lexeme *eventualita* in the corpus (28.87% is the relative frequency of the lexeme *eventualita* in the context of *tento*). In other words: approximately each fourth occurrence of the lexeme *eventualita* occurred in the collocation *tato eventualita* 'this possibility', and as such, its *mi*-score is high.

If we calculated *mi*-scores for **all** collocations of *tento*N in the corpus, such words as *liliput* 'Lilliputian', *dvojpolárnost* 'bipolarity', *zahušťování* 'thickening' and other would be included in the list of collocations. Each of these lexemes occurred just once in the corpus, and its occurrence was collocated with *tento.* Consequently, these words would have the highest *mi*-scores and their relative frequencies would equal 100%. Similarly, e.g. the word *tasemnice* 'tapeworm' occurred twice in the corpus, and one of the occurrences was collocated with *tento.* The absolute frequency of this collocation equals one, which makes 50% (= relative frequency in the context of *tento*) of its two occurrences in the corpus. - Such a collocation list would be enormously long. Taking into account the size of the corpus as well as the special, text-forming function of the studied collocation it was reasonable to cut the list, as it was done in the tables.

On the whole, the data show that many collocations with a **high** *mi*-score have **low** *t*-scores, and vice versa. Among those with high *mi*-scores are nouns with concrete lexical meanings, bound to individual texts, as well as some nouns which are stylistically marked in various ways.

The values of the scores are not independent of the corpus size. There is a discussion in literature (Church - Mercer, 1993) about assumptions which determine the "cutoff": Which *t*-scores are still significant markers of collocations (which are above/below the cutoff threshold)? We do not underestimate the importance of the discussion. However, here we profit from the advantage that we are dealing with a collocation of a special kind (functional word + meaning word) in very big corpora.

To conclude this paragraph. Both Zipf's law and the *t*-score and *mi*-score are based on the parameters/properties of the frequency distribution, and as such they both express, in principal, the same facts about using language. It is only the viewpoint (or probably better: the form of presentation) that differs, the latter being more common in corpus linguistics. One may choose either of them, or use them as two similar, complementary modeling techniques. In our case, both clearly support the above-presented argumentation concerning the text functions of noun in the collocation pattern *tento*N.

Table 11
T-scores and *mi*-scores of *tento*N collocations in the Mladá fronta corpus in the decreasing order of *t*-scores (the list is cut at f < 66).

lemma	***mi-score***	***t-score***	***relative frequency***	***absolute frequency***
týden week	5.309	88.35	9.432	8214
případ case	5.04	84.94	7.827	7675
oblast region	5.396	59.94	10.01	3770
krok step	5.933	57.98	14.53	3474
souvislost connection	6.455	54.51	20.87	3040
způsob manner	5.118	52.66	8.261	2940
chvíle while	5.211	52.14	8.811	2871

částka amount	5.373	46.44	9.859	2265
směr direction	5.67	45.19	12.11	2125
komentář commentary	7.055	44.48	31.63	2009
skutečnost reality	5.318	40.46	9.489	1722
víkend weekend	4.783	40.19	6.547	1739
druh kind	5.333	39.43	9.588	1635
rubrika column	8.034	38.38	62.35	1484
téma topic	5.33	37.72	9.568	1496
typ type	4.836	36.54	6.792	1434
opatření measure	5.238	34.92	8.976	1287
účel purpose	6.121	32.82	16.55	1109
suma sum	5.907	31.74	14.27	1042
záležitost matter	4.816	31.04	6.7	1036
instituce institution	5.144	30.97	8.411	1016
varianta variant	5.226	30.31	8.904	970
trend trend	5.932	30.19	14.52	942
ohled regard	5.509	28.93	10.83	875
záměr intention	4.98	28.71	7.508	879
fakt fact	4.851	26.53	6.866	755
tvrzení statement	5.237	24.88	8.97	653
problematika problem	5.935	23.53	14.55	572
datum date	5.352	23.06	9.713	559
lokalita locality	5.695	22.19	12.32	512
metoda method	4.83	20.63	6.765	457
fáze phase	5.183	19.45	8.641	400
peněžní monetary	4.99	18.98	7.559	384
disciplína discipline	4.911	18.75	7.156	376
odvětví branch	5.771	17.67	12.99	324
jev phenomenon	5.52	17.66	10.91	326
transakce transaction	5.523	17.55	10.93	322
ustanovení regulation	5.517	16.51	10.89	285
žánr genre	5.18	16.5	8.625	288
causa case	5.206	16.48	8.777	287
ukazatel indicator	5.607	16.21	11.6	274
choroba disease	4.797	15.18	6.612	248
kauza case	5.344	13.34	9.659	187
nařízení order	4.811	12.35	6.678	164
světec saint	5.086	10.5	8.08	117
teze thesis	5.307	10.18	9.413	109
domněnka assumption	5.244	9.881	9.011	103
formulace formulation	4.946	8.973	7.332	86
nařčení accusation	5.162	8.962	8.517	85
úkon act	4.799	8.73	6.624	82
konstatování statement	5.436	8.516	10.3	76
uskupení grouping	4.808	8.295	6.661	74
nešvar bad habit	5.972	7.995	14.93	66

Table 12
T-scores and *mi*-scores of *tento*N collocations in the Mladá fronta corpus in the decreasing order of *mi*-scores (lthe list is cut at *mi*-score < 5.21).

lexeme	***mi-score***	***t-score***	***relative frequency***	***absolute frequency***
rubrika column	8.034	38.38	62.35	1484
komentář comment	7.055	44.48	31.63	2009
resort field	7.046	6.58	31.43	44
eventualita possibility	6.924	6.35	28.87	41
souvislost connection	6.455	54.51	20.87	3040
účel purpose	6.121	32.82	16.55	1109

disproporce disproportion	6.117	4.18	16.51	18
nepoměr imbalance	6.111	5.91	16.44	36
bohulibý pleasing to God	6.074	4.93	16.03	25
končina corners	5.996	7.50	15.18	58
nešvar bad habit	5.972	7.99	14.93	66
problematika problem	5.935	23.53	14.55	572
krok step	5.933	57.98	14.53	3474
trend trend	5.932	30.19	14.52	942
suma sum	5.907	31.74	14.27	1042
logik logician	5.833	3.93	13.56	16
odvětví branch	5.771	17.67	12.99	324
komodita commodity	5.707	6.28	12.42	41
lokalita site	5.695	22.19	12.32	512
směr direction	5.67	45.19	12.11	2125
zlozvyk bad habit	5.614	5.27	11.65	29
ukazatel indicator	5.607	16.21	11.60	274
mezidobí interim	5.557	3.79	11.19	15
hypotéza hypothesis	5.548	7.71	11.13	62
transakce transaction	5.523	17.55	10.93	322
jev phenomenon	5.52	17.66	10.91	326
ustanovení regulation	5.517	16.51	10.89	285
ohled regard	5.509	28.93	10.83	875
segment segment	5.446	4.03	10.37	17
konstatování statement	5.436	8.52	10.30	76
oblast region	5.396	59.94	10.01	3770
cifra numeral	5.394	4.78	10.00	24
částka amount	5.373	46.44	9.86	2265
datum date	5.352	23.06	9.71	559
kauza case	5.344	13.34	9.66	187
druh kind	5.333	39.43	9.59	1635
téma topic	5.33	37.72	9.57	1496
skutečnost reality	5.318	40.46	9.49	1722
branže line of business	5.318	6.75	9.49	48
týden week	5.309	88.35	9.43	8214
teze thesis	5.307	10.18	9.41	109
taktik tactician	5.248	7.67	9.04	62
domněnka presupposition	5.244	9.88	9.01	103
opatření measure	5.238	34.92	8.98	1287
tvrzení statement	5.237	24.88	8.97	653
varianta variant	5.226	30.31	8.90	970
chvíle while	5.211	52.14	8.81	2871

Table 13
T-scores and *mi*-scores of *tento*N collocations in the Synek subcorpus in the decreasing order of *t*-scores (the list is cut at f < 20).

lexeme	***mi-score***	***t-score***	***relative frequency***	***absolute frequency***
případ case	4.84	23.87	8.349	612
oblast region	5.084	18.67	9.888	370
souvislost connection	6.175	18.66	21.07	358
chvíle while	4.661	17.34	7.377	326
týden week	4.716	17.13	7.664	317
směr direction	5.669	16.87	14.84	296
způsob manner	4.661	16.8	7.375	306
otázka question	4.328	15.73	5.855	274
krok step	5.058	14.16	9.713	213
skutečnost reality	4.75	13.72	7.847	203
druh kind	4.977	12.55	9.18	168

typ type	4.719	12.47	7.678	168
ohled regard	5.613	12.27	14.27	157
téma topic	5.000	11.71	9.329	146
období period	4.518	11.56	6.679	146
proces process	4.549	10.75	6.826	126
částka amount	4.906	10.54	8.744	119
příležitost occasion	4.537	10.31	6.772	116
opatření measure	4.712	9.522	7.644	98
účel purpose	5.250	9.289	11.10	91
problematika problem	5.628	8.598	14.42	77
pojem notion	4.652	7.976	7.333	69
metoda method	4.364	7.960	6.003	70
víkend weekend	4.776	7.945	7.991	68
obor field	4.495	7.881	6.576	68
tvrzení statement	5.200	7.782	10.72	64
trend trend	5.431	7.691	12.58	62
jev phenomenon	5.211	7.661	10.80	62
okolnost circumstance	4.547	7.476	6.816	61
rubrika column	5.859	7.288	16.92	55
ustanovení regulation	5.122	7.071	10.15	53
cvik exercise	6.064	6.753	19.50	47
varianta variant	4.557	6.703	6.863	49
vlastnost property	4.418	6.465	6.233	46
pojetí conception	4.81	6.397	8.178	44
příkaz order	4.679	6.374	7.470	44
komentář comment	5.06	6.361	9.729	43
suma sum	5.202	6.229	10.73	41
datum date	4.859	6.030	8.460	39
disciplína discipline	4.959	5.966	9.069	38
poloha position	4.377	5.711	6.061	36
moment moment	4.435	5.643	6.306	35
odvětví branch	5.512	5.619	13.31	33
argument argument	4.386	5.215	6.098	30
kauza case	4.729	5.182	7.733	29
tendence tendency	4.309	4.934	5.782	27
aspekt aspect	4.502	4.483	6.607	22
poznatek piece of knowledge	4.418	4.471	6.232	22
kontrolka warning light	6.477	4.422	25.97	20

Table 14
T-scores and *mi*-scores of *tento*N collocations in the Synek subcorpus in the decreasing order of *mi*-scores (low values are not shown).

lexeme	***mi-score***	***t-score***	***relative frequency***	***absolute frequency***
bohulibý pleasing to God	6.837	2.217	33.33	5
kontrolka	6.477	4.422	25.97	20
cifra	6.422	2.21	25	5
odpočet	6.199	2.416	21.43	6
souvislost connection	6.175	18.66	21.07	358
cvik	6.064	6.753	19.5	47
rubrika column	5.859	7.288	16.92	55
minilab minilab	5.766	3.104	15.87	10
kolonka column	5.721	3.102	15.38	10
směr direction	5.669	16.87	14.84	296
problematika problem	5.628	8.598	14.42	77
ohled regard	5.613	12.27	14.27	157
končina corners	5.557	2.59	13.73	7
úkaz	5.515	2.396	13.33	6

odvětví branch	5.512	5.619	13.31	33
zorný	5.452	2.394	12.77	6
trend trend	5.431	7.691	12.58	62
stať article	5.422	3.783	12.50	15
konstatování statement	5.334	3.649	11.76	14
komodita	5.334	2.758	11.76	8
dilema	5.317	3.083	11.63	10
etnikum	5.252	2.385	11.11	6
účel purpose	5.250	9.289	11.10	91
jev phenomenon	5.211	7.661	10.80	62
suma sum	5.202	6.229	10.73	41
tvrzení statement	5.200	7.782	10.72	64
lokalizace	5.189	2.175	10.64	5
údobí	5.159	2.173	10.42	5
kombinéza	5.136	2.748	10.26	8
ustanovení	5.122	7.071	10.15	53
oblast	5.084	18.67	9.89	370
lokalita	5.084	4.118	9.89	18
komentář	5.060	6.361	9.73	43
krok step	5.058	14.16	9.71	213
koncept concept	5.049	3.629	9.66	14
dialektika	5.045	3.496	9.63	13
téma topic	5.000	11.71	9.33	146
druh kind	4.977	12.55	9.18	168
branže	4.962	2.904	9.09	9
disciplína	4.959	5.966	9.07	38
síla strength	4.936	3.059	8.93	10
částka amount	4.906	10.54	8.74	119
procedura procedure	4.905	2.900	8.74	9
rozlišení	4.883	2.733	8.60	8
datum date	4.859	6.030	8.46	39
případ case	4.840	23.87	8.35	612
pojetí	4.810	6.397	8.18	44
příhoda	4.800	3.476	8.13	13

3. To sum up. The text-forming effect of *tento*N can be empirically measured in terms of Zipf's law (in the Zipf-Mandelbrot modification) and interpreted in terms of Zipf's economy and balance of the two Forces, Force of Unification and Force of Diversification: *Tento*N is used whenever the speaker intends to present (a) an element unifying the text structure (by means of *tento*), and, at the same time (b) an element diversifying it. The noun, although being co-referential and/or somehow anchored in the context, still is an element introducing/carrying a piece of new information, shifting the communication forward, and letting communication flow. *Tento*N manifests a clear tendency not to serve as a means of pure re-iteration of an old piece of information, i.e. not to carry only a unifying Force. Such a re-iteration would be inefficient from the communicative point of view, and therefore it is mostly avoided in texts. The Czech language system has other means at its disposal to express a unifying Force, e.g. personal or possessive pronouns, or ellipsis.

In full accordance with what was just said, *tento*N can be viewed as a very special kind of collocation pattern, consisting of a functional lexeme + full-meaning lexeme, which realizes a huge number of collocations in texts. Some of them have very general meanings, and as such, they are used across texts and genres, with high *t*-scores in the corpora investigated.

References

Altmann, G. (1980). Prolegomena to Menzerath's Law. In: Grotjahn, R. (Hrsg.), *Glotto-metrika 2, 1-10.* Brockmeyer: Bochum.

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

Altmann-Fitter (1997). RAM-Verlag: Lüdenscheid.

Best, K.-H. (2001a). Wie viele Wörter enthalten Sätze im deutschen? Ein Beitrag zu dem Sherman-Altmann-Gesetzen. In: Best, K.-H. (Hrsg.), *Häufigkeitsverteilungen in Texten: 167-201*. Peust & Gutschmidt Verlag: Göttingen.

Best, K.-H. (2001b). *Quantitative Linguistik. (*= Göttinger Linguistische Abhandlungen 3) Peust & Gutschmidt Verlag: Göttingen.

Čermák, F. (2000). Combination, Collocation and Multi-Word units. In: *Proceedings of The Ninth Euralex International Congress EURALEX 2000.* Heid, U. - Evert. S. - Lehmann, E. (Eds.), Stuttgart, 489-495.

Čermák, F. (2001). Syntagmatika slovníku: Typy lexikálních kombinací. In: Hladká, Z. - Karlík, P. (Eds.), *Čeština - univerzália a specifika 3: 223-232.* Brno: Masarykova univerzita.

Český národní korpus (Czech National Corpus). http://ucnk.ff.cuni.cz

Český národní korpus. Úvod a příručka uživatele (2000). Praha: FF UK - Ústav Českého národního korpusu.

Church, K.W., Mercer, R.L. (1993). Introduction to the Special Issue on Computational Linguistics: Using Large Corpora. *Computational linguistics 19/1, 1-24.*

Firbas, J. (1992). *Functional sentence perspective in written and spoken communication.* Cambridge: Cambridge University Press.

Halliday, M. A. K., Hasan, R. (1976). *Cohesion in English.* London*:* Longmans

Hrbáček, J. (1987). Srovnání dvou překladů z hlediska využití prostředků koheze textu. *Naše řeč 70, 123-130.*

Hřebíček, L. (1995). *Text levels. Language constructs, constituents and the Menzerath-Altmann law.* Trier: WVT.

Hřebíček, L. (1997). *Lectures on text theory.* Prague: Oriental Institute.

Hřebíček, L. (2000). *Variation in sequences.* Prague: Oriental Institute

Köhler, R. (1995). *Bibliography of Quantitative Linguistics.* With the Assistance of Ch. Hoffmann. Amsterdam: Benjamins

Komárek, M. (1978). *Příspěvky k české morfologii.* Praha: Státní pedagogické nakladatelství.

Mluvnice češtiny II (1986), **III** (1987). Praha: Academia.

Štícha, F. (2001). Anaforické koreferenční substantivum. In: Hladká, Z., Karlík, P. (Eds.), *Čeština - univerzália a specifika 3: 87-97.* Brno: Masarykova univerzita.

Trnka, B. (1950). Review of: G.K.Zipf, The psychobiology of language. Human behavior and the principle of least effort. *Časopis pro moderní filologii 33, 3-5.*

Wimmer, G., Altmann, G. (1999). Thesaurus of univariate discrete probability distributions. Stamm: Essen.

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse. (*= Studientextbücher Band 2). Wien: Edition Praesens.

Zipf, G.K. (1949). *Human behavior and the principle of least effort.* An introduction to human ecology. (1965: Facsimile of 1949 edition. Hafner Publishing Company: New York and London).

Supplement. Examples to Table 9 (in Czech).

*1.Tu a tam vyrůstal na písčitém dně **celý záhon žebrovaných černých dlouholistých řas**, a právě mezi **těmito záhony** žili sumýši.*

*2. V **interpretaci důstojnictva** podobala se válka jakési maturitě, kterou úspěšně a bez úhony na zdraví složí ti vojáci, kteří se v míru nevyhýbali ranním trenýrovkám a po večerech pilně studovali mechanismus brzdovratného zařízení. Ve válce, podle **této interpretace**, přicházeli o život toliko lemplové [...]*

*3. Náš pracující lid svěřuje naší lidově demokratické armádě **nejlepší zbraně** a dobře se dívá, jak se soudruzi vojáci učí **tyto zbraně** mistrovsky ovládat. A on také bděle odstraní každého, kdo by chtěl **těchto zbraní** použít.*

*4. **Nemocniční pokoj, márnice, smuteční obřad, hřbitov**. To, co o **těchto věcech** vím, se zvláštně podepisuje na jiných stránkách mého života. To, co o **těchto věcech** vím, vysvětluje, proč nemarním spoustu času sekáním trávy, mytím auta, hrabáním listí [...]*

*5. V nejstarších civilizacích existovala jako **nejvyšší bohyně** Velká matka [...]. Muž v dospělosti přenesl svou závislost na matce na **tuto bohyni**, a tím se od vlivu své matky [...]*

*6. [...] jako kdyby se zabýval nějakou velice důležitou činností, třeba vyzvědačstvím nebo tajnou službou, a na hlavě měl buřinku. Byl jsem schopen rozpoznat z docela slušné vzdálenosti, jestli se někdo chystá setkat se se mnou, a to se mi právě podařilo, když jsem zahlédl **tohoto muže** držícího v ruce kus papíru.*

*7. [...] co je obvykle pro každého, tedy i pro **komika**, prostě svaté. Nuže **komici, které mám na mysli**, nejsou sóloví baviči ze světa zábavy, ačkoli jsem poznal a dosud znám hodně **těchto komiků**.*

*8. Při oknech stál **kousek francouzského nábytku**. Přehoz na **tomto míšenci mezi křeslem a pohovkou** byl protkán fialkovými květy se čtverci v kruzích.*

*9. Mám u sebe **pistoli** a zbrojní pas. **Tato zbraň** je mým legálním vlastnictvím.*

*10. **Nemocniční pokoj, márnice, smuteční obřad, hřbitov**. To, co **těchto věcech** vím, se [...]*

*11. Formálně a důstojně, jak se sluší na dámu, nám oznámila, že **si přeje**, abychom ji pohřbili pod růžovým houštím. Novinkou na **tomto přání** byl fakt, že tentokrát si zvolila pro uložení svých ostatků místo tak snadno dostupné...*

*12. Seznámili se netradičně. Angličanka Minnie Sharpeová poslala Lukeovi svou fotografii v krajkové podprsence a připojila telefonní číslo... Taková akce vzbudila ve vyhlášeném milovníkovi pochopitelnou zvědavost. když se sešli, bylo rozhodnuto [...]. Vztah byl brzy korunován sňatkem a šťastné manželství teď už trvá tři roky. **Tento příběh** však patří k několika málo skutečnostem, které [...]*

*13. ... že jsem mu tou desetidolarovkou zamával před nosem. „Hele, kamaráde [...] Přijměte **tuto maličkost** jako projev našeho uznání [...]“*

*14. Zatímco sedím u počítače a píšu **tyto řádky**, můj manžel má puštěný televizor, kde běží fotbal.*

*15. Nevěřím, že jeden člověk na **tomto světě** hodně dokáže.*

*16. [...] kameny drásaly kůži, která byla do **této chvíle** nedotčena.*

*17. Dobrá, můj synu, předám ti **tuto moudrost**: Jdi a podívej se na [...]*

*18. Nevěděla přesně, zda i tuto knihu odnesla do **antikvariátu**. Mohla by i teď ještě podle těchto **antikvářů** nakreslit zpaměti mapu města.*

*19. A byl tu stále pes Alfreda Neugeborna, který tehdy Isabellu Goldblattovou tak poplašil. [...] Neřekla **tomuto Alfredu Neugebornovi** nikdy nic, a přece [...]*

Glottometrics 3, 2002, 61-89 *To honor G.K. Zipf*

An approach to the research of the structure of linguistic and musical texts[1]

A. Gumenyuk, A. Kostyshin, S. Simonova[2]

Abstract. The present article presents an approach to the study of specific arrangements of components in a separate linguistic or musical text. Concept and definition of the element order in the text are given and related formal models are described. The formulas for the calculation of integral numerical characteristics of element orders in texts, which take into account not only the structure but also the positional arrangement of the text components, are provided for the detailed description of a text. The distribution of numerical characteristics of element orders in a text is presented; so are the samples of research done on the construction of poems and pieces of music.

Keywords: order, arrangement, text structure

Object of research

The system of public education allows a human being to obtain the - from our point of view - unique ability to read and to write texts. While reading a real original text, there is a mental image formed in the human memory, which is not necessarily an adequate 'text-image' relationship. Writing an actual text image is anticipated by the creation of a mental "text-original" model. When the command of the relevant language is not sufficient, reading or writing the text can happen without the creation of understandable expressions. For non-experts in music, musical notes are just non-understandable musical texts. In any case, understandable or non-understandable, the expression read or recorded is reflected by a human being in the form of an imaginary "text". Thus, reading and writing are accompanied by the creation of a mental text being inaccessible for instrumental study, which is the result of special kind of human activity - mental activity or thinking. In this context a mental text is the track of mental activity in the human memory. The actual text is a track of human mental activity, which is stored in an external memory unit and, therefore, becomes accessible for instrumental processing.

The linear string structure of a text allows to assume that listening and speaking as special kinds of interaction activity between the internal and external human worlds are also strings of the events - actions which are reflected by mental pseudo-texts. Probably, a human being's interaction with the external world while reading and writing, listening and speaking does not differ principally from other kinds of his/her material activities. Therefore it is possible to assume that any separate human activity that has some objective, is a quantum process consisting of a string of interactions with the external world (events), reflected upon by a string of mental actions (thoughts) or mental pseudo-text.

[1] The authors thank their colleagues Professor V. Potapov and A. Florensov from Omsk and V. Kromer from Novosibirsk for the extensive help during the presented research.

[2] Address correspondence to: A.S. Gumenyuk or A.S. Kostyshin or S.V. Simonova, pr. Mira 11, OmGTU, Omsk, 644050, Russia. E-mail:inter@omgtu.ru, sha@omgtu.ru

At the beginning of the 20th century the study and comprehension of the pseudo-text nature of human activities and thinking were facilitated in various forms by such well-known scientists and philosophers like A. Poincaré, P. Florensky, M. Heidegger, M. Bachtin. M. Bachtin was the first Soviet literature analyst, who pointed out the following: "Humanities are sciences about the human being, about his/her specificity, rather than about a voiceless thing and natural phenomenon. A person in his/her human specificity always expresses himself (speaks), i.e. creates a text (even if it is potential). When a human is studied irrespective of the context of the text, these are not Humanities but anatomy, human physiology, " (Bachtin 1979: 285). "A text is a primary reality and the starting point of any humanitarian discipline" (Bachtin 1979: 292). In our understanding M. Bachtin defined texts as objects of research, and humanitarian disciplines as a science, which is to study human, i.e. mental activity, human nature. At present there are theories of artificial intelligence, which, from our point of view, do not exploit to a maximum M. Bachtin's research findings on the role of texts as objects of research of the thinking nature.

We will state that an actual literary text or notes of a musical composition - where not content and meaning, but similarities and differences, order and arrangement of characters, words and notes are studied - presents a special kind of a sequence of events, which represents the pseudotext-type quantum nature of a human being's physical and mental activities. Therefore, a text can be a relevant object of research of the thinking functions studied using the tools of computer science. Texts represent the most formalized outcome of human activities and, therefore, are best for computer processing of the experimental material during the analysis of the thinking nature.

History of research

However, the weak point of Humanities as defined by M. Bachtin, is the lack of quantitative methods of description, as it is not known what is to be measured in texts and how this is to be done. At present, for the quantitative evaluation and comparison of genres and authors' styles of texts the following approaches are used: random-statistical and entropic approaches, when alphabets (vocabularies), numbers (frequencies) of elements in the text structure and correlation of characters in words are taken into account.

The breakthrough in the research of texts was the works of J. Estoup, E. Condon, G. Zipf, B. Mandelbrot. As a result of these works a statistical distribution was discovered, where the rank or the number of a word in the series of different words, arranged according to the occurrence frequency descending, performed the function of "random" value. Having assumed that the rank distribution of words has the character of a language law, Mandelbrot used the hypothesis of the optimal coding of words and developed the updated analytical relation for a "statistical" rank distribution of words, in which the frequencies of their occurrence were determined by the two parameters of a text: frequency of the most frequent word and length of the text. This law is presented by the following formulas (Orlov 1980):

$$p_i = \frac{K}{(B+i)^{\gamma}}; \qquad K, B, \gamma - const \tag{1}$$

$$K = \frac{1}{\ln F_1}; \qquad B = \frac{K}{p_1} - 1, \tag{2}$$

$$v_T = KZ - B; \tag{3}$$

where p_1, p_i are frequencies of '*1st*' and *'i'* ranks of words resp. F_1 is the frequency of the most

frequent word. $\gamma \approx 1$, v_T is the theoretical volume of the text vocabulary, which has the length of ***Z.***

However, it turned out that the Zipf-Mandelbrot law was not the "limit" of the rank distribution of words tallies despite the very big text samples (i.e. the statistical law of large numbers does not work). Besides, in various text samplings words did not have the same probability of the frequency of occurrences, and vocabularies of the texts did not have the same content. Thus Mandelbrot's hypothesis on the language law in the form of rank distribution was not confirmed.

Thus, for the first time the specific character of text construction (and, in our understanding, the specific nature of human thinking) was pointed out by J. Estoup, E. Condon, G. Zipf, B. Mandelbrot, as they discovered the special frequency-rank distribution of words for various texts, which was called Zipf-Mandelbrot's law. However this law characterizes the construction of separate texts only by the special content of units (size of the alphabet, vocabulary, numbers of occurrences of text elements).

The essential breakthrough in the research of text structure took place in the seventies of the twentieth century, when the Soviet cyberneticist, Yu. Orlov, used the methods of random-statistical and entropic approaches and discovered the criterion of content of components in integral complete pieces of literature or music; that criterion is determined by the degree of coincidence of the actual rank distribution of words in a separate complete text according to the Zipf-Mandelbrot law (Orlov 1980). For fragments and conglomerates of texts such coincidence was not traced. It was quantitatively established that a complete text has a specific structure, in which its vocabulary, the length and the frequency of occurrences of identical words are bound by the Zipf-Mandelbrot law, which was supplemented by Orlov with formulas (2) and (3).

Thus, Yu. Orlov discovered that the Zipf-Mandelbrot distribution is the law applicable to a separate and integral complete text, but not to language. In this meaning Yu. Orlov, from our point of view, emphasized and specified the special construction of texts by pointing out not only to the specific objective composition, but also the systematic integrated nature of the distribution of elements under the Zipf-Mandelbrot law for a separate complete text. However, as well as his predecessors, Orlov leaves out the order of words in a text and characterizes its specific construction only by the size of the alphabet and the number of occurrence of elements.

In the seventies of the last century Polish scientist Marian Mazur was the first to indicate the importance of the sequence of word order and characters in texts for proper information. He discovered and defined the models of information and code strings of communicates[3], and developed the (qualitative) information theory on this basis (Mazur 1974). Within the framework of this theory he discovered the method and developed the formulas for counting the numbers of descriptive and identifying bits of information in information strings including linguistic texts, which generalize the formula of Claude Shannon for information quantity and entropy. However, Mazur's formulas give integral estimations only of the composition of character sequence and leave out the specific sequential order of characters in sequences (texts).

Formalism of element order in a string

Reading and writing texts is based on the apparent meaning of words, hieroglyphs, characters and symbols, which present corresponding objects or concepts. However, until recently al-

[3] The term "communicate" used as a noun has been adopted from Mazur's works.

most no attention was paid to the patterns of specific sequences of characters or words, which constitute a separate sequence of symbols, during the study of texts. Therefore, it was impossible to conduct the formal analysis of text element orders. At the same time musical texts, at first glance, do not have any content. However, the priority of the specific arrangement of musical characters is obvious for musicians. The arrangement of events in historical chronicles is essential. During the instrumental measurement of values it is important to record the natural order of data, usually it is easily represented by character sequence. In the present work the approach intended for the formal analysis of an order of elements in a separate text, any character sequence or string of messages, is considered. The methods applied to the study of the local structure of symbol sequences are not considered in the present paper.

Let's study a set of sign sequences of finite length. Generally, this countable set is part of a countable set of information strings of communicates as defined by M. Mazur. Let's select a subset of tuples with identical sets of number of occurrences of characters from its own alphabets. Being sequenced according to the descending of numbers of occurrences, each set of this kind becomes the rank distribution. Thus, the studied tuples have identical rank distributions. Let's assume that the selected set constitutes sequences with equivalent content. As we assume several occurrences of some elements, the whole set of tuples composed on the basis of a certain alphabet presents a combination of a “permutation with repetitions”, each of those differs by the original arrangement of components and their own composition or architecture.

Let the reading of each tuple be done by element-wise steps from left to right. When using this usual scan method we replace all first opposite different signs of the alphabet (characters, communicates, words) by integers, beginning with “one", and then continue increasingly depending on the number of occurrences of the next new element of the given alphabet. Other repeated signs of the given sequence are presented by numbers assigned during the first scanning of the text. A tuple in which signs are substituted by numbers, i.e. the number of the occurrence of elements in the alphabet, is constructed this way. We shall call it a **number sequence, order of elements** or **constitution of a string** (Gumenyuk 2000). As a result of this transformation a countable set of texts or strings of messages having equivalent structures (with identical, overlapped or different alphabets - dictionaries) is presented by a single finite set of number sequences, the components of which are integers increased by one unit only if a new element of the alphabet is found and the text is read from left to right. Let's divide all set of texts with equivalent content into non-intersected subsets according to the principle of correspondence to this or that numbered sequence. As a result, all the sign sequences equivalent in their content and having an identical arrangement of elements (composition), are presented by one and the same numbered sequence, which represents the original for the given subset of texts order of characters or words.

Q	***R***	***D***	***&***	***V***	***Y***	***S***	***S***	***&***	***S***	***D***	***D***	***S***	***Q***	***S***	***D***	***&***
W	***T***	***C***	***G***	***H***	***U***	***R***	***R***	***G***	***R***	***C***	***C***	***R***	***W***	***R***	***C***	***G***
T	***Y***	***M***	***H***	***T***	***O***	***S***	***S***	***H***	***S***	***M***	***M***	***S***	***T***	***S***	***M***	***H***
U	***I***	***L***	***J***	***W***	***O***	***Y***	***Y***	***J***	***Y***	***L***	***L***	***Y***	***U***	***Y***	***L***	***J***
V	***N***	***A***	***B***	***J***	***K***	***T***	***T***	***B***	***T***	***A***	***A***	***T***	***V***	***T***	***A***	***B***
1	**2**	**3**	**4**	**5**	**6**	**7**	**7**	**4**	**7**	**3**	**3**	**7**	**1**	**7**	**3**	**4**

sequence (string) of symbols
sequence (string) of symbols
sequence (string) of symbols
sequence (string) of symbols
sequence (string) of symbols
order of elements

The order of elements in a sign sequence or information string of the communicates is characterized by the power of its content (but not only by the alphabet) and by the original composition of components. All texts of a subset of the selected decomposition are isomorphic with regard to their order of elements. It is pointed out that the given countable

subset of information strings of communicates also possesses its own numbered sequence, which we call 'order of a string'; however, it is the string that - in the above mentioned meaning represents - ***the order of elements*** in a text, i.e. ***the sequence of various events***. Finally, the sign tuples with contents of various power (but, probably, with identical or equivalent alphabets or vocabularies) according to their definition differ by the order of elements in a string.

In the broader informal sense, the specific order of elements fixes the identical sequence of events. These can be chronicles of historical events, texts, sequences of data of a human being's subjective conditions, strings of the communicates, arrays of measurable values.

Thus, for the study of the actual text architecture it is offered to use a formal concept, ***order of sign sequence,*** which only represents a definite sequence, an arrangement of discernible and identical components disregarding their designation and contents.

The operation of detection in various strings of different nature in communicates of identical orders of elements expands the capabilities of interdisciplinary research; however, the outcome of this procedure is limited to the description of an order of elements in the form of an ordinary numerical tuple. In Gumenyuk (2000) the formal description of an element order is considered, which is more convenient for the numerical analysis and allows to obtain the approximated compact numerical characteristics (similar to the ones used for the description of random variables). The characteristics are useful in particular for the identification of the element order and determination of the degree of their difference. To define such formalism of the element order in a separate text using the ordinary way of its "element-wise subsequent" reading we perform two numberings:

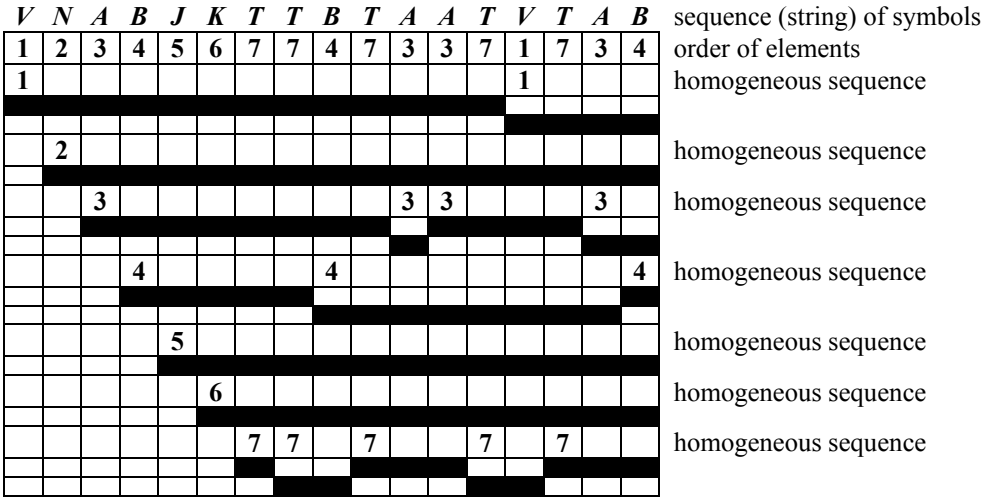
- The first one ascribes numbers to elements of its own alphabet (vocabulary) of the given sign sequence according to their first occurrence;
- The second one sets end-to-end numbering of all tuple components from the beginning to the end.

Let's decompose a complete (i.e. without empty positions) variable character sequence into ***'m'*** of the non-complete invariable tuples, in which only some positions are filled by identical signs (or signs selected according to a specific rule). An invariable sequence is considered in analogy to the flow of homogenous events studied by the probability theory. Let invariable sequences of the given decomposition be incompatible (i.e. they do not have identical numbers at the positions filled). The composition or addition of all ***'m'*** of non-complete invariable tuples provides the initial complete variable sign sequence. The decomposition of texts into invariable sequences was probably done for the first time by V. Leus (1987).

Let's define the interval of "time" (hereinafter - interval) by the number of positions from one selected component to the other nearest component, which is marked in the direction of scanning; then define the increment of number of the component (hereinafter - increment) as the difference between the numbers of components in the list of the natural sampling taken from the alphabet. Let's agree that the reading of the text will be done by steps of observation, and at each step there is the next event fixed by the increment and the interval.

Let the first scanning of a text be performed by a different, not ordinary method from the very beginning up to the end in such way that only elements with number "1" are selected; thus, the last interval is determined by the "finish" sign. Intervals of the given invariable sequence are placed in accordance with the numbers of the scanned elements in the first line of the matrix. Then, during the second scanning of the order of text elements (with the same method being used), units with number "2" are selected and the vector of intervals appropriate to the other invariable sequence is placed in the second line of the matrix. During the

new scanning the vector of intervals of invariable sequence is placed as <new> in each subsequent line of the matrix. The increment of number is increased by one unit only. Single signs, words or messages will be presented by one interval only (up to the end), which is placed in the first column of the corresponding line of the matrix. The number of columns n_{jmax} in the matrix of intervals is equal to the number of occurrences of the most frequent sign or word of the text. Vacant intervals will be filled with zeros. The number of lines ***'m'*** is equal to the power of its own alphabet or vocabulary of the text.



13	**4**	**0**	**0**	**0**
16	**0**	**0**	**0**	**0**
8	**1**	**4**	**2**	**0**
5	**8**	**1**	**0**	**0**
13	**0**	**0**	**0**	**0**
12	**0**	**0**	**0**	**0**
1	**2**	**3**	**2**	**3**

Matrix of intervals

The above presented interval description of the text structure in the form, which is obtained by the special procedure of bulk scanning of invariable sequences, can be recorded by an ordinary tuple, components of which are the "increment-interval" pairs; when the first component equals ***'1'*** only during the transition from one invariable sequence to another one, the other first components of the ***'2's,*** which describe each invariable sequence, are equal to zero. Such tuples of "increments-intervals" can present the element order in some texts in a more compact way than a matrix. A convenient formalism of the interval description of an order of elements is the tuple of complex numbers, components of which are the components of the "increment-interval" of the '2's.

The tuple of an interval order obtained by the natural procedure of scanning consists of pairs, the second component of which is always equal to '1'. During the ordinary recording of a text all intervals by default are equal to '1', that is why they are not marked by a special number.

Except for the items described above and, in some way, contrary methods of scanning, it is possible to use an optional method of reading (starting with any element), each step of which is an event marked generally both by positive and negative values of increments and intervals. The negative interval (backward interval) means reading the sign from a position, the number of which is lower than the number of the position provided at the previous step. The negative increment of number of the character means the decreasing of the number of the scanned sign in contrast to the number of the character observed at the previous step. Thus,

several identical incorrect readings of the original may occur, but, however, several identical correct readings of the original are possible as well.

Numerical characteristics of the element order in the string

We use the concept of invariable sign sequence and its vector display in the form of the corrresponding line in a matrix of intervals for the definition of the numerical characteristics of the element order in a text (Gumenyuk, 2001). The multiplication of all intervals of the selected ***j***-th invariable sequence (elements in the corresponding matrix line, except for zero) determines its absolute volume in the following way:

$$V_j = \prod_{j=1}^{n_j} \Delta(j)_i \,, \tag{4}$$

Where $\Delta(j)_i$ is the interval between ***i***-th and (***i + 1***)-th occurrence of the ***j***-th symbol, n_j is the number of occurrences of the ***j***-th symbol. The mean geometrical interval between the filled positions of an invariable string determines the volume of the selected ***j***-th symbol in the following way:

$$\Delta(j)_{me} = \sqrt[n_j]{V_j} \,. \tag{5}$$

The volume of the text is determined as the product of absolute volumes of all invariable sequences in the following way:

$$V = \prod_{j=1}^{m} V_j \,, \tag{6}$$

by substitution (4) we get:

$$V = \prod_{j=1}^{m} \prod_{i=1}^{n_j} \Delta(j)_i \tag{7}$$

where ***m*** is the size of the alphabet (vocabulary of the text).

The mean geometrical interval of the set of all invariable strings in the text determines the volume of the separate symbol in the following way:

$$\Delta_{me} = \sqrt[n]{V} \,, \tag{8}$$

where ***n*** is the text length equal to the number of all its positions.

Taking to a logarithm the presented values yield a set of convenient additive informational characteristics of the construction. At this time the interval is substituted by the distance of a certain ***j***-th symbol of the ***i***-th occurrence with regard to its ***(i + 1)***-th occurrence in the form of $\log_2 \Delta(j)_i$, the volume of invariable sequence of the selected ***j***-th symbol is substituted by the depth of arrangement of its elements in the form of $G_j = \log_2 V_j$; by the substitution (4) we get the following:

(9) $$G_j = \sum_{i=1}^{n_j} \log_2 \Delta(j)_i \, .$$

Below there is an example of the calculation of the location depth of one symbol using the method of element calculation of intervals of the homogeneous string of elements.

V	***N***	***A***	***B***	***J***	***K***	***T***	***T***	***B***	***T***	***A***	***A***	***T***	***V***	***T***	***A***	***B***	sequence (string) of symbols
1	**2**	**3**	**4**	**5**	**6**	**7**	**7**	**4**	**7**	**3**	**3**	**7**	**1**	**7**	**3**	**4**	order of elements
		3								**3**	**3**				**3**		homogeneous sequence

$G_3 = \log_2 8 + \log_2 1 + \log_2 4 + \log_2 2 = 6$ bits calculation of the location depth of the selected symbol.

The volume of the text is substituted by the depth of arrangement of all its elements in the form of $G = \log_2 V$, by substituting (6) and (4) we get the following:

(10) $$G = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log_2 \Delta(j)_i \, .$$

When comparing constitutions of different texts, the evaluation of the relative depths of arrangements is conveniently done in the form of $\delta G_j = G_j / G$.

Evaluation in the form of arithmetic average depths of the arrangements of symbols in the sequences (mapping mean geometrical intervals) is useful. The average distance of the selected ***j***-th symbol in an invariable sequence is derived from the equation (9) and calculated in the following way:

(11) $$g_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \log_2 \Delta(j)_i$$

From the equation (10) the mean distance of a separate symbol in the given text is calculated by:

(12) $$g = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log_2 \Delta(j)_i$$

The ratio of the mean distance of the selected ***j***-th symbol and the mean distance of a separate symbol gives the characteristic of the order of text elements, which supplements the frequency of occurrence of $p_j = n_j / n$ and is called ***relative distance*** of the ***j***-th symbol. And for any complete sign sequence (without empty positions) the sum of products of occurrence frequencies of the selected elements at their relative remoteness is equal to '***1***' as follows:

(13) $$\sum_{j=1}^{m} \frac{G_j}{G} = \sum_{j=1}^{m} \frac{n_j}{n} \cdot \frac{g_j}{g} = \sum_{j=1}^{m} p_j \cdot \delta g_j = 1$$

The ordering of the set of values of $\{G_j\}$ or $\{\delta G_j\}$ according to the descending depths of element arrangement of all invariable sequences is called ***rank distribution*** of depths of element arrangement of invariable sequences for the construction of a particular text. In contrast to the rank-frequency distribution, the distribution data contain fewer elements (signs, words) with identical characteristics and, accordingly, with undefined ranks. During the analysis and description of the order of text elements, the rank distributions of depths of arrangement are supplementing the ordinary frequency distributions by information on the original positional relationship of elements of the text under study.

Let's state, how the measures of the information applied by M. Mazur and C. Shannon are connected with the integral characteristics of the studied construction of a particular text. First, while studying an information string (text), Mazur considers the sequence of the communicates (signs, words) in the string essential. Thus, to get the proper message from the source to the receiver, it is necessary to keep the same order of the messages in all information strings of the given communication channel, beginning with the strings of the originals and finishing by the strings of images. So, for example, during the reading the mental "text - image" should keep the sequence of words, which is present in the real original text. However, during the calculation of the number of pieces of information in the information string containing not only various separate messages, but also groups of the identical messages, the formulas of M. Mazur only use the content of string elements (Mazur 1974). Arrangement of string components, or their ***composition,*** is not measured at this point. So, the number of pieces of descriptive information in a particular information string (text) is calculated according to the formula:

$$D = \prod_{j=1}^{m} \left(\frac{n}{n_j} \right)^{n_j / n}. \tag{14}$$

The number of the identifying pieces of information in a particular information string is calculated by the fundamental formula of M. Mazur as $J = \log D$, which by substitution (14) is as follows:

$$J = \sum_{j=1}^{m} \frac{n_j}{n} \cdot \log \frac{n}{n_j}. \tag{15}$$

Equation (15) permits the derivation of the following formula for the sequence of unlimited length, when $n \to \infty$:

$$J = H = -\sum_{j=1}^{m} P_j \cdot \log P_j \tag{16}$$

where P_j is the probability of the ***j***-th element. Shannon, in contrast to Mazur, suggests a statistical model for the transmission of communicates. In this model the source generates not a text, but an infinite random sequence, in which elements "are connected" only statistically. Text here is understood as a tuple, i.e. an ordered set of final power. Accordingly, the receiver is not capable of taking the sequence of communicates in a string into account; its function is only to identify the next communicates.

Let's calculate value V_j from (5) as follows:

(17) $$V_j = \Delta(j)_{me}^{\,n_j}$$

using formulas (8), (6), (17) we get

(18) $$\Delta_{me} = \prod_{j=1}^{m} \Delta(j)_{me}^{n_j/n} .$$

Taking to the logarithm of this value will provide the equation for mean distance of a particular symbol as follows:

(19) $$g = \log \Delta_{me} = \sum_{j=1}^{m} \frac{n_j}{n} \log \Delta(j)_{me} .$$

For a single case, when all intervals of each invariable sequence of a text are equal, i.e. $\Delta(j)_i = \Delta(j)_{me\max} = \frac{n}{n_j}$, $\forall j = \overline{1,...,n_j}$, we use formulas (18) and (19) to derive formulas (14) and (15) accordingly. Thus, Mazur's formulas only allow the correct evaluation of the element order of regular pseudo-texts. The number of pieces of descriptive information is equal to the maximum volume of a separate character, $D = \Delta_{me\max}$ and the number of the identifying pieces of information is equal to the maximum mean distance of a separate character $J = \log \Delta_{me\max}$. If there are texts and other non-regular sequences, Mazur's formulas provide an evaluation of the order of text elements only "from above", as in these cases $D > \Delta_{me}$, $H > g = \log \Delta_{me}$. Accordingly, the numerical characteristics of the order of elements take minimum values for such sequences in which all identical units are arranged one after another.

Thus, formulas for the volume of a separate symbol and its mean distance generalize the Mazur's and Shannon's formulas and, in contrast to the latter, while characterizing the order of elements in a separate text, not only elements of its content are taken into account, but also the arrangement of components (signs, words).

Let the receiver have the ability to determine the arrangement of characters in the given invariable string with regard to the positions held by signs of another invariable string (Gumenyuk 2001). Let's consider two invariable strings with identical number of positions held by signs, while the remaining positions in invariable sequences are empty. Let's designate the interval between two adjacent (i.e. with identical number of occurrences 'i') signs ***j*** and ***k*** of different invariable strings by the symbol $\Delta(j/k)_i$, if the interval for the ***j*** symbol is determined with regard to the position of the ***k*** sign; if the next interval is determined with regard to the position of the ***j*** sign, it is denoted as $\Delta(k/j)_i$. Apparently $\Delta(j/k)_i = -\Delta(k/j)_i$. This is the basis for the further calculation of deviations between positions of the adjacent signs ***j*** and ***k*** of various invariable strings as follows: $|\Delta(j/k)_i| = |\Delta(k/j)_i| = \Delta(j,k)_i$

The product of all deviations defined for two invariable strings is as follows:

(20) $$V(j/k) = \prod_{i=1}^{n_j=n_k} \Delta(j,k)_i$$

Let's call it the relative volume of the selected (***j*** or ***k***) invariable sequence.

An invariable string is a string which depends on another one, if its relative volume is smaller than absolute, i.e. when $V(j/k) < V(j)$; if $V(j/k) \geq V(j)$, then the string is composed of symbols 'j' and does not depend on string 'k'.

Let's consider the ratio between the absolute volume of the given invariable string and its own volume, defined with regard to the positions of symbols in another invariable string (applies when these sequences are dependent). The value obtained is as follows:

$$\overline{V}(j/k) = V(j)/V(j/k), \tag{21}$$

will be called the relative excess volume of the selected ***j***-th of the invariable sequence. If the symbols of the string ***'j'*** are located directly before or after symbols ***'k***', the relative volume is as follows: $V(j/k) = 1$, and its relative excess volume is equal to absolute. Thus, the absolute volume of the selected invariable sequence, measured regardlessly of other strings , is equal to the product of its relative volume and relative excess volume.

If all intervals $\Delta(j/k)_i$ have the same sign (positive or negative), then one of the invariable strings will be called induced or string of consequences, and the other will be called ***generating or reasoning string.*** The data of the string will be considered having a “cause and effect" relation. If the receiver is able to determine relative volumes of symbols in the third strings in relation to the second ones, for which the volumes in relation to the first strings are determined, and then the fourth strings in relation to the third ones and so on, there is a possibility to determine the deeper cause and effect relations. Thus, it is possible to select “main cause strings”, “immediate causes” and “direct effects", “remote causes” and “deep consequences” etc. among the strings of the messages.

If the receiver, except for the indicated one, is able to rank invariable strings of a separate text according to the descending values of their absolute and relative volumes, it is possible to calculate the volume of the whole text as follows:

$$V = V(1) * V(2/1) * V(3/2,1) * \ldots * V(m/(m-1),(m-2),\ldots,2,1) \tag{22}$$

Whereas $V(1)$ is the volume of the first in the rank of an independent invariable string, $V(j/(j-1),(j-2),\ldots,2,1)$ is the volume of the ***j***-th invariable sequence, defined in relation to the signs of the ***(j-1)*** -th string, for which, in turn, the volume is determined in relation to the filled positions of one more string etc. The strongest dependence among invariable strings is the functional dependence, when each current interval of one string is determined by an identical relation to the next interval of another string (according to Mazur — primary code). The simplest functional relationship fixes the equality of adjacent intervals of different strings $\Delta(j)_i = \Delta(k)_i$, and this can be the composition of characters to form words.

Thus, if there are actual relationships between invariable strings of communicates and the required abilities for the detection of this phenomenon, it is possible to calculate the volume of the whole separate text by the multiplication of absolute and relative volumes of symbols. If the analysis of strings which constitute texts shows their independence, then the volume of the text composed from independent strings is determined by the multiplication of absolute volumes of all invariable strings; see (6).

Examples of a numerical analysis of poetic and musical texts

The present paper briefly shows the examples of research of the full element order in poetic and musical texts on the basis of the above mentioned approach (Gumenyuk, Kostyshin, Simonova 2000). The usage of these particular types of texts is not conditioned by the capabilities of the developed approach and special methods, which are invariant as to the type of the text, i.e. are intended for the research of the element order of any sign sequences and data arrays. Our interest in the texts of such nature is to some extent conditioned by three factors.

First, it is conditioned by the problem of invariable segmentation of a separate text into information units, including word. Due to what is apparent for a specialist the separability of words in a literary text in the works of Orlov, the problem of invariable segmentation of a text into words was almost not studied. In this context he writes the following (Orlov 1980: 79): "Any work is a complex unit, consisting of sets of relatively simple elements. In literature these are words, in music - sounds, in painting - colors. It is a rare exception that one unit is easily separated from other ones... ".

In practice, however, during the separation and distinguishing of text elements certain difficulties arise, especially, when this operation has to be completely formalized and automized. So, even distinguishing words separated from each other by blanks is not possible in a fully automatic mode.

The authors of the present paper like to draw the reader's attention to the existence of the problem of invariable segmentation of texts into elementary information units even in case of completed texts.

It is necessary to note that any analysis (including informal or expert analysis) becomes practically impossible and rather subjective, when the selection of primary elements in complex structurally organized objects, and this is what texts are, is difficult. At the same time M. Boroda has developed a procedure for invariable (proper) segmentation of a musical text (a single-voice of piece of music) into basic acoustic units — F-motives (units of phonemes) (Boroda 1978), which we have applied successfully to the computer analysis of musical texts. However, invariable segmentation of polyphonic and orchestral pieces of music is still a problem.

Secondly, our attention to the architecture of poems and pieces of music is conditioned by the absence of the generally accepted vocabularies for primary elements which constitute texts. In poetry and music (except for homophone and single-voice pieces of music) it is not clear at all what units form the basis for the integrated order of elements of similar acoustic objects.

Thirdly, from our point of view, poetic and musical texts reflect a similar harmonically organized nature of acoustic objects for the best perception by a human being.

But we do not intend to completely substitute the ordinary expert analysis of poetic, musical, literary and other pieces of art by formal procedures and evaluations. We speak about auxiliary numerical-analytical tools and quantitative characteristics, which can supplement the ordinary informal professional tools.

For computerized numerical research of element order, the poems of Russian poets and musical texts of homophone and single-voice pieces of classical music were analyzed.

First of all, the problem of segmentation of poems into basic acoustic units was solved. As stated above, from Orlov's point of view, the set of elementary components, which constitute a single complete text, should be subject to the Zipf-Madelbrot law. Realizing the importance of searching the vocabulary of such components, whose rank distribution would correspond to the given law, the authors concentrated on the reverse problem: to attempt to

obtain invariable segmentation of an apparently completed poem text into elements (Gumenyuk, Kostyshin 1998).

The authors named the number of features, on the basis of which the formal identification of the "completed" text was done, after the scientist Orlov, who had discovered the criterion of content of components in a single complete piece of art, and called it the Orlov criterion. The authors have included in the structure of the given criterion the following factors according to their priority:

- Accuracy of coincidence as to the powers of actual and theoretical vocabularies of elements in a separate text;
- The degree of coincidence of actual rank distribution with the Zipf-Mandelbrot law, which can be presented quantitatively;
- By maximum relative frequency deviation of the selected element from its theoretical probability under the given law;
- By mean relative deviation of the actual frequencies from theoretical probabilities of elements in a separate text vocabulary.

Thus, when Yu. Orlov considers that there are elements of a text and determines its finality, the authors of the present article, vice-versa, had assumed in the beginning of their research the apparent finality (integrity) of a poem text, and then attempted with the help of Orlov's criterion to segment a text into elements.

Let's consider some foundations for the definition of the 'vocabulary' of basic acoustic objects in a separate text. From the point of view of Orlov's systemic criterion, in particular, the number of various elements of vocabulary in a single piece of art is connected with the text length and determined by the formula (3) of the Zipf-Mandelbrot law. This limitation does not allow to use characters of the alphabet of a language as elements of poems. Besides, not all characters of the alphabet represent sounds, a specific sequence of which makes an acoustic matter of a poem. Phonemes of a language, which upon agreement are considered minimum elements of its sound composition, also make a fixed set. Words as elements of a poem are not suitable as they are used in such short texts usually once. Text segmentation into units with an equal number of characters was rejected at once, as it was disharmonious to the acoustic nature of a poem, even if it met the Orlov criterion. Therefore, the authors have chosen the way of searching for acoustic elements of poems. The elements were in the form of phonemic units including basic sounds whose number ranged from one to several. The irregular phonemic units called consonance (combination of sounds), probably express the constitution of a poem in a more exact way by such combinations of sounds, which constitute its acoustic harmony. Let's note that in this case the length of a text of a poem is measured by the actual frequency of all combinations of sounds (consonance) in it, similar to the length of a large literary text and is determined by the frequency at which words are used.

Let's study in closer detail the method of digital computer analysis and segmentations of the texts of poems, which was developed by the authors.

The limitation to the existence of acoustic units of poems introduced by the authors did not allow to use files of the source texts directly. Therefore, first, all poems were rewritten with all the phonemes of the language being written down in their alphabetical order and checked by an expert. The Russian literary language comprises 41 phonemes: 6 vowels and 35 consonants. For "a phonetic compiler" the algorithm constructed on the basis of materials of (Grammatika russkogo jazyka 1960), which provide the rule for the reading of characters and their combinations in the whole style of Russian literary language, was used. The output of the phonetic compiler is a text, which is a phonetic record of the source alphabetic text. Due to the complexity of the algorithm of transcription it is not possible to program the text compilation into sounds of speech. And even when a self-learning algorithm is available, the

participation of a person supporting the solution of various semantic problems is required. The obtained compilation was represented as the source text, which was to be used for optimum segmentation into consonance (combinations of sounds).

The authors have developed the method for the search of the best segmentation ("best" according to Orlov's criterion) of such phonetic text (Gumenyuk, Kostyshin 1998). This method uses the known procedure of statistical solutions for the selections of words from a continuous text (Borodovskiy, Pevzner 1990).

Let's study the idea of procedure for the formation and selection of pseudo-words in a text (in our example – consonances). While doing the sequence scanning of a text to select pseudo-words, which suit the following selection criteria, the sequence of characters is considered to be a pseudo-word, if it has an unexpectedly high or low frequency in the text. When predicting the frequency of the occurrence of a pseudo-word in the text the frequency of occurrence of its pseudo-words is used. For example, the expected frequency of n-character pseudo-word $B_1,...,B_n$ is calculated through the visible frequencies of occurrences of (n-2)- and (n-1)-of the pseudo-words in letters in the form:

$$E(B_1,...,B_n) = \frac{f(B_1,...,B_{n-1}) \times f(B_2,...,B_n)}{f(B_2,...,B_{n-1})} \quad (23)$$

(Formula 2.3 in Borodovskij, Pevzner 1990), created in accordance with the marked model of (n-2)th order). To evaluate the degree of deviation of actual frequency from its expected value the following value of standard deviation is used:

$$std(B_1,...,B_n) = \frac{|f(B_1,...,B_n) - E(B_1,...,B_n)|}{(E(B_1,...,B_n))^{1/2}} \quad (24)$$

At this time the combination of symbols of $B_1,...,B_n$ is considered to be a pseudo-word, if $std(B_1,...,B_n)$ exceeds a certain limit value. The essence of the algorithm is found in the following: the text is scanned consecutively and the combinations of symbols of the n-length are selected: the first combination from the beginning of the text is taken $B_1,...,B_n$, and the value of standard deviation is calculated for it $std(B_1,...,B_n)$. If $std(B_1,...,B_n)$ is higher than the limit value, then this combination and all the other entries to the text are selected as pseudo-texts, and they are put into the vocabulary. Then the next combination is taken $B_{n+1},...,B_{2n}$. If $std(B_1,...,B_n)$ is lower than the limit value then the following combination is taken: $B_2,...,B_{n+1}$, for this combination again the standard deviation is calculated: $std(B_2,...,B_{n+1})$, when the calculation of the actual frequencies is done, the number of actual occurrences of sub-words does not include already selected pseudo-words, as they are already entered in the vocabulary, and cannot be used in the formation of pseudo-words. When using this method, the 'window' with the length 'n' symbols is moved to the end of the text. When it reaches the end of the text, the size 'n' of the window for a pseudo-word decreases and is reduced by 1, then the return to the beginning of the text materializes. The algorithm works that way as long as the size of the window of a pseudo-word 'n' is more than 2. When n = 2, the expected frequency of the occurrence of a two-letter pseudo-word B_1,B_2 is calculated according to the formula

$$E(B_1,B_2) = f(B_1) \times f(B_2), \quad (25)$$

which represents the variant of the formula (23). In formula (25) the expected frequency of a two-letter pseudo-word is determined as the multiplication of true statistical probabilities of the first and the second letters.

The experiments have shown that segmentation of the texts of poems was done in units consisting of no more than four phonemes.

The attempts of optimum segmentation of eleven texts into consonances were done within the framework of the two models of poem reciting: without spaces between words and with spaces only between the last line of a verse and the first of the following one (the example of segmentation with pauses only between verses is given in Attachment 1). Thus, it was possible to obtain the insignificantly different consonance 'vocabularies' in the studied poems.

Using the formal procedure of segmentation (Kostyshin 1998) for the search of the best segmentation of a text into consonances and assuming the absence of its natural division into words, it was possible to find a separation meeting Orlov's criterion for all eleven poems. The parameters of text segmentation are shown in lines (a) of Table 1.

As during the recital a poem is no continuous acoustic sounding and pauses are heard at the end of each verse (line), pauses at the end of lined were used for more adequate reflection of the structure of the acoustic matter of a poem. The results of segmentation of the same eleven poems, but without pauses are shown in lines (b) of Table 1. For all poems it was possible to find segmentation with pauses meeting Orlov's criterion.

Table 1

	Title and Author	Z	F_1	p_1	v	v_T	δ_V
a	St.Peterburg's verses. O. Mandelshtam	438	84	0.192	99	99	0.00%
b	**St.Peterburg's verses. O. Mandelshtam**	**438**	**83**	**0.189**	**97**	**99**	**2.02%**
a	The memory of Marina Tsvetayeva. B. Pasternak	608	38	0.063	165	164	0.61%
b	**The memory of Marina Tsvetayeva. B. Pasternak**	**591**	**29**	**0.049**	**170**	**170**	**0.00%**
a	Winter morning. A. Pushkin	404	36	0.089	111	111	0.00%
b	**Winter morning. A. Pushkin**	**400**	**35**	**0.086**	**111**	**110**	**0.91%**
a	Liberty. A. Pushkin	1085	45	0.041	277	280	1.07%
b	**Liberty. A. Pushkin**	**1098**	**54**	**0.049**	**271**	**271**	**0.00%**
a	Blessing. A. Pushkin	931	44	0.047	237	241	1.66%
b	**Blessing. A. Pushkin**	**908**	**38**	**0.042**	**245**	**244**	**0.41%**
a	Winter. A. Pushkin	771	28	0.036	225	224	0.45%
b	**Winter. A. Pushkin**	**785**	**32**	**0.041**	**221**	**220**	**0.45%**
a	Cleopatra. A. Pushkin	965	52	0.054	240	241	0.41%
b	**Cleopatra. A. Pushkin**	**942**	**40**	**0.042**	**247**	**250**	**1.20%**
a	Horse riders. A. Pushkin	823	41	0.050	217	217	0.00%
b	**Horse riders. A. Pushkin**	**829**	**41**	**0.049**	**218**	**219**	**0.46%**
a	Autumn morning. A. Pushkin	451	23	0.051	141	139	1.44%
b	**Autumn morning. A. Pushkin**	**437**	**19**	**0.043**	**143**	**142**	**0.70%**
a	A song on the Blessed Oleg. A. Pushkin	1322	60	0.045	326	319	2.19%
b	**A song on the Blessed Oleg. A. Pushkin**	**1373**	**86**	**0.063**	**308**	**306**	**0.65%**
a	Parting. A. Pushkin	574	42	0.073	151	151	0.00%
b	**Parting. A. Pushkin**	**566**	**33**	**0.058**	**159**	**158**	**0.63%**

Z - Length of a poem defined by the number of consonances;
F_1 - Number of occurrences of the most frequent consonance;
p_1 - Relative frequency of the most frequent consonance;
v - Power of the actual vocabulary (number of different consonances in the given poem);
$\boldsymbol{v_T}$ - Size of the estimated vocabulary calculated using formula (3) of the Zipf-Mandelbrot law;
δ_V - Deviation of power of the actual vocabulary of consonances from the calculated;

The analysis of the results obtained gives some reasons for assuming relative stability of the vocabulary of consonances in the Russian poetic language within the framework of two models — continuous poems and poems broken into lines. The selection of the representative vocabulary of consonances in the poetic language requires research of a greater sampling of poems by various authors.

Table 2 below shows the results of computer processing of several more poems composed by Russian poets. In this experiment the merging of phonemes into consonances was done only within the limits of one line, that is the existence of pauses in a poem was taken into account.

Table 2

Title of a poem and its author	**Z**	**F_1**	**p_1**	**v**	**v_T**	**δ_V,%**	**δ_{cp},%**
«Siplet cheremukha…» ("The tree in blossom….") S. Yesenin	185	15	0.081	63	64.7	2.8	10
«Krai ti moi…» ("My land…."). S. Yesenin	199	12	0.060	70	74.4	6.3	10
«V tom krayu…» ("In that Land…"). S. Yesenin	366	20	0.055	110	117.1	6.4	11
«Za tyimnoy pryadyu ……» ("Behind the dark…"). S. Yesenin	306	17	0.056	94	102.7	9.2	9
«Ulibka tomitelnoi skuki……» ("A smile of boredom..") A. Fet	277	18	0.065	90	91.5	1.7	6
«Mi odni iz sada…». ("We are alone, from the gardenr…") A. Fet.	527	28	0.053	142	153.5	8.1	9,3
«Kogda mechtatelno…» ("When dreaming…") A. Fet.	352	21	0.060	112	111.1	0.79	8,6
Village. A. Fet.	370	29	0.078	104	107.1	2.97	11

Z, F_1, p_1, v, v_T, δ_V — see the relevant explanations under Table 1**;**
δ_{cp} **-** mean relative deviation of the actual rank distribution of consonances in texts from the Zipf-Mandelbrot distribution;

As it is seen in the tables, it was possible to achieve such splitting of poems into consonances that under the Zipf-Mandelbrot law their actual and computational vocabularies of consonances display good agreement. The deviation does not exceed 10 %. The best agreement between empirical and theoretical vocabularies (1-2 %) is obtained for poems by A.S. Pushkin. Actual and calculated curves of rank distributions also have a high degree of correspondence. (The mean relative deviation for all texts does not exceed 11 %). It is also illustrated by curves shown in Fig. 1 (for the poem “Winter” by A.S. Pushkin.)

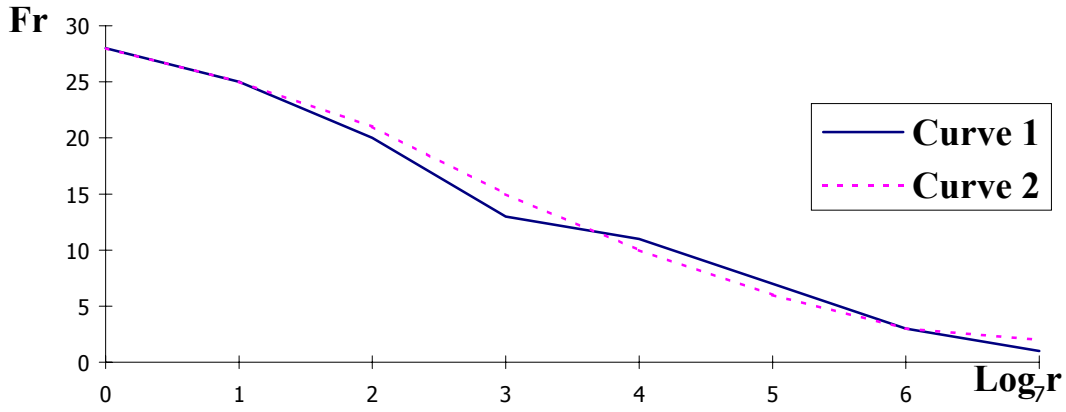


Fig. 1. Rank distributions of numbers of occurrences of consonances in the poem "Winter" by A.S. Pushkin.

Fr - Frequency of a consonance of rank r;
r - Rank or serial number of an element included in the vocabulary of consonances of the text in rank distribution, ordered descendingly with respect to the value of the numerical characteristic (here - number of the occurrences Fr);
Log r - Binary log of a rank;
Curve 1 - Actual rank-frequency distribution;
Curve 2 - Computed distribution (according to the Zipf-Mandelbrot law).

Table 3 presents the results obtained by the processing of the same poems by the way of entropic and numerical characteristics of an order of text elements.

Table 3

The title of a poem and its author	Z	Hs	G	H	g
A song about St. Oleg. A. Pushkin	1322	9753.01	8466.94	7.38	6.40
Freedom. A. Pushkin	1085	7804.86	6659.05	7.19	6.14
Cleopatra. A. Pushkin	965	6777.87	5816.12	7.02	6.03
Blessed. A. Pushkin	931	6543.76	5598.91	7.03	6.01
Horse riders. A. Pushkin	823	5704.28	4854.29	7.0	5.9
Winter. A. Pushkin	771	5445.71	4685.22	7.06	6.08
Parting A. Pushkin	574	3728.03	3170.89	6.5	5.5
Autumn evening. A. Pushkin	451	2963.9	2555.87	6.6	5.7
Winter morning. A. Pushkin	404	2438.52	2036.10	6.04	5.04
"V tom krayu…" («In that land …»). S. Yesenin	366	2238.9	1899.9	6.12	5.2
"Za tyumnoi pryadiyu…" («Behind the dark……»). S. Yesenin	306	1813.86	1549.2	5.93	5.06
"Krai ti moi.." («My land…»). S. Yesenin	199	1153.32	966.6	5.8	4.9
"The tree in blossom…" («Fall of the berries …»). S. Yesenin	185	1007.31	842.7	5.45	4.56
"Mi odni, iz sada…" («We are alone,	527	3414.11	2910.26	6.48	5.5

from the garden …»). A. Fet					
Village. A. Fet	370	2234.81	1868.49	6.04	5.05
"Kogda mechtatelno…" «When dreaming …». A. Fet	352	2145.41	1798.24	6.1	5.11
"Ulibka tomitelnoi…" («Smiling ……». A. Fet	277	1636.48	1396.6	5.91	5.04

Hs - amount of information in the text;

H - (unconditional) entropy or amount of information in one consonance; is calculated assuming that the statistical independence of consonances is calculated by the division of *Hs* by *Z*;

G - the depth of the layout of the entire consonance of the text; is measured in bits, as well as entropy;

g - the average remoteness of separate consonance in the text; is calculated by the division of *G* by *Z*;

It is obvious that values *Hs* and *G* increase when the length of the poem increases as well. The noticeable increase of *Hs* values above the *G* values and *H* values above those obtained for 'g' is explained by the following:

- firstly, by the model of stochastic source of consonance. Because of the statistical dependence on consonances, *Hs* and *H* values are lower;

- secondly, the *G* and *g* values are determined by the degree of regularity of the orders of consonances; they have the maximum values $\boldsymbol{G_{max} \approx Hs}$ and $\boldsymbol{g_{max} \approx H}$ for a trivial pseudo-text with a regular order of consonances.

As the indicated characteristics of entropy and average remoteness of a consonance depend less on the length of a poem, it is possible to use them for a comparison of different works of one author, and also works of different poets. Based on this fact, it is possible to assume that the higher the value of *g*, the more regular the rhymes of the poem. According to Table 3 the entropy and average remoteness of a consonance in poems of the great Russian poet A.S. Pushkin are higher than the respective values in the verses by Yesenin and Fet.

Besides, it is assumed that with the help of such characteristics like the depth of layout of consonances and their average remoteness in the text it is possible to do better editing and control of safety of the text content, as they, in contrast to frequency and entropy characteristics, take into account not only the content of the text but also the arrangement of elements in it.

To verify the last assumption a special experiment was conducted. Several poems (4 by Yesenin and 4 by Fet) were tested in the following way: the length of the text and the element order of the poem consonances were not changed, one or two lines of the verses were interchanged in a way that

- the rhyme was not infringed and the sense was maintained (1st deformation);
- the sense was saved, but the rhyme was infringed (2nd deformation);
- the sense was not saved and the rhyme was infringed (3rd deformation).

Table 4 below presents the results of entropic and numerical characteristics of an order of elements for several poems and their deformations under the number of variant. For two of the poems we present more detailed characteristics: rank distribution of the depths of consonance layouts (original and deformations).

Table 4

The title of a poem and its author	**Z**	**Hs**	**G**	**H**	**G**
«Siplet cheryomukha snegom …». ("The tree in blossom…..") S. Yesenin	185	1007	842.7	5.45	4.56
1st deformation	185	1007	850.15	5.45	4.60
2nd deformation	185	1007	840.36	5.45	4.543
3rd deformation	185	1007	840.01	5.45	4.541
«Krai ti moi zabroshenniy… ». ("This is my land…") S. Yesenin	199	1153	966.63	5.8	4.86
1st deformation	199	1153	966.51	5.8	4.86
2nd deformation	199	1153	976.97	5.8	4.91
3rd deformation	199	1153	969.8	5.8	4.87
«V tom krayu gde zheltaya krapiva…» ("In the land where the grass is green .. .") S. Yesenin	366	2239	1899.9	6.12	5.191
1st deformation	366	2239	1900.3	6.12	5.192
2nd deformation	366	2239	1900.8	6.12	5.193
3rd deformation	366	2239	1907	6.12	5.21
«Za tyomniy pryadyu perelesits…....». ("Behind the dark woods….") S. Yesenin	306	1814	1549.2	5.9	5.06
1st deformation	306	1814	1554.2	5.9	5.08
2nd deformation	306	1814	1555.5	5.0	5.08
«Ulibka tomitelnoy skuki...». ("The smile of a boredom ….") A. Fet	277	1636	1396.6	5.9	5.04
1st deformation	277	1636	1390.3	5.9	5.02
2nd deformation	277	1636	1401.9	5.9	5.06
3rd deformation	277	1636	1400.7	5.9	5.05
«Mi odni; is sada v styokla okon...». ("We are alone, from the garden into the windows…" A. Fet	527	3412	2910.3	6.47	5.522
1st deformation	527	3412	2905.7	6.47	5.15
2nd deformation	527	3412	2911.4	6.47	5.524
3rd deformation	527	3412	2903.8	6.47	5.510
«Kogda mechtatelno ya predan tishine ...». ("When I am dreaming in silence…. ") A. Fet	352	2145	1798.2	6.1	5.109
1st deformation	352	2145	1799.9	6.1	5.114
2nd deformation	352	2145	1796.5	6.1	5.104
3rd deformation	352	2145	1795.2	6.1	5.100
«Village». A. Fet	370	2235	1868.5	6.04	5.050
1st deformation	370	2235	1869.1	6.04	5.052
2nd deformation	370	2235	1871.0	6.04	5.057
3rd deformation	370	2235	1863.3	6.04	5.036

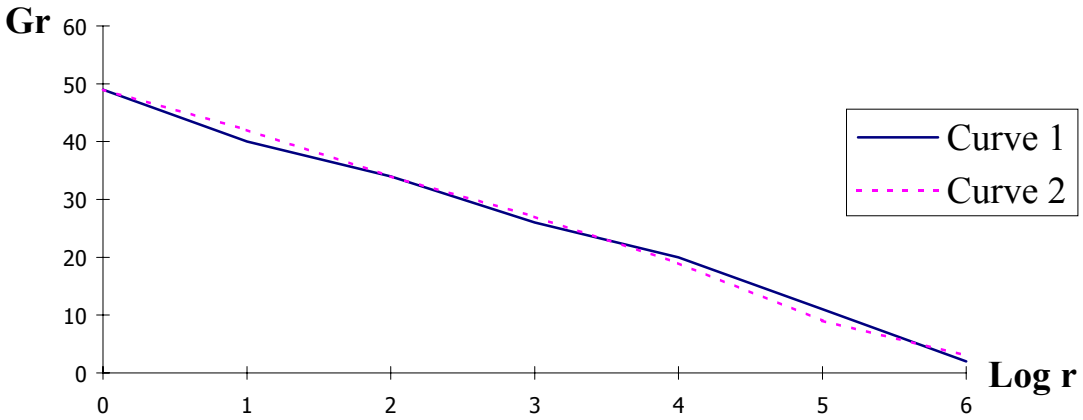


Fig. 2. Rank distribution of depths of consonance layout of invariable strings of the poem «Siplet cheryomukha snegom …». ("The tree in blossom….."} by S. Yesenin and its deformations of the third variant.

Gr - depth of consonance layouts of rank r in an invariable string;
Log r - binary log of a rank (see Fig. 1);
Curve 1 - rank distribution of depths of consonance layouts of invariable strings for the original text;
Curve 2 - the same for deformation of the text according to the third variant.

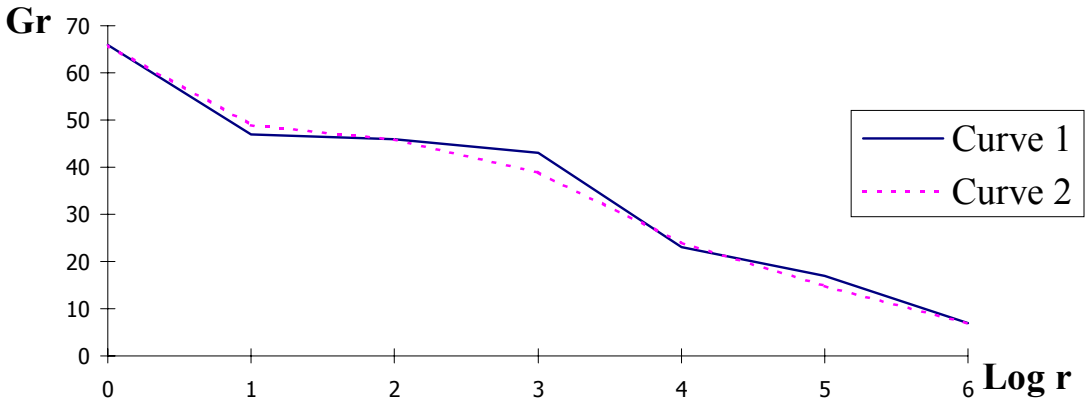


Fig. 3. Rank distribution of depths of consonance layouts of invariable strings of the poem «Ulibka tomitelnoy skuki...». ("The smile of a boredom ….") by A. Fet.

Gr, Log r, Curves 1 and 2 - see fig. 2.

During the research we once again were convinced that "the best is the enemy of the good". All interchanges of lines/verses, which were directed towards the keeping of meaning and rhyme (according to variant 1), have been successful, if they were evaluated by the comparison of the originals and deformations and also the appropriate numerical characteristics of their order of elements. Only in two cases those values appear to be a little lower than in the originals (A. Fet. «Ulibka tomitelnoy skuki...». ("The smile of a boredom ….") and «Mi odni; is sada v styokla okon...». ("We are alone, from the garden into the windows…"). Subjectively, those deformations of poems are hardly perceptible. However, only for two interchanges aimed at the breach of a rhyme with the meaning being being kept (according to

variant 2) the expected decrease of the characteristics («Siplet cheryomukha snegom …». ("The tree in blossom….."} by S. Yesenin, «Kogda mechtatelno ya predan tishine ...». ("When I am dreaming in silence…. ") A. Fet.) was obtained. In other cases those characteristics increased in comparison with the original.

Three interchanges of lines aiming at the distortion of meaning and breach of rhyme (according to variant 3) did not show the expected results, but also resulted in the increase of these characteristics «Krai ti moi zabroshenniy… ». ("This is my land…") by S. Yesenin, «V tom krayu gde zheltaya krapiva…» ("In the land where the grass is green .. .") by S. Yesenin, «Ulibka tomitelnoy skuki...». ("The smile of a boredom ….") by A. Fet). For four poems those characteristics in the given variant of deformation have decreased as predicted. Small samplings of poems do not allow substantial conclusions concerning the behavior of numerical characteristics during deformations of element orders in poems according to the second and third variants. Besides, it is necessary to consider that the simple increase of numerical characteristics of element order results in the trivial case of completely regular sounding and does not reflect other properties of poems.

Thus, for each poem it was possible to select its own vocabulary of consonances. This fact has two aspects. On the one hand, the number of such vocabularies, in our opinion, constitutes the element-acoustic basis of the Russian poetic language. This is proved by comparison. The sets of consonances of different poems are hardly overlapped, especially in the area of the most frequent elements (vocabularies of the processed poems are not presented here due to the limited size of the article). On the other hand, the alphabet of consonance of the poetic text probably creates a basis for the composition of a concrete poem as a completed piece of art (according to Orlov's criterion).

The proven fact obtained in the conducted experiment on the deformation of order of elements in poems without changing their component construction is the fixed changes of the numerical characteristics of an element order, when frequency and entropy characteristics are not changed.

With the method represented above being applied ten pieces of classical music were studied as well. It is necessary to mention that all those works are either homophonic (with the distinct main voice) or are written especially for single voice musical instruments, as for this class of musical note texts there exists a good and tested procedure of segmentation into elementary motives (F-motives) developed by M. Boroda.

Table 5 includes the frequency characteristics of processed texts. The works are ordered in compliance with the increase of their overall length measured by the number of F-motives.

The table shows that the majority of pieces of music is subject to the Zipf-Madelbrot law. Deviations of the actual alphabet from the designed one do not exceed 20 %. For most of the pieces of music theoretical and actual curves for rank allocation coincide with a large degree of large accuracy. Fig. 4 shows graphs of rank distribution for the piece of music of Saint-Saëns's "Introduction and Rondo-Capricciozo" (in a logarithmic scale).

Table 5

Author and Title	**Z**	**F_1**	**p_1**	**v**	**v_T**	**δ_V,%**	**δ_{cp},%**
Shostakovich, Dance.	92	15	0.163	34	32.7	3.8	18
Scarlatti, Sonata.№1.	257	22	0.086	93	80.4	13.5	7,8
Scarlatti, Sonata №67.	396	50	0.126	91	100.2	10.1	19
Bach, Prelude and Fugue №25.	525	45	0.086	165	135.9	17.6	9,3
Hindemith, Sonata №2.	910	33	0.036	316	253.3	19.8	4,7
Hindemith, Sonata №1.	1267	63	0.05	364	301.9	17.0	11

Saint-Saëns, Rondo and…	1154	88	0.076	242	255.8	5.7	15
Chopin, Ballad №1.	1145	82	0.072	228	257.7	13.0	21
Levitin, Sonata.	1250	78	0.062	375	284.2	24.2	7,6
Chopin, Ballad №2.	1591	99	0.062	325	344.5	5.9	14

Z-length of the piece of music (see below for its deformation and fragment), defined by the number of F-motives;

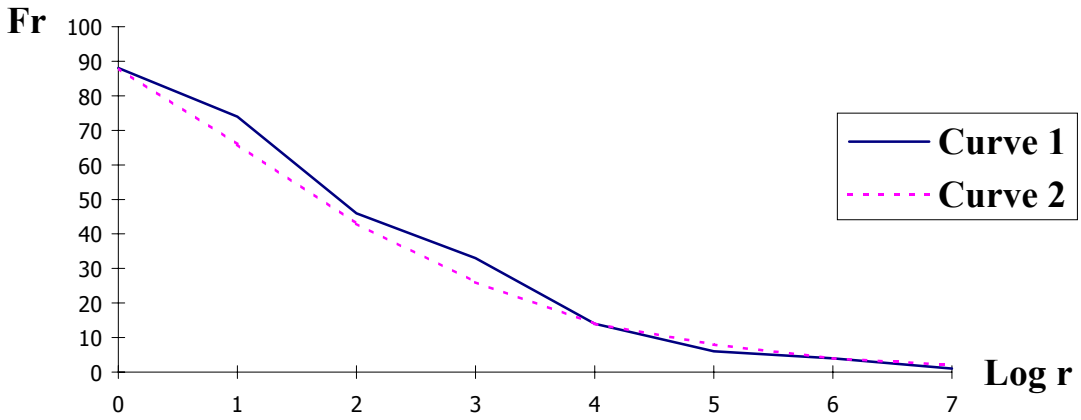


Fig. 4. Rank distribution of the number of occurrences of F-motives in Saint-Saëns's "Introduction and Rondo-Capricciozo".

Curve 1 - actual frequency - rank distribution;
Curve 2 - calculated distribution (according to the Zipf-Mandelbrot law).

Thus, M.G. Boroda comes to the conclusion that the F-motive construction of the text does not depend on the style of the piece of music and is only determined by its length and frequency of the most frequent F-motive.

The picture changes essentially, if a part of the piece of music is taken instead of the whole. Table 6 shows that the deviations of the theoretical prognosis from the actual "F-motive " stock are much higher than in case of single-piece texts, and in some cases this value is almost 100 %.

Table 6

Composer and title	**Z**	**v**	**v_T**	**δv,%**
Scarlatti, Selections from sonatas	800	93	180.8	94.4
Bach, Prelude №25	171	50	58.8	17.6
Bach, Fugue №25	354	149	99.1	33.4
Hindemith, Sonata №1, part 1	243	96	86.5	9.9
Hindemith, Sonata №1, part 5	393	61	101.4	66.2
Livitin, Sonata, part 1.	446	160	137.1	14.3
Saint-Saëns "Introduction…" final part.	498	95	121.1	27.4

But there also are exceptions, e.g. the music by Hindemith and Levitin, some parts of which are also subject to the Zipf-Mandelbrot law. Fig. 1 shows graphics of actual and calculated frequency rank distribution for the 1st part of Sonata №1 by Hindemith. It is seen

that those distributions coincide, as well as actual and calculated alphabets. This allows to assume that the parts selected by the author can be independent, finished pieces of music.

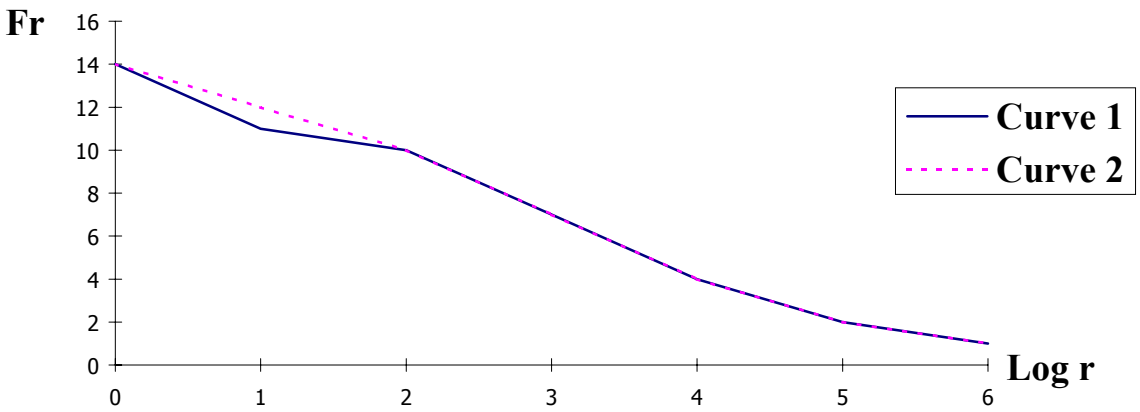


Fig. 5. Rank allocations of numbers of occurrences of F-motives in the 1st part of Sonata №1 by Hindemith.

Curve 1 - actual frequency rank distribution;
Curve 2 - calculated distribution (according to the Zipf-Mandelbrot law).

Thus, the conclusions by Orlov and Boroda that the Zipf-Mandelbrot law is the law of integrity of a separate piece of art are proved again.

Table 7 presents the results of the processing of the same ten pieces in the form of entropy and numerical characteristics of element order.

Table 7

Author and piece of music	Z	V	Hs	G	Hs-G
Shostakovich, Dance.	92	34	400	272	128
Scarlatti, Sonata №1.	257	93	1527	1060	467
Scarlatti, Sonata №67.	396	91	2132	1463	669
Bach, Prelude and fugue №25.	525	165	3349	2571	777
Hindemith, Sonata №2.	910	316	6865	4744	2121
Hindemith, Sonata №1.	1267	364	9185	5985	3200
Saint-Saëns, "Introduction…"	1154	242	7535	5153	2381
Chopin, Ballad №1.	1145	228	7309	4909	2399
Levitin, Sonata.	1250	375	9420	6782	2637
Chopin, Sonata №2.	1591	325	11137	7440	3697

Hs - amount of information in the text which is determined by assuming the statistical independence of F-motives;
G - depth of layout of F-motives of invariable strings of a piece of art.

It is easy to notice that the value Hs tends to grow with the increased length of a piece of art. Besides, during the more detailed study, we found out that if we take some pieces of art of similar length, the amount of information in the text depends on the richness of its "F-motive stock".

For example, Sonata № 1 by Hindemith is a little bit longer than Levitin's Sonata, but it has less alphabet and, accordingly, a lower amount of information as well. However, this characteristic is not enough, as it only takes frequency F-motive construction of the musical text into consideration. The value indicated in the next table column, the depth of layout of G, except for the frequency of occurrences of F-motives, characterizes their position in the text. Our following conclusion has been confirmed: this value is always lower than the amount of information, and the difference between them increases during the increase of the length of a piece of music. The noticeable exceeding of Hs values above the value of G and H above C in all investigated pieces of music is explained by the same two factors, which are indicated, should someone want to analyze the results of the studies. It might be desirable to mark one more interesting moment. If in a work arbitrarily selected fragments of identical size (such that the length of the text and its structure have not varied) are swapped, the depth of layout of F-motives varies. For this, the frequency characteristics remain constant (see Table 8).

Table 8

№		Z	v	Hs	G	Hs-G
1.	Saint-Saëns "Introduction…."	1154	242	7535	5153	2382
2.	Two randomly selected fragments with the length of 10% of the total length of the text have been inter-replaced.	1154	242	7535	5152	2383
3.	Two randomly selected fragments with the length of 25% of the total length of the text having been rearranged.	1154	242	7535	5147	2388
4.	The text is divided into two halves and these halves have been rearranged.	1154	242	7535	5085	2450

Table 9 contains the average characteristics: entropy, relative remoteness and difference between them.

Table 9

Author and Title	Z	v	H	g	H-g
Shostakovich, Dance.	92	34	4.35	2.95	1.4
Scarlatti, Sonata №1	257	93	5.9	4.1	1.8
Scarlatti, Sonata.№67	396	91	5.4	3.7	2.7
Bach, Prelude and fugue №25	525	165	6.4	4.9	1.5
Hindemith, Sonata №2	910	316	7.5	5.2	2.3
Hindemith, Sonata №1	1267	364	7.2	4.7	2.5
Saint-Saëns "Introduction….."	1154	242	6.5	4.5	2.0
Chopin, Ballad №1	1145	228	6.4	4.2	2.2
Levitin, Sonata	1250	375	7.5	5.2	1.7
Chopin, Sonata №2	1591	325	7.0	4.6	2.4

It may be noted that above the indicated characteristics are described much less and depend on the length of a piece of music. The difference of these characteristics for pieces of music, which are similar in length, is presumably determined by a genre, and the highest values of entropy and mean remoteness of F-motives occur in the music of Hindemith and Levitin. According to music experts this music is characterized by a more complex construction and less repeatability. The music is more serious and more informative, it conveys more information to think about, but, at the same time, is less predictable, keeps the greater tension during listening, and, certainly, "is acquired" with more difficulty. Usually, in such pieces, the size is more complex, it is much more difficult to detect any melody lines while listening, reprises are practically not used. At the same time it is possible to consider such compositions like "Dance" by Shostakovich, Ballad №1 by Chopin or Sonatas by Scarlatti to be popular, as they are more oriented to the broad audience, which, certainly, does not reduce the greatness of this music. The motives of these compositions are easily remembered, as in them the found soundings are used once, frequently there are reprises.

The interesting picture is observed during the deformations of compositions: when there is a simple rearrangement of the randomly selected fragments of identical length, the mean remoteness of F-motives of the composition varies, and it matters where the fragments are taken. The entropy remains the same, as the F-motive composition of the musical note text does not change. But if we simply delete some of its fragments, it is possible to see that the entropy begins to decline considerably. (Table 10).

Table 10

№		Z	v	H	g	H-g
1.	Saint-Saëns. Introduction.	1154	242	6.5	4.46	2.04
2.	Units with the total length of 10% of the whole text, exchange	1102	242	6.5	4.44	2.06
3.	25%, exchange	1154	242	6.5	4.38	2.12
4.	50%, exchange	1154	242	6.5	4.41	2.09
5.	10%, removal	1039	222	6.44	4.37	2.07
6.	25%, removal	874	165	5.98	4.1	2.88
7.	50%, removal	577	125	5.8	3.7	2.1

So, the above mentioned research results allow the assumption that the characteristics of an element order - the depth of layout and mean remoteness - are sensitive to the interlocation of F-motives in the text. In this connection the question is whether it is possible to use the given numerical characteristics to improve sounding of a piece of music by some deformation of its element order. Such an attempt has been undertaken. Some of its results are listed in Table 11 based on the example of Sonata №1 by Hindemith. It should be emphasized that if second and third parts of the composition are swapped, the depth of layout and the mean remoteness of F-motives increase. When playing this sonata in such a sequence, the composition as a whole sounds slightly different in case of a subjective listening by a not very experienced researcher, but it is not worse than the original, and, probably, it is even a little better. For comparison Fig. 6 shows the characteristics of the original and the deformation by the solid line (for the case that the fifth (last) part, which, by the way, is the least informative of all the parts, is played right at the beginning of sonata, before the 1st part. The change of sounding, which does not mean an improvement, is obvious even to a musical non-expert. It also is confirmed by both, the decrease of mean remoteness and the depth of layout of F-

motives resp. When two small fragments being logically and melodically completed are rearranged in the 1st part of the Sonata such that the melody as a whole has – at least - not worsened, we just obtain some fewer characteristics of an element order in comparison with the original. Such an analysis is impossible if only frequency characteristics of the structure of the musical text are used.

Table 11

№	Title	Z	v	H	g	G
1.	Hindemith, Sonata №1.	1267	364	7.25	4.724	5985
2.	2nd and 3rd parts are presented	1267	364	7.25	4.726	5988
3.	The 5th part is arranged before the first one.	1267	364	7.25	4.605	5835
4.	The 1st part comprises some small units.	1267	364	7.25	4.722	5983

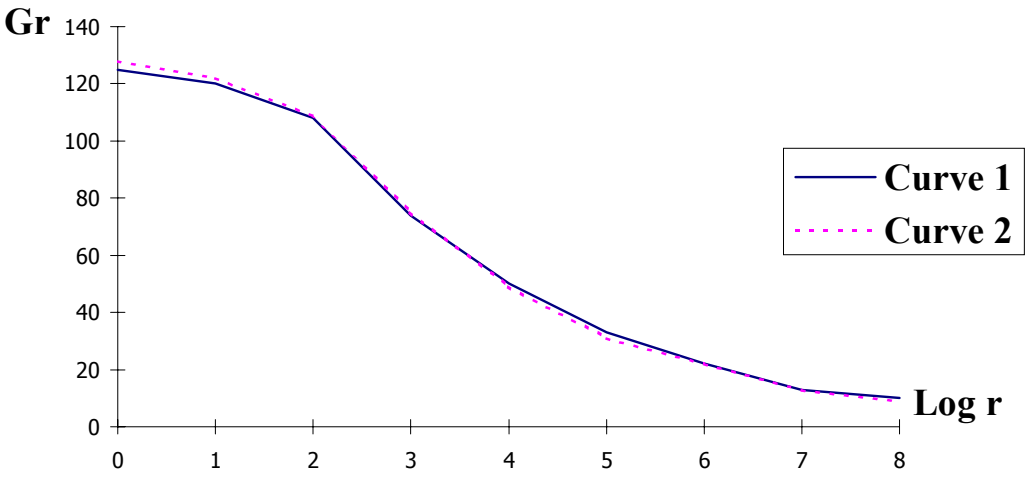


Fig. 6. Rank distribution: depths of layout of F-motives of invariable strings in Sonata №1 by Hindemith (curve 1) and its deformation (curve 2).

Except for the analysis of pieces of music, the characteristics of an element order can be used as criteria for the reliability of the musical information during transmission and saving.

Thus, we could show that a more detailed research of the architecture and properties of a piece of music is possible with the help of numerical characteristics of an element order, which take into account the rearrangement of F-motives in a musical text.

The quantitative comparisons of the constructions of poems (consisting of consonances) and musical texts (consisting of F-motives) provide for a better foundation of our assumption as to the existence and quantum nature of audio objects.

In our opinion, the models and characteristics of an element order in a sign sequence will help establish the direction of research of the nature of thinking; such research will be based on the study of the quantitative characteristics of the construction of the element order, characters and words in texts and notes, with one of its founders being was the American linguist George Zipf.

Attachment 1.

Segmentation of «Petersburg Verses» by O. Mandelshtam into consonances

На т жел ть и з н ойп р а вь и тье ль ст вье н ны хз д а ньи й
К р уж ы л а сь до л г а м у т на й а мье тье ль
И п р а в а вье т а пьать с а дь итсь а ф с а ньи
Шы р о кьи м жес т а м з а п а хну ф шы нье ль

Зьи м у й у т п а р а х о ды на п рьипь о кье
З а жг ло сь к а й у т ы т о л ст а йе с тье к ло
Чу до вьищ на к а к бр а нье н о сье цф до кье
Р а сьсь и й а а т ды х а йе т ть а жел о

А на т нье в ойп а с о ль ст в а п а л у мьир а
А д мьир а л тье й ст в а с о лнц еть и шы на
И г а су д а р ст в а жос т к а й а п а р фьир а
К а к в л а сь а ньи цаг р уб а й а бьед на

Тьаш к ааб у з а сье вье р на в а с н об а
А нье гь и на ст а рь и н на й а т а ск а
На п ло щадь и сье на т а в а л су г р об а
Ды м о к к а ст р аих а л а до кш т ы к а

Че р п а льи в о д у й а льи кь иич а й кьи
М а р с кьи йе п а сье щаль и ск л а т пье нь кьи
Г дьеп р а д а в а й а збь и тье ньи льи с а й кьи
Льи шопь е р ны йе бр о дь а т м уж ы кьи

Ль еть итф т у м а н м а т о р а ф вье рье ньи ц а
С а м а ль убь и в ы й ск р о м ны й пье шех о т
Чу д а к йе вгь е ньи й бьед на сть и ст ыдь итсь а
Бье н зьи н в ды х а йе т и су дь бук ль а ньо т

Vocabularies of consonances obtained during the segmentation of «Petersburg Verses» written by O. Mandelshtam: pauses between the lines not being taken into account (1st and 3rd columns) and pauses being considered.

А 84	А..... 83	С 9	С..... 9
Т 21	Т 21	СТ..... 8	Н..... 8
Р..... 18	Р 18	У 8	СТ 8
И..... 16	Й..... 16	Н 7	У..... 8
Й 16	И..... 14	Д 6	ЛЬ 7
К 13	О..... 14	ЙЕ..... 6	Д..... 6
О 13	К..... 13	КЬИ 6	ЙЕ 6
В 12	НА..... 13	НЬИ..... 6	КЬИ..... 6
НА 12	В..... 12	ЛЬ..... 5	НЬИ 6
М 11	М..... 11	ЛЬИ..... 5	ТЬЕ 5
П..... 10	Л..... 9	ТЬЕ..... 5	Ы 5
Л 9	П..... 9	Ы 5	ВЬЕ 4

ВЬЕ 4
ДО 4
ДЫ..... 4
НЬЕ 4
СЬЕ 4
Х 4
Г..... 3
ДЬ..... 3
Е 3
З..... 3
ЛО 3
НЫ 3
СК..... 3
СУ 3
СЬ..... 3
Ф..... 3
ШЫ 3
БР 2
БЬЕД..... 2
ЕТЬ..... 2
ЖЕЛ 2
ИТСЬ..... 2
МЬИР..... 2
ОБ..... 2
ОЙП 2
ПЬЕ 2
ТЬ 2
УЖ 2
ЧУ 2
ААБ..... 1
БУК 1
БЬЕ..... 1
ВГЬ..... 1
ВЬ..... 1
ВЬИЩ..... 1
ГДЬ..... 1
ГЬ 1
ЖГ 1
Г 4
ДО..... 4
ДЫ 4
З..... 4
ЛЬИ..... 4
НЬЕ..... 4
СЬЕ 4
ДЬ 3
ЛО..... 3
НЫ 3
СК 3
СУ 3
СЬ..... 3
Ф 3
Х..... 3
ШЫ 3
БР 2
БЬЕД..... 2
Е 2
ЕТЬ 2
ЖЕЛ..... 2
ЗЬИ 2
ИТСЬ 2
КЬЕ 2
МЬИР..... 2
ОБ 2
ОЙП..... 2
ПЬЕ..... 2
ТЬ..... 2
УЖ 2
ЧУ 2
ААБ 1
АИХ..... 1
БУК..... 1
БЬЕ 1
ВГЬ 1
ВЬ..... 1
ВЬИЩ..... 1
ЖЕС 1
ЖОС..... 1
ЗБЬ 1
ИИЧ 1
ИТФ 1
КШ 1
КЬ..... 1
КЬЕ 1
КЬЕЗ 1
ЛНЦ 1
ЛЬЗЬ..... 1
МЬЕ..... 1
НЗЬ..... 1
НЬ..... 1
НЬО..... 1
ОАН 1
ПЬАТЬ 1
РЬ 1
РЬЕ..... 1
РЬИПЬ 1
СЬСЬ..... 1
ТЬАШ 1
УБ..... 1
УБЬ 1
ФЬИР 1
ХЗ..... 1
ХНУ 1
Ц..... 1
ЦАГ..... 1
ЦФ..... 1
ЧЕ..... 1
ШЕХ 1
ШОПЬ..... 1
ЩАДЬ..... 1
ЩАЛЬ 1
ЫДЬ 1
ГЬ..... 1
ДЬЕП 1
ЖГ..... 1
ЖЕС..... 1
ЖОС 1
ЗБЬ..... 1
ИИЧ 1
ИТФ..... 1
КШ..... 1
КЬ..... 1
ЛНЦ..... 1
МЬЕ 1
НЬ 1
НЬО 1
ПЬАТЬ..... 1
РЬ..... 1
РЬЕ..... 1
РЬИПЬ..... 1
СТЬ..... 1
СЬСЬ..... 1
ТЬАШ..... 1
УБ 1
УБЬ..... 1
ФЬИР..... 1
ХЗ..... 1
ХНУ..... 1
Ц..... 1
ЦАГ 1
ЦФ 1
ЧЕ 1
ШЕХ 1
ШОПЬ 1
ЩАДЬ..... 1
ЩАЛЬ..... 1
ЫДЬ 1

References

Bachtin, M.M. (1979). *Estetika slovesnogo tvorčestva*. Moskva: Iskusstvo.

Boroda, M.G. (1978). Častotnye struktury muzikalnych tekstov. In: *Izmerenie i prognoz v kulture.* Moskva.

Borodovskiy, M.Yu., Pevzner, P.A. (1990). Statističeskie metody analiza genetičeskich tekstov. In: *Kompjuternyj analiz genetičeskich tekstov, 33-80*. Moskva: Nauka

Grammatika russkogo jazyka, t.1. Fonetika i morfologija, 1960, s. 47-100. Moskva: Izd. AN SSSR

Gumenyuk, **A.S.** (2000). O formalizme, izmerenii i izčislenii stroenij cepej soobščenij. In: *Materialy Meždunarodnoj naučno-techničeskoj konferencii "Informacionnye sistemy i technologii", t.3, 499-502*. Novosibirsk: Izdatel´stvo NGTU.

Gumenyuk, A.S. (2001). O podchode k formalizacii celostnogo vosprijatija teksta. In: *Kvantativnaya lingvistika i semantika: Sbornik naučnych trudov, Vypusk 3: 3-12.* Novosibirsk: Izdatel´stvo NGPU.

Gumenyuk, A.S., Kostyshin, A.S. (1998). O kompjuternom analize tekstov i odnom formalizme segmentacii stichotvorenij russkoj literatury na sočetanija fonem. In: *Kvantativnaja lingvistika i semantika: Sbornik naučnych trudov, Vip. 1: 3-17.* Novosibirsk: Izdatel´stvo NGPU.

Gumenyuk, A.S., Kostyshin, A.S., Simonova, S.V. (2000). Issledovanie stroenij lingvističeskich tekstov na primere stichotvorenij russkogo jazyka i muzykalnich proizvedenij klassiki. In: *Kvantativnaya lingvistika i semantika: Sbornik naučnych trudov, Vypusk 2: 12-40.* Novosibirsk: Izdatel´stvo NGPU.

Kostyshin, A.S. (1998). Ob odnom algoritme segmentacii tekstov na strukturnye edinicy. In: *Kvantativnaja lingvistika i semantika: Sbornik naučnych trudov, Vypusk 1: 18-21.* Novosibirsk: Izdatel´stvo NGPU.

Leus, V.A. (1987). Čislennyj kriterij blizosti tekstov. In: *Vyčislitelnye sistemy. Vypusk 123: 61-83.* Novosibirsk. (Izdatel´stvo IM SO AN SSSR)

Mazur, M. (1974). *Kačestvennaya teorija informacii*. Moskva: Mir.

Orlov, U.K. (1980). Nevidimaja garmoniya. In: *Čislo i mysl. Vypusk 3: 70-106*. Moskva Znanie.

Glottometrics 3, 2002, 90-109 *To honor G.K. Zipf*

Speaker's information content: frequency-length correlation as partial correlation

Simone Andersen[1]

Abstract. A new variable is presented which is demonstrated to be of influence on the correlation between lengths and frequencies of words occurring in texts. The variable is called speaker's information content ("sic").

Keywords: Length, frequency, information, conspicuousness, speaker's information content

Information and conspicuousness

The relation between word lengths and frequencies as stated by Zipf (1935, 1949) is derived from Zipf's earlier investigations in phonology (1929, 1932). The basic construct in his phonological hypotheses is the "conspicuousness" of a sound – resp. a phoneme – which he showed to be inversely related to its relative frequency. The idea refers to principles of information theory emerging in those days (Hartley 1928; Shannon 1948; Wiener 1948). The amount of information $h(x_i)$ of an event x_i is defined as a function of its relative frequency:

(1) $h(x_i) = f(p(x_i)) = -\operatorname{ld} p(x_i).$

During a long time, when early information theory was applied to human information processing (Birkhoff 1933; Attneave 1955; Berlyne 1958, 1960; Zuckerman 1979; 1984), the concept of information (as of "complexity") was equally used to characterize both, the degree of size, strength, extension and the degree of distinctness, unexpectedness, novelty of an investigation object - two groups of variables which were treated as equivalents and gathered in some idea of intensity. (The information theoretical quantification of "prägnanz" (Attneave 1954) can be taken for example.)

Correspondingly, Zipf developed his concept of conspicuousness which merges the components of both groups. Indeed, intensity on the sounds or phoneme level can be considered equally the size of production effort (size of articulation effort, number and kind of articulatory features, amount of stress, difficulty of articulation) as well as distinctivity facilitating perception (salience, discernibility, ease of acoustic perception and identification). So Zipf succeeded in showing that lower frequency of a sound, and therefore higher amount of information, is positively related to greater "conspicuousness". Well-founded by Zipf's model of the balancing forces of unification and diversification (speaker's and hearer's "interests"; Köhler 1987; Altmann, Köhler 1995) the idea means an important step in quantitative linguistics (Altmann 1993; Altmann, Köhler 1995).

Trying to expand his considerations by applying the laws derived from phonological

[1] Address correspondence to: Simone Andersen, Loogestieg 19, D-20249 Hamburg, Germany.
E-mail: AndersenSC @aol.com

investigation to units at lexical level, however, some problems arise. Now Zipf considers the length of a word, measured by the number of syllables (sometimes number of phonemes) to be analogous to the conspicuousness of a sound.

To maintain his analogy transfer he needs a couple of implicit untested hypotheses on classification and information processing: statements concerning effort and difficulty, emphasis, salience and discrimination in understanding and producing words. There is room for doubts if the results of investigation of sounds can simply be transposed. It could e.g. be questioned if speakers' and hearers' interests really are opposed to another at word level. If unification means a small inventory - "one word for all cases" - or repeating the same all the time, it could be asked if this should be facilitation for the text producer. If diversification means a gigantic inventory - for every case its own word – it could be asked if this and/or greater length really means facilitation for the hearer, even if some writers with bad writing skills typically seem to believe it... (A more plausible continuation of Zipf's economic considerations towards the level of semantics: Köhler & Altmann 1993.)

The critical problem, however, concerns Zipf's translation of his construct of conspicuousness. Word length cannot be treated like intensity or "size" of a sound: especially, it cannot be replaced by or equalled to the variables mentioned above.

For example, the longer words are not necessarily those that are better identified and easily perceived, as readability research has shown. The shorter words are not necessarily less salient or surprising - or perhaps even harder to understand. On the contrary: many examples can be found where a longer word (e.g. two-syllable words like "ever", "any", "many") can be contrasted to a shorter word which is more unusual or unexpected ("whilst"), more specific ("quarks"), more salient or activating ("stop", "sex", "crime" – as demonstrated by advertising or newspapers); in an actual text, however, where the context decides very much on salience and expectancies, this is even more evident. Where do these problems come from?

In the recent decades, it has turned out that an important step is required for the proper application of information theory to human information processing: the distinction of the two independent constructs of intensity (strength, size, extension) and uncertainty (or novelty, surprise potential, unexpectedness, salience, "difference in sameness" (Berlyne 1960).). Only the latter bears the name of "information" in its mathematical sense. The development of Zuckerman's psychobiology of personality and his construct of sensation seeking (Zuckerman 1979, 1984, 1991), probably the highest developed work in this area, and the critical reception of his work (Wohlwill 1984) have proven this.

Therefore, an important part of Zipf's argumentation for the length-frequency correlation at word level is lacking: according to Zipf, conspicuousness of a word is determined by its frequency: this can only be in the meaning of unexpectedness or unusualness (the "real" information), while on the other hand, this in turn is just independent of size (or intensity).

Unquestionably, however, there is some correlation, even though not always perfect, shown by empirical data. Köhler (1986) and Hammerl (1991) found correlations between frequency classes and lengths with $D = 0.3276$, $D = 0.7633$ or $D = 0.8544$ (which means correlation coefficients of $r = \pm 0.5724$, $r = \pm 0.8737$ or even $r = \pm 0.9243$).

Apart from Zipf's own argumentation, there have been other explanations for the Zipfian word correlations which are very plausible (Guiraud 1963; Mandelbrot 1954). The aim of this paper is not to discuss the appropriate explanation. We rather want to take up Zipf's idea of looking for information content related to words in texts.

When the amount of information cannot simply be inferred from the length of a word as explained above, the question is where it is and how it is related to length and frequency. So, we are dealing here with three entities: length, frequency and information.

We want to show that and how information content of words affects their frequency-length correlation and why it must be taken into account.

Information content of words in texts

If we want to apply the relation cited above

$$h(x_i) = - \text{ld}\, p(x_i)$$

to human information processing, it is a non-trivial question, how p is determined.

If we look at language as a whole, the "a priori"- frequencies, as they are noted down in frequency lexica, can be entered for p. Thus, for every word an „a priori" information content would result (the majority of all words would yield an enormous h, because of their small relative frequencies).

The situation becomes very different when we look at words occurring in specific texts. "Information" can be determined in more than one way.

Of course, one could first think of calculating the uncertainty of the text as a whole: we consider the text as a system, and we determine the state of uncertainty of the system (text entropy) I.

The information $I(X)$ of a system X is

(2a) $$I(X) = \sum_{i=1}^{n} p(x_i) h(x_i)$$

or with $h(x_i) = - \text{ld}\, p(x_i)$

(2b) $$I(X) = - \sum_{i=1}^{n} p(x_i)\, ld\ p(x_i)\,.$$

Determining text entropy I, the $p(x_i)$ are just the relative frequencies at which the words occur in the text. The difference of the calculated text entropy from maximum text entropy (= state of equal probabilities) is the increase of knowledge we get by the complete text compared to the knowledge we would have if we only got a list of all words occurring without knowing if they occur once, twice or very often etc. It increases with a smaller number of different words (types) and more differing probabilities. Measuring text entropy, however, gives us no "information" on single words.

To examine the frequency-length correlation we need a measure for the amount of information assigned to each word (or word class) separately.

In the following, we will concentrate on the information content single words do have in a special text.

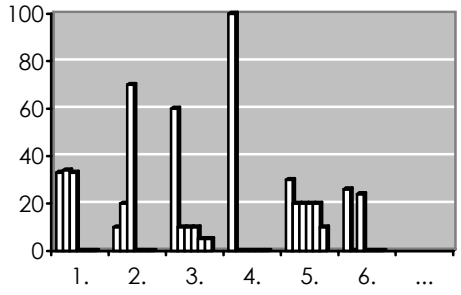
There are two ways of looking for this kind of information words have in texts; they will be called here: "speaker's information content" (*sic*) and "hearer's information content" (*hic*).

As we are going to show, for calculating *sic* and *hic*, p has to be determined in a different way than by the relative frequency at which it occurs in the text.

A word in a text can be thought of as a realization of a number of different alternative possibilities (see the figure below).

The number of alternatives is varying for different words, and strictly speaking the probabilities for the alternatives are distributed in varying ways (in detail: Andersen, 2002). In the following, we will only consider the number of alternatives, regardless of their differing probability distributions.

Thus for every word in a text the degree of uncertainty can be determined. Instead of "information" we now prefer to use here the classical concept of "uncertainty" U, according to Hartley, where the principles of freedom of choice and decision content are pointed out.



...nth word

Into $h(x_i) = -$ ld $p(x_i)$ we enter $p = 1/n$, (in Hartley's measure the alternatives are supposed to be equal), so the uncertainty U or $h(x_i)$ is $h(x_i) =$ ld n as a function of the number of alternatives n.

Now it has to be observed that *hic* ≠ *sic* for many of the words.

The Fitts-Garner controversy (Fitts et al. 1956; Garner 1962, 1970; Garner & Hake 1951) on information content, structure and the amount of redundancy in metric figures (in Coombs, Dawes, Tversky 1970) has shown that information in artefacts or objects constructed by humans has to be defined in two different ways, depending on two different ways for determining p – from the construction perspective (also Evans, 1967) and from the reception perspective. Attneave (1959; 1968) recognized that we are dealing here with two descriptive systems which are not necessarily congruent.

For our aims this means: information content of words in texts has to be determined for speaker and hearer separately.

An example: Suppose that in the sentence "Tomorrow he will be in Cologne" one word is deleted and replaced by a blank. Trying to fill in the blank, we are able to determine the uncertainty of the missing word. From the hearer's perspective, trying to anticipate what could be intended to be said, the word "will" would be very expectable if missing: there are only few alternatives conceivable ("may","might","must","could", "should"), its information content is relatively low. A lot of more possible words are conceivable replacing "tomorrow": for example "now", "then" etc., but also "anyway", "surely", and many other words, which means a far greater information content - for the hearer. "Cologne" at last stands in a "hearer's distribution" with a nearly infinite number of alternatives – presumed, there is no prior pragmatic knowledge –, its information is extremely high. There are methods in readability research using the different amounts of information with p being defined from hearer's view (Taylor 1953; Andersen 1985).

However, if p shall be defined from the speaker's distributions, other values will result. For the speaker, the conditions for determining p are not the possibilities of inferring the future message from incomplete data. His uncertainty h depends on the decision content of the specific position in the text. That means: the speaker's conditions for determining p are in his intentions of transmission, not in the incomplete knowledge by the received part of the message. Provided, he wants to transmit the message that a certain person will be in Cologne tomorrow, then there is no alternative for "tomorrow" – except for (perhaps) the name of the weekday. There is no alternative for "Cologne". Compared to the entire words in the sentence, the greatest number of alternatives (= uncertainty or decision content) for the speaker can be found for the words "he" (here the different possible names of the person are conceivable: e.g., Robert, Rob, Robbie, Bob, Bobby, etc.) and "be" (possible alternatives like "stay", "lodge", etc.).

Notabene: The p by which h is determined here, is not the p denoting the relative frequency in the total text. It is the p related to the possibilities a specific word has at a specific

position, here defined as either zero or 1/*n*, where n is not the total sum of words in the entire text, but the set of words conceivable at the specific position. What is neglected when correlating the lengths and the frequencies of words in real texts is the fact that there is not at all free choice out of all existing words at every moment for the text producer. Why it is of importance to regard the speaker's information content when looking at the correlation between words and their lengths will be demonstrated by the following example.

Imagine the following situation: in a seminar room a number of individuals are sitting; their names are differing in the number of syllables (e.g., Zipf, Benford, Csikszentmihályi etc). A teacher poses questions to the auditory. After every question some persons raise their hands, she calls one of them to give his/her answer. After a number of questions (e.g. 20 questions) the names called out by the teacher are counted and the distribution of their word lengths is determined. As we can easily imagine, we hardly can infer from our results any substantial statements on possible reasons for the observed frequencies (not even on preferred lengths). The data do not tell us "the whole truth".

The distribution of uttered name lengths depends on several other distributions: (a) on the "original" distribution of the persons' names sitting in the seminar room, (b) it also depends on the distributions of the participants putting their hands up after the particular questions. (c) At last, one can think of all those influences resulting from the situation in which the choice is made itself (e.g. "preference distributions", etc.).

Illustrating our lacking knowledge when counting the word lengths alone and ignoring the speaker's uncertainty: We do know the number of answers: 20 names were uttered which symbolizes the number of occurring words in a text. We know the distribution of their lengths. However, what we are not told is the distribution of the "original" name lengths of the students. Nor do we know the distributions of persons raising their hands after the particular questions; the teacher, however, is forced to choose just among them. The example was constructed to illustrate the important aspects in looking for speaker's information content: The teacher confronted with the number of persons raising their hands stands for the speaker at the moment, when he is going to say what he means to say. We locate the uncertainty situation just at the moment the decision is made. That means: the circumstances leading to this point (the distribution of name lengths being present in the room, the reasons determining over who raises his hands, and even the reasons for the number of persons raising their hands) are considered to be "conditions" and separated from the choice situation.

For the problem of word length frequencies this means that for determining *sic*, we are not concerned with the questions of how the speaker percepts the world and the necessity for producing a text, why the speaker decided to talk, why he decided to talk about this subject, this particular feature of it , why he decided to transmit just this message, and why he decided to plan just this structure of content. Thus, for the present measure, the processes developing what Köhler calls "requirement of application" (1987, 1990) have already been working. We consider them to be forces having led to this moment, forming the conditions. We now have to regard "the number of people raising their hands", that means we have to regard the uncertainty for the speaker at every moment or the amount of *sic*.

We believe that *sic* is a third variable, which is related to at least one of the variables frequency *f* and length *L* of a word, and which is of influence on the relationship between them; we have to examine what happens if we adjust for the effects of *sic* on the original frequency-length correlation. The effects can be supposed to be linear as well, so we use partial correlation.

Method

Partial correlation:
The partial correlation coefficient $r_{xy\cdot z}$ is the measure for the linear association r_{xy} between two variables x and y, from which the linear effect of a third variable z has been eliminated.

$$(3) \qquad r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1-r_{xz}^2}\sqrt{1-r_{yx}^2}} .$$

To demonstrate the method the following text taken from a German newspaper ("Hamburger Morgenpost", 24-4-2002; for complete text see appendix) was examined. The example is meant to show the main principles, not to collect empirical data.

First, we recorded three variables for every word:

its length L_i, 2) its frequency in the text f_i, 3) the number of alternative words a_i.

To determine a_i some preliminary definitions have to be made: The number of alternatives a_i includes the word that was chosen in the text, so if $a_i = 1$ is counted, it means „no other word possible than the one chosen". The decision content or uncertainty then is $h_i = \mathrm{ld}\ 1 = 0$.

The values of a_i were determined subjectively in this example, according to the author's intuitive competence of writing texts. To demonstrate the principle of the method this may be sufficient. (Anyway, it can be taken for evident that different words in a text differ considerably in the number of their possible alternatives.)

Possibly a_i can be determined objectively as well. A conceivable method would be: letting a greater number of subjects (native speakers) report which and how many other words are possible alternatives at the particular place in the text, provided that the transmitted content will stay exactly the same.

This method corresponds to the structuralistic contrast or opposition criterion in identifying phonemes. Two phones are realizations of one phoneme if they can be replaced by one another in a word without changing the meaning of the word. The phone corresponds to the single word or token. If two words (or tokens) can be replaced for one another in a particular text without changing its meaning according to speaker's intention, they are realizations of one "tokeme".

In order to simplify the counting method and to show the principle as clearly as possible here it was decided: each word has to be replaced always by one word only. So, every alternative consists of one word. Additionally it holds that the replacement has to be carried out with all things being equal, that means, provided the entire rest of the text remains unchanged. (Example: "They stayed *at* home". Only the word "*at*" is possible at the underlined position. It is not allowed to replace "*at*" by "*in their*", changing the sentence into "They stayed in their house", etc.)

On the other hand, the possibility of „no word" instead of the considered word– recorded as Ø - is counted and included in the number of alternatives a.

These constraints may represent an insufficiently realistic model of the writing or text production process. The aim here, however, is not a performance description, but the demonstration of counting *sic*. (Some refinings of the method will be described later.)

Results

First: Recording of words

We started by separately counting for every individual word (type): its length *L*, its frequency *f*, the number of alternatives *a*, and by *a* we calculated the speaker's information content *sic*. (If for one word *a* was not equal at all positions, the mean was determined.) (See Table 1.)

Table 1
Lengths, frequencies, number of alternatives
and *sic* of individual words

L	f	a	sic
1	1	1	0.0
3	1	16	4.0
1	2	3	1.6
1	1	2	1.0
1	1	1	0.0
...	…	…	…

etc. (for complete data see appendix)
$\sum$ words = 102
L = length, *f* = frequency,
a = number of alternatives

Then the correlation between length and frequency r_{fL} with and without controlling for *a* and for *sic* was determined (separately):

zero-order correlation: $r_{fL} = -0.2644$
controlling for *a*: $r_{fL \cdot a} = -0.2371$
controlling for *sic*: $r_{fL \cdot sic} = -0.2071$
zero-order correlations:
number of alternatives and length: $r_{aL} = 0.3079$
number of alternatives and frequency: $r_{af} = -0.1325$
sic and length: $r_{sicL} = 0.3124$
sic and *f*: $r_{sicf} = -0.2340$

The results are depicted in the Figures a.1 and a.2.

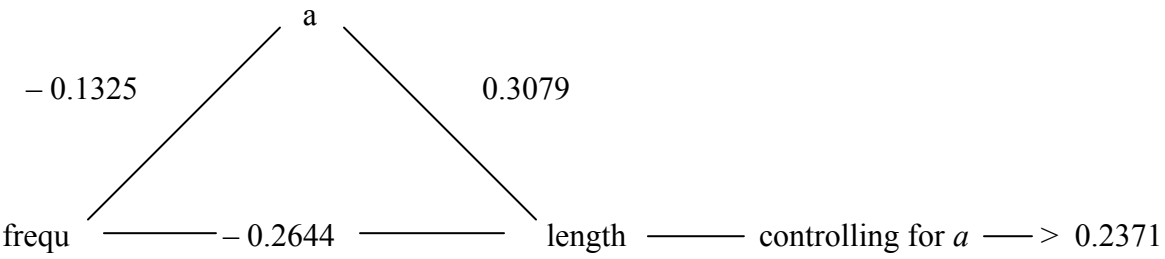


Figure a.1. Partial and zero-order correlations of frequency, length and number of alternatives

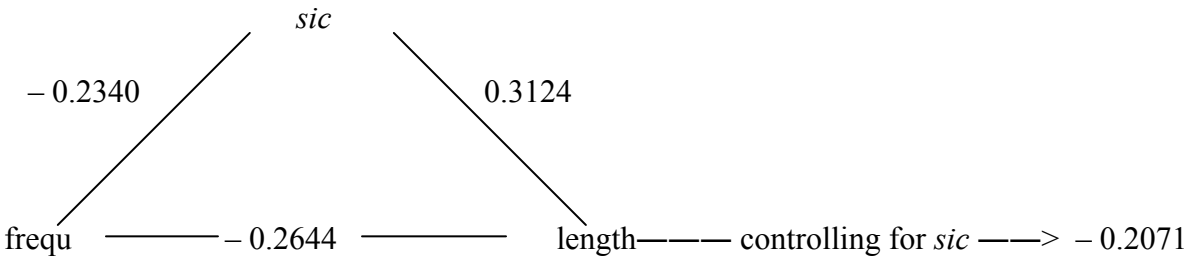


Figure a.2. Partial and zero-order correlations of frequency, length and sic

The data for single words already show: if we control for the number of alternatives *a*, thus eliminating its influence on the relationship between frequency and length, the correlation decreases. If *sic* is held constant, the correlation gets remarkably lower. An interpretation of this is that the number of alternatives – and with it *sic* – are partly responsible for the correlation r_{fl}.

The number of alternatives – and with it *sic* - influences the frequency (negatively). It also influences length in a positive way: if there are more alternatives, the speaker is able to choose a longer word. So the *fL*-correlation is partly due to *sic*.

Now we group the data according to their frequencies, like Zipf used to do.

Second: Data grouped according to their frequency (times of occurrence)

We recorded

f = frequency (times of occurrence), class size = number of types,
f x *cls* = number of cases or tokens (frequency *x* class size),
meanL = mean length of words in the class,
a = total number of alternatives for all cases (tokens) in the class,
mean a = *a*/number of cases = *a* / (*f* x *cls*),
sic (of *mean a*) was calculated for the entire class, by $h(x_i) = \mathrm{ld}(mean\ a(x_i))$

Table 2.1
Data grouped according to their frequencies

f	class size	f x cls	mean L	a	mean a	sic (of mean a)
1	88	88	2.26	315	3.58	1.84
2	11	22	1.27	58	2.64	1.4
3	1	3	1.00	4	1.33	0.41
4	0	0	0	-	-	-
5	2	10	1.00	13	1.30	0.38
6	1	6	1.00	7	1.17	0.23

We can see (Table 2.1) that the mean number of alternatives decreases when frequency increases.

The very frequently used words have fewer alternatives on average – perhaps this is why they are used more often?

Hapax legomena (*f* = 1) show more decision content, on average. The speaker has more freedom of choice. So they can tell us more about the text producer and his choice behavior or

his preferences or properties than the words of the other frequency classes, which are chosen less voluntarily.

Figures I.a - d show the number of types (class size), the mean length, the mean number of alternatives and the mean *sic* in dependence on frequency.

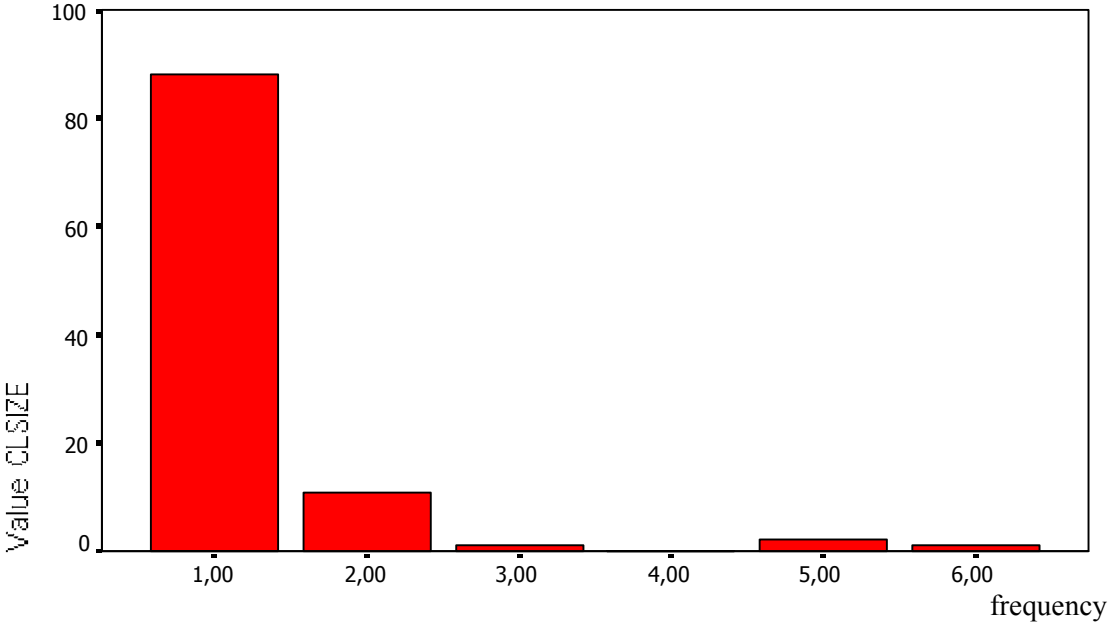


Fig. I.a. Frequency and class size

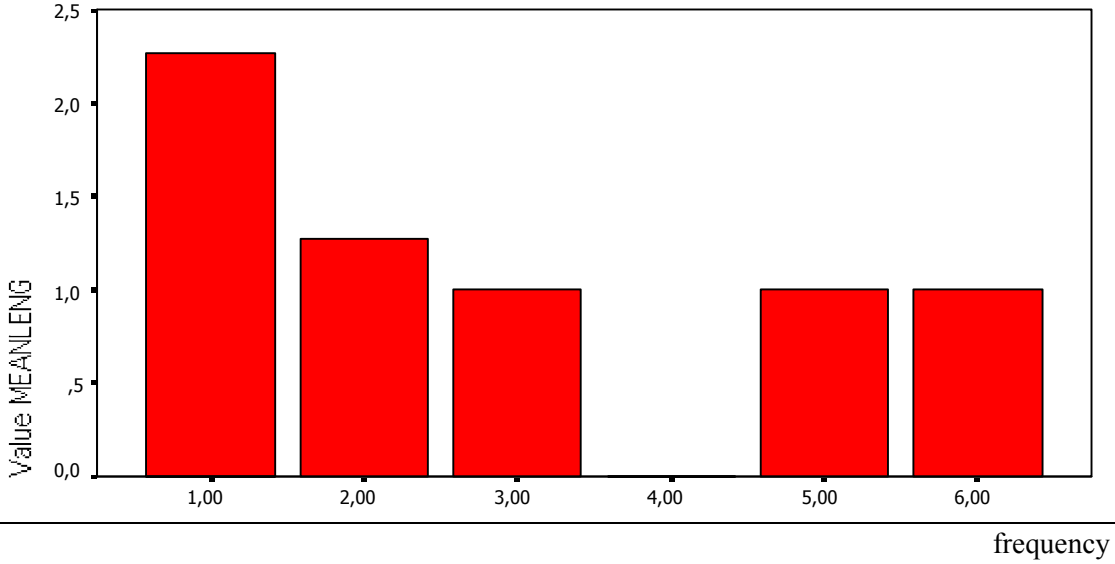


Fig. I.b. Frequency and mean length

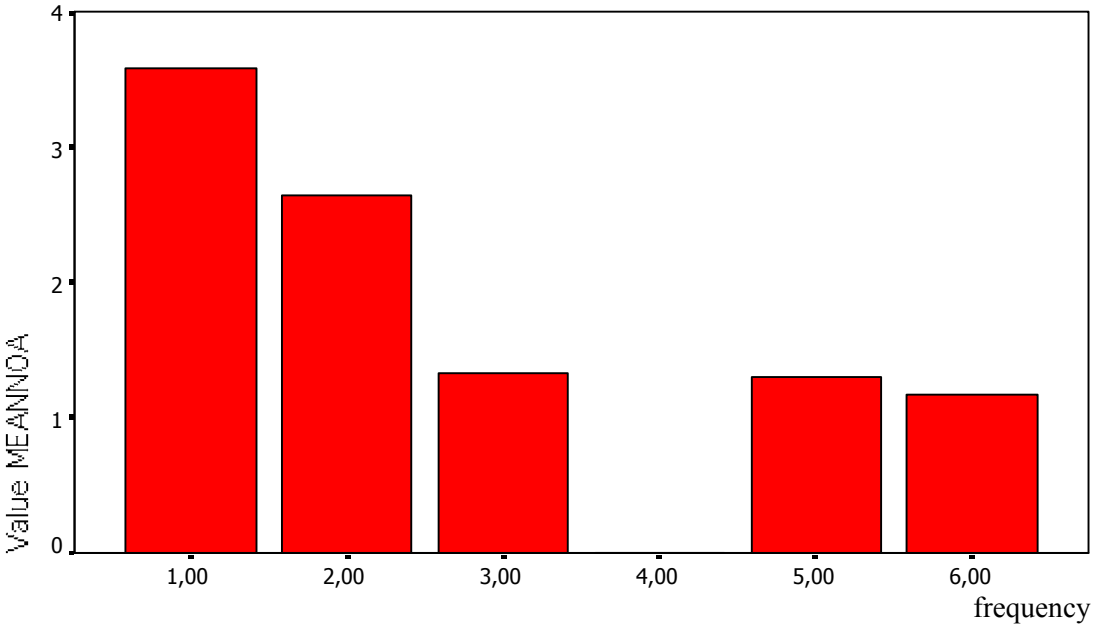


Fig. I.c. Frequency and mean number of alternatives (noa)

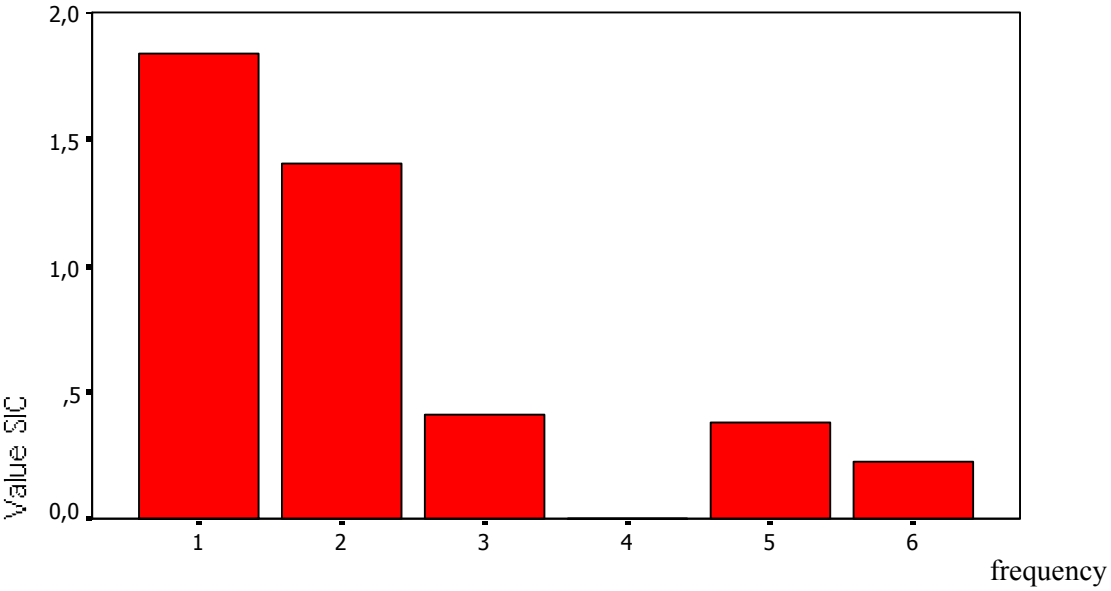


Fig. I.d. Frequency and mean *sic* (speaker's information content)

Now we determined zero-order and partial correlations between frequency, mean length, mean number of alternatives and mean *sic*: The correlation r_{fL} = -0.7512 between frequency

and mean length for word classes shows the inverse relationship much more distinct than the correlation between frequency and length (-0.2644) for single words. Controlling for the number of alternatives or for *sic*, this correlation changes drastically; it gets even positive: $r_{fL \cdot a} = 0.3242$ controlling for *a*, and $r_{fL \cdot sic} = 0.1771$ controlling for *sic*.

This is a very clear hint that the original *fL*-correlation is very much effected by or due to this third variable, measured either as the number of alternatives or as *sic*, that in turn has very high zero-order correlations with frequency and mean length (see Figures b.1 and b.2). By knowing the mean number of alternatives or *sic*, the mean length can be inferred better than by knowing its frequency.

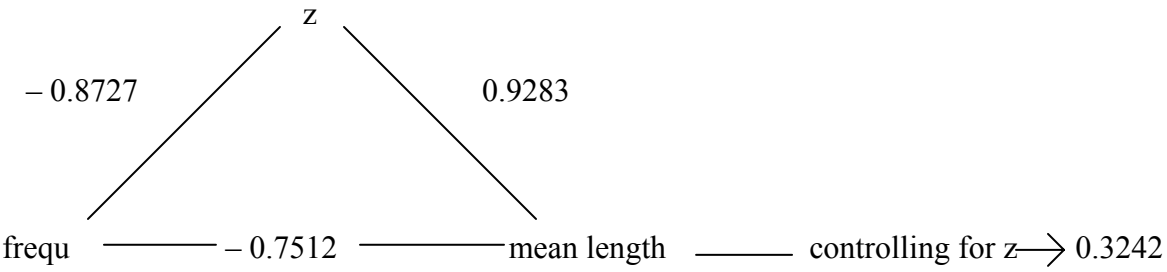


Figure b.1. Partial and zero-order correlations of frequency, mean length and mean number of alternatives (z = mean number of alternatives)

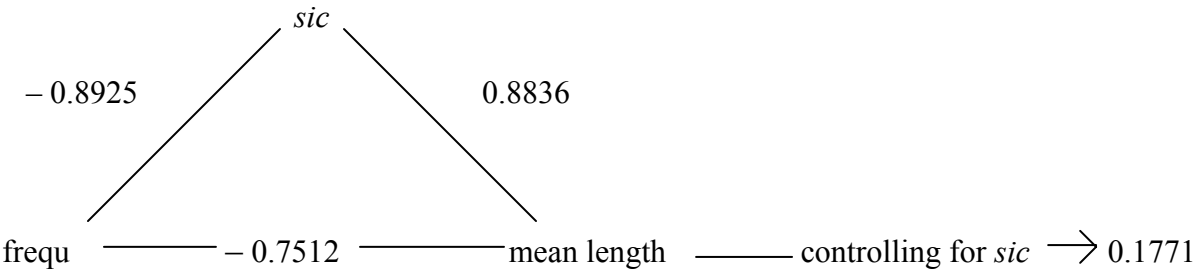


Figure b.2. Partial and zero-order correlations of frequency, mean length and mean sic

Controlling for the effect of a third variable can be considered holding this variable constant. This leads us to the idea of taking a detailed look at all those cases with $a = 1$ and regarding the proportional amount of them.

We counted the number of those cases with "no other alternative" (that means: no other word was possible at the specific position than the one that was chosen), and called them "forced choice" (*fc*).

Table 2.2
Data grouped according to their frequency, showing number and proportion of forced choice

f	class size	f x cls	m. length	fc	mean fc
1	88	88	2.26	40	0.45
2	11	22	1.27	14	0.64
3	1	3	1.00	3	1.00
4	0	0	0	0	-
5	2	10	1.00	8	0.80
6	1	6	1.00	4	0.67

fc = number of cases with forced choice ($a = 1$);
mean fc = *fc*/(*f* x class size) = *fc*/tokens

Figure b.3 shows the zero-order correlations and the partial correlation.

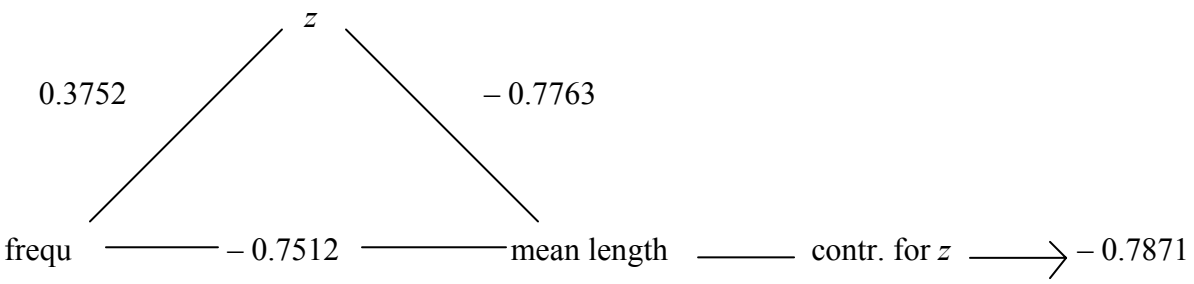


Figure b.3. Partial and zero-order correlations of frequency, mean length and mean *fc*

If the proportion of *fc* in a class ("mean *fc*") is held constant, the *fL*-correlation of – 0.7512 increases: $r_{fL \cdot mfc}$ = – 0.7871.

It becomes visible that the differing shares of *fc* in the different classes here slightly suppress the relationship between frequency and mean length. We observe a strong negative relationship between the proportion of *fc* and the mean length in a class: the longer the words are on average in a class, the smaller becomes the *fc*-proportion, this means, the more voluntary or "free chosen" they are. Again we find reason for the supposition that long words are connected with free choice.

The *fc* proportion seems to distort the *fL*-correlation. So we will group the data from another perspective.

We now look at length classes.

Third: Data grouped according to their length

The words were classified according to their lengths. For each class the following was recorded:

f = frequency; here this means class size or number of tokens,
types = number of types,
a = number of alternatives
mean a = mean number of alternatives *a/f*
sic (of mean *a*) was calculated for the entire class, by $h(x_i)$ = ld(*mean a*)

Table 3.1
Data grouped according to their lengths

length	f	types	a	mean a	sic (mean a)
1	60	37	100	1.67	0.74
2	42	39	157	3.74	1.90
3	15	15	80	5.33	2.41
4	5	5	14	2.80	1.49
5	6	6	45	7.50	2.90
7	1	1	1	1.00	0.00

Additionally we counted

fc = the number of forced choice (number of cases with no alternative than the one chosen),

ffree = the frequency or number of "free chosen" words, this is the number of all cases with any more alternative than the one chosen ($a > 1$),

f/t = the mean frequency of a word in the class, calculated as frequency/number of types.

Table 3.2
Data grouped according to their lengths, showing the number of cases with forced choice

length	f	f/t	fc	ffree
1	60	1.62	42	18
2	42	1.08	24	18
3	15	1.00	6	9
4	5	1.00	2	3
5	6	1.00	1	5
7	1	1.00	1	0

f/t = frequency/ number of types = mean frequency
fc = forced choice
ffree = *f* – *fc*

We are interested here in class sizes or numbers of cases. The correlations between the number of class members *f* (here *f* = class size), lengths of class members *L* and number of forced choice members *fc* were determined (see Fig. c).

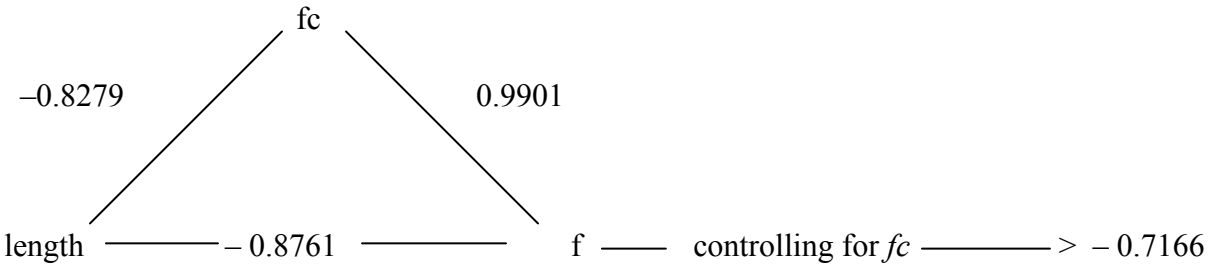


Fig. c. Partial and zero-order correlations of length, frequency (class size) and number of *fc* members (*fc* = frequency of cases with forced choice)

The zero-order correlation between the length and the frequency (number of cases) is $r_{fL} = -0.8761$.

If the influence of *fc* is eliminated from it or the number of *fc*-cases is held constant, the correlation decreases: $r_{fL \cdot fc} = -0.7166$. That means the *fL*-correlation is partly due to the number of *fc*.

If we now look at the numbers of *fc* and *ffree* we can see that among the 129 words (tokens) there are only 53 "free chosen". In contrast, there are 76 *fc* cases.

$\sum fc = 76$
$\sum ffree = 53$
$\sum f = 129$

The Figures II.a,b show the distributions of *fc* and *ffree* cases among the length classes. We can see: The two-syllable words are not "less popular" than those with one syllable. But the one-syllable words are very often cases of forced choice.

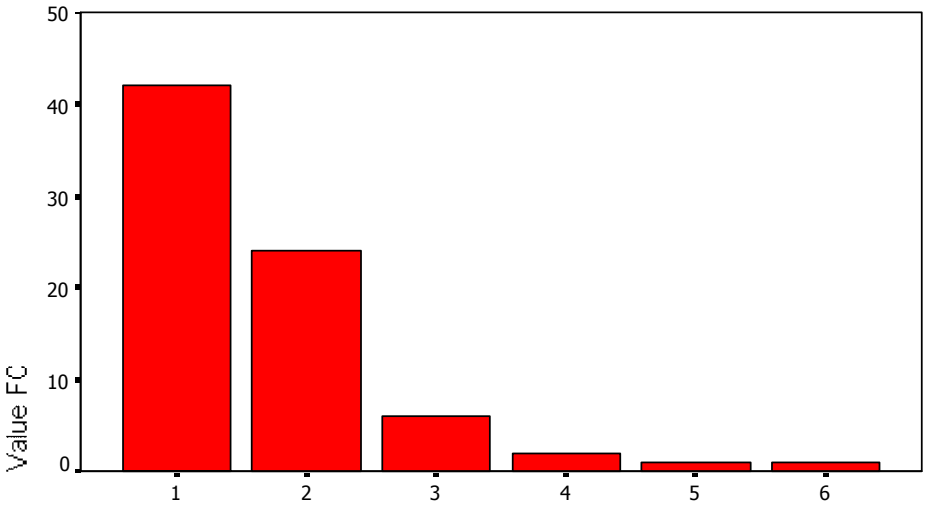


Fig. II.a. Cases of forced choice among length classes

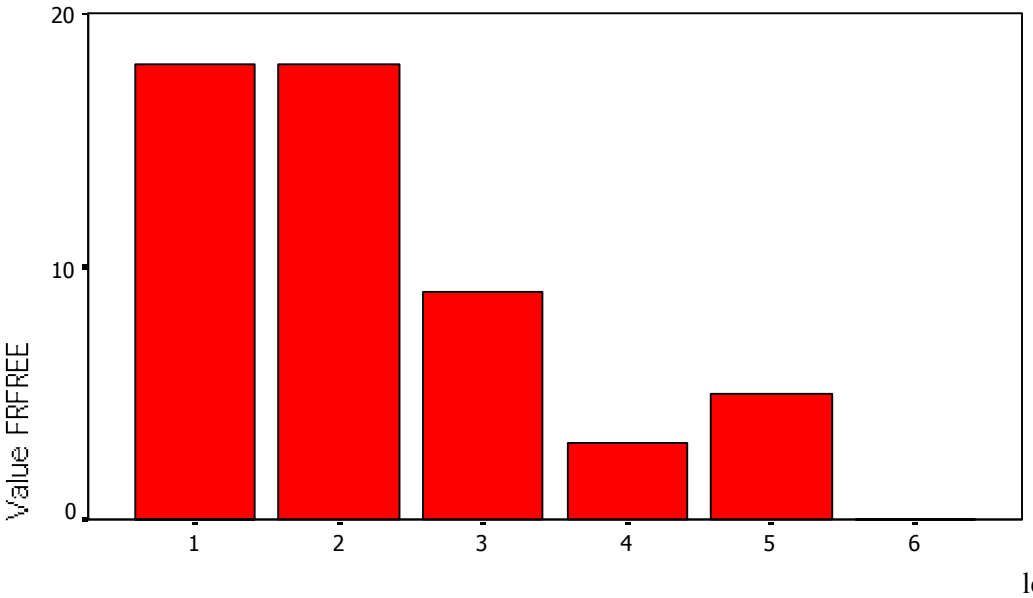


Fig. II.b. Frequencies of free chosen words among length classes

The data suggest a hypothesis:
The word frequencies *f* result from two sets of words with differing distributions:

- the distribution of *fc* which correlates nearly perfect with class frequencies *f*
- the distribution of *ffree* which approximates equal distribution if we categorize into "one-syllable","two-syllable" and "three or more-syllable words" with each ≈ 18.

Table 3.3
frequencies of free chosen words (*ffree*), of forced choice-words (*fc*),
together with total frequencies (*f*)

length	f	fc	ffree
1	60	42	18
2	42	24	18
≥ 3	27	10	17

$r_{f,fc}$ = 0.999

So the frequency distribution is obviously due to the forced choices, which are prescribed for the text producer, who in turn "tries to" distribute the frequencies of self-chosen words equally to the three length classes. This hypothesis should be examined in further research.

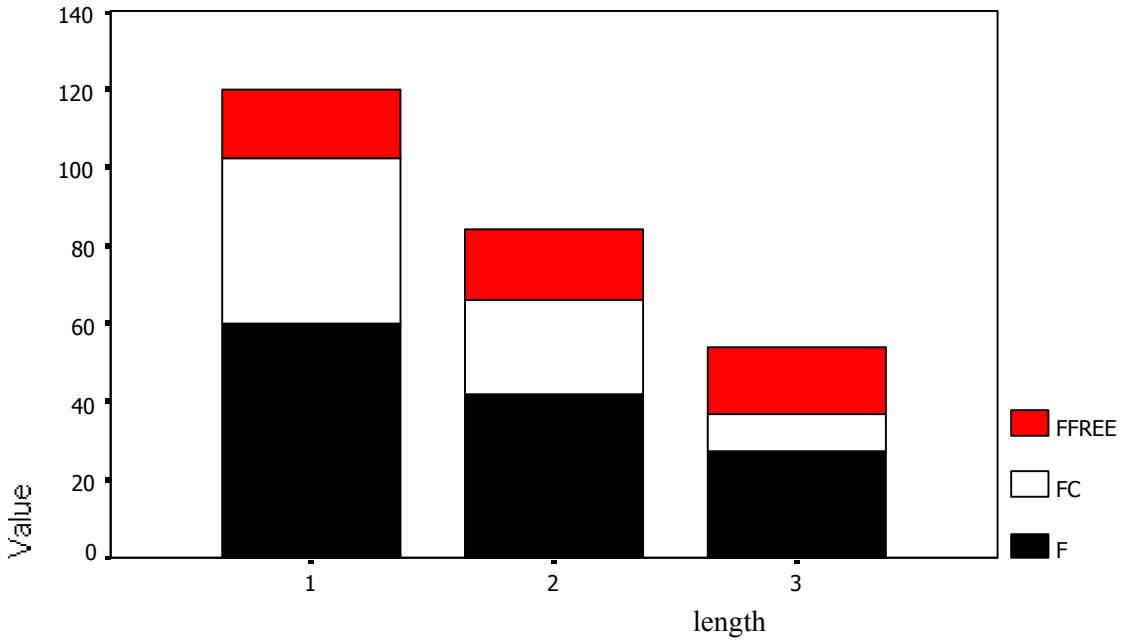


Fig. II.c. Proportions of forced choice and free chosen words in three length classes.
Length case numbers: 1 = 1 syllable; 2 = 2 syllables; 3 = 3 and more syllables

Summary: conclusion and outlook

Word lengths do not reveal information in a direct way. Other than the conspicuousness of a sound, the length of a word cannot be equalled to its information content. Speaker's information content (*sic*) has to be regarded as a third variable connected with length and frequency. The correlation between frequency and length is partly determined by it. Controlling for this third variable and its effects are necessary.

Regarding the amount of *sic* provides with new methods for the analysis of texts. A lot of refinements of the method presented above are conceivable.

Comparing "manifest" and "latent tokens": We could compare the words realized to the non-realized but conceivable alternatives (e.g. do the alternatives of a word always differ in length?)

Including the "latent frequency" when counting: We could consider - and count - the non-realized words, too: For example, a German text in which the word "sowie" is used consequently instead of "und", reveals information about author's characteristics; the procedure would resemble to certain – intuitively used - observation methods in psychoanalysis.

Determining the "popularity" of a word: We could ask how often and when a word is realized, provided it is among the alternatives.

Considering the varying probabilities of the alternatives: Up to now the p forming *sic* has been treated as to be equal for a particular position: $p = 1/n$. One could (and will have to) take into account the differing probabilities for the alternatives for a word. Appropriate ways of counting the number of alternatives together with weightings of their probabilities are necessary.

If we assume that $p \neq 1/n_i$ for the different alternatives, we can take up again the remark made above. We are dealing here with two different sorts of p: p_1 being the relative frequency of the word in the specific text and p_2 being the probability for the word to be chosen at the specific position in the text.

If we define $p(y)$, the overall frequency of a word, as a function of $p(x)$, the frequency of favourable occasions for it, and $p(y|x)$, the particular probability of being uttered at a specific occasion x:

$$p(\mathrm{y}) = \sum_x p(x)p(y \mid x)$$ (detailed explications see Andersen, 2002),

we can use the formula to describe the relative frequency in a text and demonstrate how it is related to the information content of a word. The y are words, while the x are abstract entities, »occasions», positions in the text or decision situations.

We enter here p_1 for $p(y)$. p_2 corresponds to $p(y|x)$. If $p(y|x)$ could be estimated, it would give a more precise measure for p_2 than $1/n$, so h_i could be determined with even more precision.

References

Altmann, G. (1993). Phoneme counts. In: Altmann, G. (ed.), *Glottometrika 14, 54-68.*

Altmann, G., Köhler, R. (1995). Language forces and synergetic modelling of language phenomena. *Glottometrika 15, 62-76.*

Andersen, S. (1985). *Sprachliche Verständlichkeit und Wahrscheinlichkeit.* (= Quantitative linguistics 29). Bochum: Brockmeyer.

Andersen, S. (2002). Freedom of choice and the psychological interpretation of word frequencies in texts. *Glottometrics 2, 45-52.*

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review 61, 183-193.*

Attneave, F. (1955). Symmetry, information and memory for patterns. *American J. of Psychology 68, 209-222.*

Attneave, F. (1959). *Application of information theory to psychology*. New York: Holt.

Attneave, F. (1968). Triangles as ambiguous figures. *American J. of Psychology 81, 447-453.*

Berlyne, D.E. (1958). The influence of complexity and novelty in visual figures on orienting responses. *J. of Experimental Psychology 55 (3), 289-296.*

Berlyne, D.E. (1960). *Conflict, arousal and curiosity.* New York: McGraw-Hill.
Birkhoff, G.D. (1933). *Aesthetic Measure*. Cambridge: Harvard University Press.
Coombs, C.H., Dawes, R.M., Tversky, A. (1970). *Mathematical Psychology: An elementary introduction.* Englewood Cliffs, N.J.: Prentice-Hall.
Evans, T.G. (1967). A program for the solution of a class of geometric-analogy intelligence-test questions. In M. Minsky (Ed.), *Semantic Information Processing: 271-353*. M.I.T. Press, Cambridge, Mass.
Fitts, P.M., Weinstein, M., Rappaport, M. Anderson, N., Leonard, J.A. (1956). Stimulus correlates of visual pattern recognition – a probability approach. *J. of Experimental Psychology 51, 1-11.*
Garner, W.R. (1962). *Uncertainty and Structure as Psychological Concepts*. New York: Wiley.
Garner, W.R. (1970). Good patterns have few alternatives. *American Scientist 58, 34-42*.
Garner, W.R., Hake, H. (1951). The amount of information in absolute judgements. *Psychological Review 58, 446-459.*
Guiraud, P. (1963). Structure des répertoires et répartition fréquentielle des éléments: La statistique du vocabulaire écrit. In: Moles, A.A. & Vallancien, B. (eds.), *Communications et langages: 35-48.* Paris: Gauthier-Villars.
Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells.* Trier: Wissenschaftlicher Verlag.
Hartley, R.V.L. (1928). Transmission of Information. *Bell System Technical Journal 7, 535-563.*
Köhler, R. (1986). *Struktur und Dynamik der Lexik (*= Quantitative linguistics 31). Bochum: Brockmeyer.
Köhler, R. (1987). System theoretical linguistics. *Theoretical linguistics 14, 2/3, 241-257.*
Köhler, R. (1990). Elemente der synergetischen Linguistik. *Glottometrika 12, 179-187.*
Köhler, R., Altmann, G. (1993). Begriffsdynamik und Lexikonstruktur. In: F. Beckmann & G. Heyer (eds.), *Theorie und Praxis des Lexikons*: *173-190.* Berlin, New York: De Gruyter.
Mandelbrot, B. (1954). Structure formelle des textes et communication. *Word 10, 1-27.*
Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal 27, 379-423 & 623-656.*
Taylor, W.L. (1953). Cloze Procedure: A new tool for measuring readability. *Journalism Quarterly 30, 4, 415-433.*
Wiener, N. (1948). *Cybernetics. Or control and communication in the animal and the machine.* Cambridge, Mass.: M.I.T. Press.
Wohlwill, J.F. (1984). What are sensation seekers seeking? Open Peer Commentary to Zuckerman: Sensation seeking. *Behavioral and Brain Sciences 7, 453.*
Zipf, G.K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology 40.*
Zipf, G.K. (1932). *Selected studies of the principle of relative frequency in language.* Cambridge, Mass.: Harvard Univ. Press.
Zipf, G.K. (1935). *The psycho-biology of language. An introduction to dynamic philology.* Cambridge, Mass.: M.I.T. Press.
Zipf, G.K. (1949). *Human behavior and the principle of least effort.* Cambridge, Mass.: Addison-Wesley Press.
Zuckerman, M. (1979). *Sensation seeking: Beyond the optimal level of arousal.* Hillsdale, NJ: Erlbaum.
Zuckerman, M. (1984). Sensation seeking: A comparative approach to a human trait. *Behavioral and Brain Sciences 7, 413-417.*

Zuckerman, M. (1991). *Psychobiology of Personality*. Cambridge: Cambridge University Press.

Appendix

I. Analysed text:

129 words (tokens) without headline
(„Hamburger Morgenpost“, 24. April 2002, p. 11):

Das dickste Spürnasenbuch der Welt

Ein Taschenbuch ist das nicht gerade, was gestern auf dem Bahnhof Altona als “dickstes Kinderbuch der Welt“ vorgestellt wurde: Fünf Meter dick ist der Monster-Wälzer und wiegt 250 Kilo. Die Stiftung Lesen hatte zum “Welttag des Buches“ Schüler in ganz Deutschland aufgerufen, einen spannenden Kinder-Krimi zu schreiben.
Kinderbuchautor Andreas Steinhöfel dachte sich den Anfang aus, und 18000 Jungen und Mädchen “strickten“ die Geschichte rund um eine geheimnisvolle “Weltuhr“ und einen bösen Gauner zu Ende. 47000 Seiten kamen zusammen - die Aktion soll ins Guinness-Buch der Rekorde aufgenommen werden.
Zur Präsentation des Gemeinschaftswerkes kamen viele der Nachwuchsautoren nach Altona. Und weil es in dem Buch um zwei junge Spürnasen geht, spielten Stella-Musical-Stars in der Bahnhofshalle für die Kinder Szenen aus “Emil und die Detektive“ nach.

II. First recording of words: length *L*, frequencies *f*, *sic* and number of alternatives

L	f	sic	a
1.00	1.00	.00	1.00
3.00	1.00	4.00	16.00
1.00	2.00	1.60	3.00
1.00	1.00	1.00	2.00
1.00	1.00	.00	1.00
3.00	1.00	3.00	8.00
1.00	1.00	1.00	2.00
2.00	1.00	.00	1.00
1.00	1.00	1.00	2.00
1.00	2.00	.00	1.00
2.00	1.00	.00	1.00
3.00	1.00	.00	1.00
1.00	1.00	.00	1.00
2.00	1.00	3.00	8.00
3.00	1.00	.00	1.00
1.00	5.00	.00	1.00
1.00	1.00	.00	1.00
3.00	1.00	4.00	16.00
2.00	1.00	.00	1.00
1.00	1.00	.00	1.00
2.00	1.00	.00	1.00

1.00	1.00	1.60	3.00
2.00	1.00	3.00	8.00
2.00	1.00	2.00	4.00
1.00	6.00	.00	1.00
1.00	1.00	.00	1.00
5.00	1.00	.00	1.00
2.00	1.00	1.00	2.00
1.00	5.00	.00	1.00
2.00	1.00	.00	1.00
2.00	1.00	.00	1.00
2.00	1.00	.00	1.00
1.00	1.00	.00	1.00
2.00	1.00	.00	1.00
1.00	2.00	.00	1.00
2.00	1.00	.00	1.00
2.00	1.00	2.00	4.00
1.00	3.00	.00	1.00
1.00	1.00	1.00	2.00
2.00	1.00	.00	1.00
4.00	1.00	2.00	4.00
2.00	2.00	.00	1.00
3.00	1.00	3.00	8.00
2.00	2.00	.00	1.00
2.00	1.00	1.60	3.00
1.00	2.00	.00	1.00
2.00	1.00	3.00	8.00
5.00	1.00	2.00	4.00
3.00	1.00	.00	1.00
3.00	1.00	.00	1.00
2.00	1.00	.00	1.00
1.00	1.00	1.60	3.00
1.00	1.00	1.00	2.00
2.00	1.00	1.60	3.00
1.00	2.00	.00	1.00
4.00	1.00	.00	1.00
2.00	1.00	.00	1.00
2.00	1.00	.00	1.00
2.00	1.00	4.00	16.00
3.00	1.00	2.00	4.00
1.00	1.00	1.00	2.00
1.00	2.00	.00	1.00
2.00	1.00	.00	1.00
5.00	1.00	3.00	8.00
2.00	1.00	1.00	2.00
2.00	1.00	4.00	16.00
2.00	1.00	3.00	8.00
2.00	1.00	.00	1.00
7.00	1.00	.00	1.00
2.00	1.00	.00	1.00
2.00	2.00	.00	1.00

3.00	1.00	.00	1.00
3.00	1.00	2.00	4.00
1.00	1.00	1.00	2.00
1.00	1.00	.00	1.00
2.00	1.00	.00	1.00
1.00	2.00	.00	1.00
3.00	1.00	.00	1.00
4.00	1.00	2.00	4.00
2.00	1.00	.00	1.00
1.00	1.00	.00	1.00
5.00	1.00	3.00	8.00
5.00	1.00	3.00	8.00
2.00	1.00	3.00	8.00
5.00	1.00	4.00	16.00
1.00	2.00	.00	1.00
3.00	1.00	1.00	2.00
1.00	1.00	1.00	2.00
1.00	1.00	.00	1.00
1.00	1.00	3.00	8.00
2.00	1.00	3.00	8.00
3.00	1.00	4.00	16.00
1.00	1.00	.00	1.00
2.00	1.00	3.00	8.00
2.00	1.00	1.00	2.00
3.00	1.00	1.00	2.00
1.00	1.00	3.00	8.00
4.00	1.00	2.00	4.00
1.00	1.00	.00	1.00
2.00	1.00	2.00	4.00
2.00	1.00	.00	1.00
4.00	1.00	.00	1.00

Glottometrics 3, 2002, 110-135 *To honor G.K. Zipf*

A conceptualization of the configurational and functional organization

Vladimír Majerník[1]

Abstract. Configurational and functional organizations are distinguished and analyzed. Information theoretical measures are proposed and examples of organization form physics, biogenesis, music and language are presented.

Keywords: General systems, system property, elements of interaction, organization, goal directed systems, information

1. Introduction

One of the most prominent tasks of science is to reduce the variety of events and phenomena occurring in nature to a few entities, ideas or models having simple logical and mathematical structure such as motion, field, randomness or orderliness. While motion and field are described by the causal and deterministic laws, randomness obeys the indeterministic and 'stochastical' laws.[2]

To describe the orderliness one needs a collective of the *interacting* objects. Such a collective forms a **system**. The notion of **system** became a unifying concept and an important paradigm in science with large applicability to the human sciences and technology as well. For example, the physical statistical ensemble is, in fact, a system of the physical stochastical objects (spins, molecules, etc.) between which a physical interaction and the statistical dependencies exist.

One of the most important properties of a whole class of physical, cybernetic, biological and sociological systems appears to be their **organization**. The term "**organization**" as used in different disciplines often has different meanings. This is so because the organization of a probabilistic system is generally related to its **probabilistic uncertainty**, while in the goal-directed systems, it is related to the way in which the individual subsystems of these systems **cooperate** in order to achieve a specific **goal** or **function.**[3] We attempt to precise the often confusing use of the term 'organization' in that we make a difference between two types of organization: the configurational and functional organization.

---

[1] Address correspondence to: Vladimír Majerník, Dept. of Theoretical Physics, Palacky University, Tr. 17. listopadu 50, CZ-72007 Olomouc, Czech Republic. E-mail: majerv@prfnw.upol.cz

[2] The progress of science often arises when a concept of everyday life is precised and exactly defined. In physics, the precised notions as, e.g. work and force form the basic paradigms of mechanics. A great progress of science materialized in 1948 when Shannon defined the hitherto vague concept of information. In the last decades the concept of system has been put more precisely in the framework of a general systems theory.

[3] It is remarkable that many authors use the terms "organization" and "order" interchangeably as if they were synonyms. This may lead to serious errors. A classic example is a living cell versus a crystal. A crystal has a large **configurational order**, i.e. it has an ordered spatial structure. Due to this ordered structure it is often said that it has a large organization. In contrast, in the cell there is less spatial and temporal orderliness, but its subsystems exhibit to a great extent a coherent and cooperative behaviour which enables it to pursue a certain biological function. We see that the term "organization" is used in a different sense in this example.

The configurational organization is typical for the probabilistic systems, i.e., for collectives of interacting objects obeying natural laws without following a specific goal. A general probabilistic system represents a suitable framework for quantifying of randomness as well as orderliness. The characteristic property of all probabilistic systems is their **probabilistic uncertainty**. Since the most important measure of this uncertainty is **entropy** the randomness and orderliness of probabilistic systems can be described quantitatively by using the concept of entropy.

The functional organization is linked with goal-directed systems, i.e. those systems which tend to achieve the a-priori goals. To describe such systems quantitatively one needs to find an appropriate measure for their ability to reach their goals. While the configurational organization of a probabilistic system is uniquely quantified by means of its probabilistic uncertainty, the quantification of the functional organization is a very complicated problem so far not being solved unambiguously. This situation still materializes, in spite of the fact that several authors have tried to quantify the degree of the functional organization of the goal-directed systems. For example, Denbigh (1975) attempted to express the degree of the functional organization of biological objects by means of the so-called **integrality**. The amount of integrality is derived from the fact that any **functionally** organized system is necessarily an assembly of parts which are interconnected. Integrality is essentially the product of the number of connections which "facilitate" the function of the system and the number of its different parts. This measure does not adequately quantify the organization of the goal-directed systems because it takes into account only the number of their inter-connections and parts without their exact quantification with respect to their functions and goals. Therefore, we will try to find another measure for the degree of the functional organization of the goal-directed systems which also takes into account the magnitudes of the cooperative behaviour between the system elements which is related to their goal.[4]

This article is devoted to the notion of organization which is frequently confused with the concept of order. In my paper below configurational organization - where entropy and information play a central role - is differentiated from functional organization. For the latter, a measure is defined by introducing the notions of functional connection and co-operative dependencies. Examples are given how these concepts can be applied to different areas of science. This article is organized as follows: Part I is devoted to the configurational organization of the probabilistic systems. We will introduce and describe the quantities which are necessary to construct an appropriate measure of their configurational organization. Part II is devoted to the goal-directed systems which are organized functionally. There, we will extensively describe the behaviour of such systems and try to quantify the functional organization by introducing the notions of the functional connection and cooperative dependencies between their elements. We will describe some important goal-directed systems, as living objects, language and music, in greater details.

[4] There may be an interrelation between both types of organization, which is especially important for systems in which the functional organization is realized through a certain order of physical objects. A collective of interacting physical objects obeying physical laws, may strongly affect the functional organization of a goal-directed system.

2. Elements of general system's theory

A general complex system represents itself as a collection of objects between which an interaction exists
C. I. Shenman

Alles hängt mit allem zusammen
Karl Marx

The interdisciplinary approach to various natural phenomena is the main trend of modern sciences. This approach has been applied to sciences more and more due to new achievements in the recent decades in biology, chemistry and physics. The need for a more rigorous approach to the study of large, complex and interrelated phenomena in mathematics, physics and the social sciences led to the development of the general systems theory (see, e.g. Klir 1970). A general system is defined as a set of **statements and propositions** involving a collection of objects and expressing relationships between them (Eckman and Mesarovic' 1961). A more detailed definition can be found in Bunge (1979: 7). Mathematically, a system is determined by the set of its elements, $S \equiv \{s_1, s_2, ..., s_n\}$, together with the set of relations between them, $R \equiv \{r_i(j)\}$, where $r_i(j)$ is a function defined by the elements s_i and s_j. A system is a set of the interacting parts creating through their interaction new **system property** not existing in a loose collection of individual objects. First, due to the interaction between the system elements, a system is formed.[5]

To describe a system quantitatively one needs to find quantities which can serve as measures for its system´s properties. These measures should fulfill certain reasonable mathematical requirements which can be summarized as follows: (i) They should contain the quantities characterizing the individual system elements. (ii) They should contain terms describing element interaction. (iii) They should depend on the structure of a given system, i.e. on how its elements are arranged and how they mutually interact. By means of the suitable con-structed measures for a system property, one can make quantitative statements about its magnitude.

There are two essential types of system in science: probabilistic and goal-directed. While in a probabilistic system, the system property asked is its probabilistic uncertainty, in a goal-directed system, this system property is a-priori given by its goal. To organize a system means to form such a structure in it that the required system property assumes its desired value. Mathematically, a probabilistic system is generally described by a set of statistically dependent random variables $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$.[6] These random variables, called element´s random variables, are characterized by their probability distributions and their statistical dependencies (statistical linkages) which are given by the conditional probability distributions that an element random variable $\tilde{x}_i$ assumes its k-th values when element´s random variable $\tilde{x}_j$ has its l-th value. An important property of a probabilistic system is its **probabilistic uncertainty** whose measure can exactly be quantified. The well-known measure for the probabilistic uncertainty is the Shannon entropy.

---

[5] Although the general systems theory was developed approximately forty years ago, there are systems which have not been fully integrated. This is the case, because many systems were introduced and studied in sciences (physics, biology, etc.) long before the existence of the general systems theory. Particularly, the concept of thermo-dynamic systems, determined by the values of its state variables (e.g. internal energy, temperature, entropy, etc.), and that of a physical statistical system, characterized by its configurational order, although intensively studied since the beginning of thermodynamics and statistical physics, are not fully integrated into the general systems theory.

[6] A random variable is a mathematical quantity assuming a set of values with the corresponding probabilities.

Here, the simple rule holds that the smaller the orderliness of a system (given by means of magnitudes of the mutual statistical dependencies of their elements), the larger its entropy. The entropy of a probabilistic system appears to be the most appropriate measure of its configurational organization (Watanabe 1969). The orderliness and entropy of a probabilistic system are interrelated with each other (inversely and logarithmically) so that any *increase* in the degree of configurational organization must necessarily result in the *decrease* of its entropy. The measure for the configurational organization, constructed by means of entropies, is called the Watanabe measure and defined as follows (Watanabe 1969):

(configurational organization of a probabilistic system) =
(entropy of whole system)-(sum of entropies of parts of this system)

In quantitative terms, this measure expresses the property of a configurationally organized system to have order between its elements which causes the system as a whole to behave in a more deterministic way than its individual parts. If a system only consists of elements which are statistically independent, then its Watanabe measure becomes zero. If a system consists only of elements which are deterministically dependent, then its Watanabe measure becomes maximum. Generally, the configurational organization of a system lies between these extremes. The prominent systems the Watanabe measure can be applied to include the physical statistical ensembles, above all, the Ising systems of spins (Kubo 1974). The high configurational organization especially exhibits those systems which have some spatial, temporal or spatio-temporal structures arisen in them when they are far from thermal equilibrium (e.g. laser, fluid instabilities, etc.) (Haken 1983). These systems can only be sustained by a steady flow of energy and matter (Ke-Hsueh Li 1986; Bertalanffy 1953; Nicolis and Prigogine 1976).

A goal-directed system has an a-priori given goal and one seeks for such a structure that the requested system property becomes optimal. Here, the set of relations $\{r_i(j)\}$ does generally not consist of the conditional probabilities but other quantities characterizing the cooperative interaction between the system elements. The structure and the cooperative interaction between the system elements determine the organization of goal-directed systems. Such an organization is called functional. A general goal-directed system is functionally well-organized if the success of the system toward attaining its goal exceeds the sum of the successes of its parts. The functional organization became an important concept of science because the goal-directness is typical of many systems in biology, robotics or cybernetics.[7]

3. Configurational organization

Die physikalische Forschung hat klipp and klar bewiesen, dass die gemeinsame Wurzel der beobachteten strengen Gesetzmäßigkeit der ***Zufall*** *ist.*
E. Schrödinger

Hinter dem Ordnungsbegriff des Mathematikers steht vor allem der Aspekt der (eindeutigen) Anordnung, während 'Ordnung' in Physik eher im Kontrast zur 'Unordnung' gesehen wird.
M. Eigen

[7] It is interesting that many authors dealing with physical or biological systems are not sufficiently aware of the existence of the general systems theory which they consider as an abstract and formal theory not able to describe the majority of systems. However, as we will show later on, even this abstractness of systems theory forms a suitable framework in which the different systems can be adequately described.

At the most fundamental level, all our further considerations are based on the concept of **probability**. Although there is a well-defined mathematical theory of probability, there is no universal agreement about the meaning of probability. Thus, for example, there is one view that probability is an objective property of a system, and another view that it describes a subjective state of belief of a person. Then there is the frequentist view that the probability of an event is the relative frequency of its occurrence in a long or infinite sequence of trials. This latter interpretation is often employed in the mathematical statistics and statistical physics. In everyday life probability stands for the degree of ignorance as to the outcome of a random trial. Commonly, probability is interpreted as the degree of the subjective expectation of an outcome of a random trial. Both subjective and statistical probabilities resp., are "normed", i.e. the degree of expectation that an outcome of a random trial occurs, and the degree of the "complementary" expectation, that it does not, is always equal to one.[8]

The word 'entropy'[9] was first used in 1884 by Clausius in his book *Abhandlungen über Wärmetheorie* to describe a quantity accompanying a change from the thermal to mechanical energy and it continued to have this meaning in thermodynamics. Boltzmann (1896) in his *Vorlesungen über Gastheorie* laid the foundation of the statistical theory of entropy by providing its measure. He linked entropy with molecular disorder. The concept of entropy as a measure of uncertainty was first introduced by Shannon (1948). Wiener (1948) and Shannon (1948) are credited for the development of a quantitative measure of the amount of information. Shannon's entropy may be considered as a generalization of entropy, defined by Hartley (1928), when the probability of all events is equal. Nyiquist (1924) was the first author who introduced a measure of information. His paper has largely remained unnoticed. After the publication of Shannon's seminal paper in 1948, the theory grew rapidly and was applied with varying success to most areas of human endeavour.

Mathematicians were attracted by the possibility of providing an axiomatic structure of entropy and its ramification. The axiomatic approach to the concept of entropy at-tempts to formulate a postulated system which, apparently, provides a unique numerical char-acteristic of entropy[10] and which adequately reflects the properties asked from the probabil-istic uncertainty measure in a diversified real situation. This has been a very interesting and thought-provoking area for information scientists. Khinchin (1957) was the first who gave a clear and rigorous presentation of the mathematical foundation of entropy. A good number of works has been published on the axiomatic characterization of different measures of probabilistic uncertainty and information based on various postulates and axioms. An extensive list of works in this field can be found in the book of Aczél and Daróczy (1975).

In the introduction, we claimed that randomness is intimately connected with the concept of entropy. In order to discuss this subject in more details let us introduce the so-called **stochastical object** which represents an object obeying probabilistic laws (Welsh 1970). A stochastical object is given by the set of its possible states, *S*, their probability distribution, **P**, and the values defined on these states, i.e., to any stochastic object is assigned a random variable $\tilde{x}$. A simple stochastical object characterized by a discrete random variable $\tilde{x}$ is usually given by a probabilistic scheme

---

[8] Although the concepts of probability are generally governed by a sophisticated mathematical language, it expresses only the commonly familiar properties of probability used in everyday life. For example, each number of spots at the throw of a simple die represents an elementary random event to which a positive real number is associated called its probability (relation(i)). The probability of two (or more) numbers of spots at the throw of a simple die is equal to the sum of their probabilities (relation(iii)). The sum of probabilities of all possible numbers of spots is normed to one.

[9] The word 'entropy' stems from the greek word '*τροπη*' which means 'transformation'.

[10] Entropy is sometimes called "missing information".

S	S_1	S_2	.	.	.	S_n
P	$P(x_1)$	$P(x_2)$	.	.	.	$P(x_n)$
x	x_1	x_2	.	.	.	x_n

The states of a stochastical object are given in the first row. In the second row, there are *components* of the probability distribution of $\tilde{x}$ and the third row gives the values of $\tilde{x}$ defined on its states. By means of the concept of the stochastical object many important probabilistic phenomena can be simulated, especially, the so-called **random experiment**. A random experiment is an experiment whose outcomes are given only with certain probabilities.[11]

A random experiment is fully described by the set of its outcomes, their probabilities and the values of the random variable defined by its outcomes. It is already intuitively clear that with any random experiment a probabilistic uncertainty in its outcomes is linked which depends on the number of its outcomes and their probabilities. The more alternative outcomes a random experiment has, the larger its probabilistic uncertainty is.

An important problem is the determination of a measure of the probabilistic uncertainty of a random experiment. In realization of a random experiment two stages are to be distinguished: the stage *before* and the one *after* its performance. Within the first stage one does not know its outcome, only the relevant probability distribution. Therefore, there exists in this stage an uncertainty in its outcomes. Within the second stage, after the performance of the random experiment, its result is known and its probabilistic uncertainty is completely removed. The mathematical requirements asked from the uncertainty $H(P_i)$ of the i-th **outcome** of a random experiment are the following (Faddejew 1957):

(i) It should be a monotonically decreasing continuous and unique function of the probability of this outcome;

(ii) The common value of the uncertainty of a certain outcome of two statistically independent experiments $H(P_iP_j)$ should be additive, i.e.,

$$H(P_iP_j) = H(P_i) + H(P_j)$$

where P_i and P_j is the probability of the i-th and j-th outcome, respectively.

(iii) $H(P_i = 1/e) = 1$ (determination of unit of the entropic measure of probabilistic uncertainty).

It was shown that the only function which satisfies these requirements has the form (Faddejew 1957)

$$H_i = H(P_i) = -\log P_i.$$

The quantities $H_1, H_2, \ldots, H_n$ attached to outcomes of random experiments used to be added to a complete probabilistic scheme of a random experiment. Their mean value is given by the formula

$$(3.1) \quad H(\mathbf{P}) = H(\tilde{x}) = \sum_i P_i H(P_i) = -\sum_i P_i \log(P_i)$$

[11] A typical random experiment is the throw of a single die whose probabilistic scheme is

S	S_1	S_2	S_3	S_4	S_5	S_6
P	1/6	1/6	1/6	1/6	1/6	1/6
x	1	2	3	4	5	6

The expression (3.1) is called *information-theoretical* or *Shannon's entropy*.

The determination of an entropic measure of the probabilistic uncertainty of a random experiment was a milestone in the history of science. Although this measure could have been found in many branches of science, it was found in the framework of communication theory. In 1948 Claude E. Shannon introduced the information-theoretical entropy in a rigorous manner. Since then it has been applied to many areas of science and become a powerful tool in describing stochastical objects and probabilistic systems.

We remark that, except of the entropic measures of uncertainty, there is a class of the so-called moment uncertainty measures given mostly as the **higher statistical moments** of $\tilde{x}$. The statistical moments are often used as the uncertainty measures of the probabilistic system, especially in experimental physics, where e.g. the standard deviation of measured quantities (the square root of the second central statistical moment) characterizes the accuracy of a physical measurement. The moment uncertainty measures are also used by formulating the uncertainty relations in quantum mechanics (Messiah 1961).

3.1. Information

In the theory of probability the statistical dependence of two random variables $\tilde{x}$ and $\tilde{y}$ expresses a property similar to the functional dependence of two variables in the calculus. Therefore, it is important to have an adequate measure for the extent of the mutual statistical dependence (linkage) between random variables $\tilde{x}$ and $\tilde{y}$. In an extreme case, there is an exact correspondence between the values of the random variables $\tilde{x}$ and $\tilde{y}$. This, in fact, represents a functional dependence between the values of both random variables. Another extreme case arises when the values of both random variables are absolutely independent of each other. Here, the random variables $\tilde{x}$ and $\tilde{y}$ are fully statistically independent. In a general case, the degree of the statistical dependence of $\tilde{x}$ and $\tilde{y}$ is between these extremes.

Let us consider two random variables, $\tilde{x}$ and $\tilde{y}$, given by their probabilistic schemes

S	S_1	S_2	S_3	S_4	S_5	S_6
P_x	$P(x_1)$	$P(x_2)$	$P(x_3)$	$P(x_4)$	$P(x_5)$	$P(x_6)$
$\boldsymbol{x}$	x_1	x_2	x_3	x_4	x_5	x_6

and

Z	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6
P_y	$P(y_1)$	$P(y_2)$	$P(y_3)$	$P(y_4)$	$P(y_5)$	$P(y_6)$
$\boldsymbol{y}$	y_1	y_2	y_3	y_4	y_5	y_6

The probability of the simultaneous occurrence of the values of $\tilde{x}$ and $\tilde{y}$ is called their joint probability.

The random variables, $\tilde{x}$ and $\tilde{y}$ are said to be statistically independent, if (Welsh 1970)

$$P(x_i, y_j) = P(x_i)P(y_j), \quad i, j = 1,2,...,n$$

otherwise they are statistically dependent. The joint probability of $\tilde{x}$ and $\tilde{y}$ can be also written in the form

$$P(x_i, y_j) = P(x_i) r_i(j),$$

where $r_i(j)$, i,j =1,2,...is the set of the conditional probabilities that the random variable $\tilde{y}$ assumes its j-th value given the i-th value of the random variable $\tilde{x}$. A measure of the statistical dependence $\tilde{x}$ and $\tilde{y}$ should have the following reasonable properties:

(i) It should be equal to zero if $\tilde{x}$ and $\tilde{y}$ are statistically independent.

(ii) It assumes its maximal value if $\tilde{x}$ and $\tilde{y}$ are functionally dependent.

(iii) It should be a scalar function of components of the probability distributions of $\tilde{x}$ and $\tilde{y}$

A measure which has the above requested properties is the ratio

$$R_{ij}(\tilde{x}, \tilde{y}) = \frac{P(x_i, y_j)}{P(x_i)P(y_j)} = \frac{P(x_i) r_i(j)}{P(x_i)P(y_j)},$$

where the marginal probabilities $P(x_i)$ and $P(y_j)$ are defined as

$$P(x_i) = \sum_{j=1}^{n} P(x_i, y_j) \text{ and } P(y_j) = \sum_{i=1}^{n} P(x_i, y_j)$$

A suitable mathematical quantity for the measure of the statistical dependence of $\tilde{x}$ and $\tilde{y}$ is the *logarithm* of $R_{ij}(\tilde{x}, \tilde{y})$. The condition for the statistical independence of $\tilde{x}$ and $\tilde{y}$ implies

$$\log[R_{ij}(\tilde{x}, \tilde{y})] = 0 \quad \text{for every } i \text{ and } j.$$

Any deviation of $\log[R_{ij}(\tilde{x}, \tilde{y})]$ from 0 means that $\tilde{x}$ and $\tilde{y}$ are in some extent statistically dependent.

The larger this deviation, the larger the statistical dependence of $\tilde{x}$ and $\tilde{y}$. Therefore, the mean value of $R_{ij}(\tilde{x}, \tilde{y})$. is taken as a measure of the statistical dependence of the random variables $\tilde{x}$ and $\tilde{y}$. This quantity is called **information**

$$(6.1) \quad I(\tilde{x}, \tilde{y}) = \sum_{i,j=1}^{n} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} = \sum_{i,j=1}^{n} P(x_i, y_j) R_{ij}(\tilde{x}, \tilde{y})$$

It is well-known that another more familiar moment-like measure for the statistical dependence of two random variables $\tilde{x}$ and $\tilde{y}$ is their coefficient of correlation defined as

$$r(\tilde{x}, \tilde{y}) = \frac{E[(\tilde{x} - E\tilde{x})(\tilde{y} - E\tilde{y})]}{\sqrt{E[\tilde{x} - E\tilde{x}]^2 \, E[\tilde{y} - E\tilde{y}]^2}},$$

where the symbol E stands for the mean value for the expression which follows it. In the theory of probability, it has been shown that the coefficient of correlation has an unpleasant short-coming consisting in the fact that, in contrast to **information**, it is equal to zero in some cases when random variables are even strongly statistical dependent (Vajda 1982). Therefore, the coefficient of correlation does not fulfil the first requirement asked from a measure of the

statistical dependence. On the other hand, information (6.1) becomes always equal to zero if the random variables are statistically independent. From this point of view, the information is a more adequate measure of statistical dependence than the coefficient of correlation.

The information $I(\tilde{x},\tilde{y})$ is closely linked with Shannon's entropies of $\tilde{x}$ and $\tilde{y}$. Let us consider the total entropy of $\tilde{x}$ and $\tilde{y}$

(6.2) $$H(\tilde{x},\tilde{y}) = -\sum_{i,j} P(x_i y_j)\log P((x_i y_j).$$

Eq.(6.2) can be rearranged into the following form (note that $\sum_j P(x_i)r_i(j) = P(x_i)$)

$$H(\tilde{x},\tilde{y}) = -\sum_{i,j} P(x_i)r_i(j)\log P[(x_i)r_i(j)]$$

(6.3) $$= -\sum_i P(x_i)\log P(x_i) - \sum_j P(y_j)\log P(y_j) - \sum_{i,j} P(x_i)r_i(j)\log\frac{r_i)j)}{P(y_j)} \quad .$$

Denoting Shannon's entropies of $\tilde{x}$ and $\tilde{y}$ by the symbols $H(\tilde{x})$ and $H(\tilde{y})$ and the corresponding information by the symbol $I(\tilde{x},\tilde{y})$, Eq.(6.3) can be written as

$$H(\tilde{x},\tilde{y}) = H(\tilde{x}) + H(\tilde{y}) - I(\tilde{x},\tilde{y}).$$

We see that $H(\tilde{x},\tilde{y})$ is given as the sum of entropies of the random variables $\tilde{x}$ and $\tilde{y}$ minus the information $I(\tilde{x},\tilde{y})$. Hence, the information can be expressed as the difference between the total entropy of two random variables and the sum of their entropies , i.e.,

(6.4) $$I(\tilde{x},\tilde{y}) = H(\tilde{x}) + H(\tilde{y}) - H(\tilde{x},\tilde{y}).$$

This shows that information belongs to the same class of the mathematical quantities as Shannon's entropy and it represents its natural complement.

3.2. Measures of the configurational organization

In chapter 3.2, a measure of the configurational organization called the Watanabe measure was introduced. In what follows we rewrite this measure, denoted by the symbol $\mathbf{O}(\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n)$**,** in the mathematical form for a probabilistic system with *n* statistically dependent random variables $\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n$ whose ranges contain *N* numerical values by means of the corresponding entropies

(7.1) $$\mathbf{O}(\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n) = \sum_{i=1,2,...,n} H(\tilde{x}_i) - H(\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n),$$

where

$$H(\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n) = -\sum_{i_1,i_2,...,i_n} P_{i_1,i_2,...,i_n}\log(P_{i_1,i_2,...,i_n})$$

and $P_{i_1,i_2,...,i_n}$ is the joint probability distribution of random variables $\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n$.

$$H(\tilde{x}_1) = -\sum P_{i_1}\log P_{i_1},\ H(\tilde{x}_2) = -\sum P_{i_2}\log P_{i_2},...,H(\tilde{x}_n) = -\sum P_{i_n}\log P_{i_n},$$

are entropies of the individual random variables. Equation (7.1) can be rearranged in the form

$$\text{(7.2)} \quad \mathbf{O}(\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n) = \sum_{i_1,i_2,...,i_n} P_{i_1,i_2,...,i_n} \log P_{i_1,i_2,...,i_n} - \sum_{i_1,i_2,...,i_n} P_{i_1,i_2,...,i_n} \log P_{i_1} P_2,...,P_{i_n} \quad ,$$

where the marginal probability of the *j*-th random variable $\tilde{x}_j$ is

$$P_{i_j} = \sum_{i_1,i_2,...,i_{j-1},i_{j+1},...,i_n}^{n} P_{i_1,i_j,...,i_n}.$$

Eq.(7.2) can be rewritten in a more compact form

$$\text{(7.2)} \quad \mathbf{O}(\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n) = \sum_{i_1,i_2,...,i_n} P_{i_1,i_2,...,i_n} \log\left(\frac{P_{i_1,i_2,...,i_n}}{P_{i_1} P_2,...,P_{i_n}} \right)$$

Similarly, as in Chapter 2, we consider the ratio

$$R_{i_1,i_2,...,i_n}(\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n) = \left[\frac{P_{i_1,i_2...i_n}}{P_{i_1} P_{i_2}...P_{i_n}} \right].$$

The mean value of its logarithm gives the degree of the statistical dependencies of random variables $\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n$. According to (3.2), the mean value of $\log(R_{i_1,i_2,...,i_n}(\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n))$ represents just the Watanabe measure

$$\text{(7.3)} \quad \mathbf{O}(\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n) = \sum_{i_1,i_2,...,i_n} P_{i_1,i_2,...,i_n} R_{i_1,i_2,...,i_n}(\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n)$$

The quantity $\mathbf{O}(\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n)$ represents a generalization of the information between two random variables $\tilde{x}$ and $\tilde{y}$. Hence, $\mathbf{O}(\tilde{x}_1,\tilde{x}_2)$ is also identical with $I(\tilde{x},\tilde{y})$ (see Eq.(6.4)). The Watanabe measure is basically given by the mutual statistical dependencies of random variables which are defined on the elements of a probability system. If these variables are statistically independent, then knowing the value of the random variable $\tilde{x}_i$ does not mean the value of the neighbouring random variable $\tilde{x}_{i+1}$ can be predicted. The orderliness in this system is minimal. If these variables are strongly dependent then, knowing the value of $\tilde{x}_i$ means that the value of $\tilde{x}_{i+1}$ can be predicted with high accuracy, i.e., the orderliness of system is much larger.[12] We see that the quantity $\mathbf{O}(\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n)$ represents the appropriate measure of the configurational organization (orderliness) of a system which fulfils all desired requirements.

The configurational organization of a probabilistic system is a function of the probability distributions of the individual element random variables, which we denote by the symbol **P**, and of the statistical dependencies between them, which we denote by the symbol **R**. That is

$$\mathbf{O}(\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n) = F(\{\mathbf{P},\mathbf{R}\}).$$

[12]Intuitively it applies that the larger the configurational order in a probabilistic system is, the more information its random variables $\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n$ mutually contain. If the values of the random variables of a probabilistic system are *functionally* dependent then it represents a deterministic system with highest orderliness.

The value of **O** changes with any change in the probability distributions of random variables $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ and/or with any change of the statistical dependencies between them.

Symbolically, we can write

$$\delta \mathbf{O} = F\{\mathbf{P} + \delta\mathbf{P}, \mathbf{R} + \delta\mathbf{R}\} - \mathbf{F}\{\mathbf{P}, \mathbf{R}\}.$$

A classical system with the configurational organization is the so-called Ising system (Majerník, 2001).

4. Goal-directed systems

4.1 Introduction

The systems whose system properties are prescribed a-priori form the class of goal-directed systems. This class of systems as well as that of the probabilistic systems can be described in the frame of general systems theory. While the interelement relations in a probabilistic system are given by their statistical dependencies, expressed in the form of conditional probabilities, the interelement relations in a goal-directed system are given by their goal and may be communicative, social, etc. In spite of the fact that a goal-directed system has a prescribed goal, it often achieves it not fully but only partly. The degree of which a goal-directed system reaches its goal depends on the internal structure and interelement cooperative interaction. There is a difference between the 'ideal' and 'real' values of the organization of a goal-directed system, and our task is to find an appropriate measure for this difference.

There are many goal-directed systems with functional organization in various areas of science and technology. The common property of all these systems is the cooperative interaction among their elements in achieving their goal and functions. In order to describe the goal-directed systems with functional organization in the framework of general systems theory some new concepts have to be introduced. The first concept is the **functional connection** among the elements of a goal-directed system. If one can establish that a cooperative interaction exists between two or more elements, then these elements are functionally connected; otherwise they are functionally disconnected. If elements of a goal-directed system are functionally connected then they can exhibit different degrees of the **cooperative dependence** (a notion which is analoguous to that of the statistical dependence in the configurationally organized systems) with respect to the goal of the system. Elements of a goal-directed system can be functionally connected by many ways. The actual way in which the elements of a goal-directed system are functionally connected is called its **functional organization**. The problem with the functional organization is its quantification, i.e., the determination of the degree in which an actual system fulfils its goal and function.

Below we present an elementary theory of functional organization based on the foregoing concepts. Since the meaning of the term "goal" may be rather vague, we shall use it only in that case when it can be expressed in the form of exact mathematical requirements im-posed on certain system parameters (e.g. the system should reach certain prescribed values of some of its system parameters). A goal-directed system is given by the set of its elements (parts or units) which is called its **basic set** $\mathbf{S} \equiv \{s_1, s_2, s_3, ..., s_n\}$. The functional connection of the elements of **S** means that there exists an interactive behaviour (e.g. energetic or informational) among them with respect to the goal of the system. Some of the elements of the basic set are functionally connected (i.e., there is an interaction between them), others are functionally disconnected (i.e., there is no interaction between them). The set of relations $r_i(j)$

between the elements of a system indicates the degree of their cooperative dependencies.[13] There are several ways of how to express the structure of a goal-directed system in the theory of general systems, e.g. by means of the algebraic incidence matrices. We do it by means of binary oper-ations between these basic elements. The symbol $\cup$ between the elements means that they are functionally connected, while the symbol $\subset$ between them means that they are functionally disconnected.[14]

4.2 Cooperative interaction in a goal-directed system

The functionally connected elements can have different degrees of cooperative interaction with respect to the goal of a system. Sometimes, such elements affect the organizational function only insignificantly, sometimes they do it considerably. Hence, the functionally connected elements can have different degree of *cooperative dependence* with respect to the goal.[15]. The cooperative dependence of two elements of a goal-directed system is defined by the following way: Let Ψ be a general finite and single-valued positive organizational function defined on the subset of the basic set of a goal-directed system containing two elements s_1 and s_2 which are functionally connected or disconnected.

The elements s_1 *and* s_2 *are said to be cooperatively independent if*
$\Psi(s_i \cup s_j) = \Psi(s_i \bigcirc s_j)$:
otherwise they are cooperatively dependent.

The considered catalytic system has only two elements (s_1 chemical reactants, s_2 additive substance). If s_2 is a catalyst or anticatalyst the elements s_1 and s_2 are cooperatively dependent, since it holds

$$\Psi(s_1 \cup s_2) \neq \Psi(s_1 \bigcirc s_2).$$

When s_2 is a neutral substance, the elements s_1 and s_2 are cooperatively independent. We see that the cooperative dependence expresses the measure of mutual cooperative interaction of

---

[13] As an example of a simple system which is functionally organized, we consider a system in which a catalytic chemical reaction takes place, i.e., a chemical reaction occurring in the presence of a chemical substance (e.g. catalyst or anticatalyst) which may affect (accelerates or decelerates) it. Two reactants and the chemical substance form a two-element system whose goal is to reach the largest reaction velocity. This simple system can demonstrate many typical properties of an organized system. The reactants and the substance can be separated (i.e., elements of this system are functionally disconnected) or brought together (i.e., they are functionally connected). Since the goal of this system is to perform the chemical reaction in the shortest possible time, the following three cases are to be distinguished if reactants and the substance are functionally connected. They are: reaction time is (i) the same, (ii) shorter, (iii) longer **than** in the case when the reactants and substance are functionally disconnected. In the first, second and third cases, the system exhibits no organization, it is well-organized and it is disorganized with respect to its goal, respectively. Mathematically, the reactant and substance form a goal-directed system whose basic set S has two elements: (i) two reactants s_1 and (ii) the chemical substance s_2. These elements can be functionally connected or disconnected depending on whether the reactants and substance are separated or brought together.

[14] Denoting the reaction time of the chemical reaction by the symbol T the following three cases occur: $T(s_1 \cup s_2) = T(s_1 \bigcirc s_2)$, $T(s_1 \cup s_2) > T(s_1 \bigcirc s_2)$ and $T(s_1 \cup s_2) < T(s_1 \bigcirc s_2)$. The value of T for the considered system is determined by the kind of the functional connection between its elements. We refer to this function as the **organizational function** of a goal-directed system.

[15] The value of organizational function T in our example is the same in case (i) when the elements s_1 and s_2 are functionally connected or disconnected. Here it is already intuitively clear that these elements are cooperatively independent. In both the latter cases (ii) and (iii) they are cooperatively dependent

two elements of a system independently of whether this cooperative dependence is favourable or unfavorable with respect to the goal of system.

The concept of cooperative dependence can be easily extended also for systems comprising more than two elements. Let $s_{1,}\ s_{2,\ldots,}\ s_{\mathrm{k}}$ be elements of goal-directed system. If the organizational function Ψ defined on the set of elements $s_{1,}\ s_{2,\ldots,}\ s_{\mathrm{k}}$ fulfils the condition

$$\Psi(s_1 \cup s_2 \cup s_3 \cup \ldots \cup s_k) = \Psi(s_1 \bigcirc s_2 \bigcirc s_3 \bigcirc \ldots \bigcirc s_k).$$

then the elements $s_{1,}\ s_{2,\ldots,}\ s_{\mathrm{k}}$ are cooperatively independent; otherwise they are cooperatively dependent.

One can construct different measures for the cooperative dependence of elements of a goal-directed system. The simplest measure represents the absolute value of the difference between the values of the organizational function for the functionally connected and disconnected elements $s_{1,}\ s_{2,\ldots,}\ s_{\mathrm{k}}$, i.e.,

$$\epsilon(s_1, s_2, \ldots, s_k) = |\Psi(s_1 \cup s_2 \cup s_3 \cup \ldots \cup s_k) - \Psi(s_1 \bigcirc s_2 \bigcirc \ldots \bigcirc s_k)$$

A system with cooperatively dependent elements can have different structures according to how its elements are functionally connected. A value of the organizational function is assigned to every actual functional connection of the basic elements. The **actual** manner in which the functional connections among the elements of a system are realized, is called its **functional organization**. Here, the functional organization of system is given by means of the structure of the functional connections among its elements. Let us denote one actual way of the functional connections between the elements of a goal-directed system by the symbol σ. The set of all possible functional connections of a system we denote by the symbol *D*. If we further denote by the symbol *N* the state of a system in which all its elements are mutually functionally disconnected, then three subsets of *D* are to be distinguished (*g* is the goal value of the system):

(i) the subset D_d of *D* whose elements satisfy the inequality $(\Psi(\sigma)\text{-}g) > (\Psi(N)\text{-}g)$. The corresponding systems are disorganized with respect to the given goal.

(ii) the subset D_z of *D* whose satisfy the inequality $(\Psi(\sigma)\text{-}g) < (\Psi(N)\text{-}g)$. The corresponding systems are well-organized with respect to the given goal.

(iii) the subset D_n of *D* whose elements satisfy the equation $(\Psi(\sigma)\text{-}g) = (\Psi(N)\text{-}g)$.. These systems have no functional organization with respect to the given goal.

The basic concepts of the theory of functional organization can be summarized as follows. (Majerník 1985):

> [1] A goal-directed system is given by the sets **S** and *D*.
> [2] Two elements s_i and s_j are *functionally* connected if a cooperative interaction between them is established. Functional connection or disconnection of two elements are denoted by the symbol ∪ and ○, respectively .
> [3] Functionally connected elements exhibit different degrees of the *cooperative dependence* with respect to the goal of system.
> [4] The function which assigns the magnitude of the desired system property of a goal-directed system to the actual functional connections of system elements represents the organization function and the actual way, in which the functional connections among the elements of a goal-directed system are realized, represent its functional organization.

Hence, we need for the fully characterization of the functional organization of a system Σ the following data: (i) the goal value g of Σ; (ii) the elements of its basic set, $\mathbf{S} \equiv \{s_1, s_2, \ldots, s_n\}$;

(iii) set D of all functional connections among the elements of the basic set; (iv) the organizational functions Φ defined on elements *D*.

There are many examples of goal-directed systems in science and technology, especially in biology. A simple system is the above mentioned catalytic system containing two reactants and a substance affecting the rate of the chemical reaction. A more complicated system arises when - among these simple catalytic systems - further cooperative dependencies are established. Each such catalytic subsystem produces the reactant or catalyst for the next catalytic system. Such a system is e.g. the Krebs cycle, well-known in bioenergetics. Another system of this kind, involving simple RNA and DNA, called the hypercycle, was studied by Eigen and Schuster. The most sophisticated cooperative interaction between the system elements occurs in the cell. The characteristic feature of the complex biological catalytic reactions in the cell is an interplay between the information carriers (DNS) and the active substances (proteins). Such an interplay between the information and action is also typical of other goal-directed systems occurring in science.[16]

4.3 The goal-directed string systems

Except of some systems with the abstract basic elements, all the goal-directed systems consist of a collective of physical objects forming their physical backgrounds. As an example, let us mention music which can be seen as a sequence of music signals with the goal to generate sensory-aesthetic perception in music listeners. The physical background of music forms the acoustic signals. The sequence of these signals can be seen as a purely statistical system with only a configurational organization. A robot which has no aesthetic feeling would be considered by a music sequence as a system only with configurational organization and it would measure it by its entropy. Generally, it is believed that systems with a large functional organization are also in a state characterized by high order and, hence, by high configurational organization. However, the high configurational organization of a system does not automatically imply that it has a functional organization. Hence, the interrelation between functional and configurational organization in an actual system appears to be one of the central problems in the theory of organization.

To illustrate this problem we study a simple system which consists of one-dimensional strings of elements representing its alphabet. The most important string systems are (i) the human language with an alphabet of about 50 letters (graphemes in the written form or phones in the spoken form), (ii) the classical music with an alphabet of about 150 letters (notes), (iii) the nucleic acid (DNA) with an alphabet consisting of four letters and (iv) proteins having an alphabet of 20 letters (amino acids). Besides the Watanabe one the generally accepted measure for the configurational organization is redundancy defined as

$$R_n = 1 - \frac{H_n}{(n+1)H_0},$$

where H_n is the entropy of the n-dimensional frequency distribution and H_0 is the entropy of the uniform frequency distribution. A large class of string systems are goal-directed and perform certain functions. A string of graphemes which represents a word in a given language represents a functional string in this language. The set of all functional strings of phonemes

[16] A classical example of such interplay is a simple goal-directed system containing the so-called Maxwell demon which is capable of separating hot and cold gas molecules in a gas cylinder. Its goal is to decrease the entropy of a thermodynamic system by means of an interplay of information gaining and mechanical action.

constitutes a spoken language. A sequence of musical notes which has an aesthetic content is a functional string in music. A string of four elements in genetic DNA is functional if it carries useful information for the activity of a cell. A string of amino acids is functional if it forms a functional protein. We can determine the configurational organization of music, language and other systems realized by strings if we know the statistical distributions for their letters by means of their Watanabe's measure for configurational organization or by means of their redundancy.

Which of the sets of all possible strings are functional depends on the a-priori goal of the string system. Let us next consider certain systems in which strings of different lengths occur. We denote the number of all possible strings of the length n and the number of the functional strings of length n resp. by the symbols S_n and Z_n. An appropriate quantitative characteristic of the system's effective use of its letters is called its *letter utility* and is given by the formula (Majerník, 1998)

$$K = \frac{1}{\ln 2} \ln \left(1 + \frac{Z_n}{S_n}\right).$$

This measure gets its values from the interval [0,1]. If **all** possible strings of the length n are functional, the system has letter utility $K = 1$, if a system has no functional strings its letter utility is $K = 0$. A string system which expresses the numbers whose elements are figures has utility $K = 1$, since any string of figures represents a number. In linguistics the functional strings of the length n represent only such strings that express the words, therefore its letter utility is $K < 1$. Although the number of functional strings are not at our disposal we can already estimate that the string system of proteins has very small letter utility. This points to the fact that the biological structures are extremely complicated as compared to the systems generated by man.[17] The relation between the value of the functional organization and configurational organization is different for different string systems and must be determined separately for each individual system. Intuitively, one already has a feeling that systems with a strong functional organization are also in a state characterized by high value of configurational organization. Systems with the strong configurational organization are in a low-entropy state therefore their orderliness tends to decay quickly often destroying their functional organization as well.

5. Some examples of functionally organized systems

5.1 Living objects

Life as a whole, from the simplest organisms to the most complicated ones - is a long sequence of counterbalancing with the environment, a sequence which becomes eventually very complex. The time will come - even if it is still rather far away - when mathematical analysis supported by scientific analysis can put this counterbalancing into majestic formulas of equations, thus finally putting life itself into an equation.
I.V.Pavlov

As said in the introduction, the functional organization is related to the goal of a system, therefore to interpret the biological objects as goal-directed systems it is necessary to determine their goal with respect to which they are organized. As biological objects generally consists of many subsystems, it is useful to distinguish between their **top goal** and the **subgoals**

[17] Generally, the probability to find a functional string among all possible strings is the smaller, the smaller the corresponded letter utility is.

of their subsystem. We assume that the top goal of a system is that goal from which all subgoals of its subunits can be derived. We assume, for the time being, that a top goal of biological objects exists and we will try to find it[18] by means of the so-called functional isomorphism between the information-conserving systems and biological objects.

5.2 Functional isomorphism of two systems

To treat living objects one has to find another model system that is functionally isomorphic with it. According to Majerník (1987) two types of isomorphism of goal-directed systems are to be distinguished: (i) the strong isomorphism and (ii) the simple functional isomorphism. Two systems $\Sigma_1 = [g_1,S_1,D_1,\Psi_1]$ and $\Sigma_2 = [g_2,S_2,D_2,\Psi_2]$ are then *strongly isomorphic* if there exists a one-to-one mapping F between them such that:

$$F : g_1 \to g_2; \quad F_2 : S_1 \to S_2; \quad F_3 : D_1 \to D_2; \quad F_4 : \Phi_1 \to \Phi_2$$

Any goal-directed system which is strongly isomorphic with another system represents its *model*. Two systems are then **simply functionally** isomorphic if there exists a one-to-one mapping only between *their goals*. Many goal-directed systems with a hierarchical structure possess a top goal from which the goals of their subsystems can be derived. We will try to show that there exists a *functional isomorphism* between the biological objects and the goal-directed system called *information-conserving device* which represents its model. An information-conserving device is a system with the goal to conserve and/or to accumulate its internal information content in the course of time.

It is relatively simple to find an isomorphism between the subgoals of subunits of an information-conserving device and those of a living object. Due to a lack of clear definition of the top goal of living objects[19] we assume that the information-conserving device and a living object are *functionally isomorphic* with regard to their top goals. Hence, we take the top goal of living objects, the conservation and the accumulation of their internal information contents, against the destroying effects by entropy-decreasing processes occurring in them and in their environments.

5.3. Physical objects (signals) that are able to carry information

Any physical medium capable of establishing a communication linkage between the information source and receiver is called **signal**. A signal can have different realizations, it can e.g. be an ensemble of photons or a modulated physical field. From the mathematical viewpoint, a signal represents a stochastical object whose states occur with different probabilities. The more information a signal carries, the more complicated its physical structure is as to space and time. Generally, complicated physical structures are mathematically improbable and physically unstable so that a signal carrying a large amount of information is necessarily in a low-entropic state. Due to this fact, an isolated signal tends, in course of time, to decay through the increase of its internal entropy by a simultaneous decrease of its information content. The physical processes decreasing the information content of a signal are the

[18]The existence of a top goal of living objects is by far not evident and we must admit the possibility that such a top goal does not exist at all.

[19]Referring to living objects one can vaguely speak of survival as their top goal, however, without any specification of what it exactly means.

following: (i) energy dissipation due to the decrease of signal's intensity. (ii) the energy and entropic currents within the signal itself due to its unstable and non-equilibrium state. (iii) deformation and destruction of signal due to the interaction with its environment.

To protect a signal against the said processes the use of a device, linked with it, is needed. This, so-called information-conserving device, has the following subunits: (i) An energy supplying and amplifying mechanism compensating the dissipation of signal's energy (amplifying subunit). (ii) A stabilization mechanism being able to conserve signal's structure (stabilizing subunit). (iii) A mechanism being able to gain information of signal's environment in order to protect it against the external destruction (information-gaining subunit). The amplifying and a stabilization subunits of the information-conserving device are aiming at the conservation its energy and its spatial and temporal form. The information gaining subunit is aiming to acquire information on the environment of the signal and to utilize it for protecting the said signal against the external destructive effects.

Despite all the mentioned activities, the information-conserving device generally cannot completely prevent the decay of a signal and decrease its information content. To maintain the information content of a signal over large time period an information-conserving device, coupled with it, shall, in addition, contain a replication subunit. This makes it possible to generate a new signal with the same information content as its original.

Let us now compare the goals of the individual subunits of information-conserving device with those of living object. It is well-known the basic activities of a living object are: (i) metabolism, (ii) homeostasis, (iii) information-gaining and protection and (iv) replication. The corresponding activities of the subunits of an information-conserving device are: (i) the energy supplying and amplifying subunit, (ii) the stabilization subunit, (iii) the information-gaining and protecting subunit and (iv) the replication subunit. A functional isomorphism between the subunits of living object and those of information-conserving device is indicated in the following table which shows a one-to-one correspondence between the activities (subgoals) of subunits of biological objects and those of information-conserving devices.

Metabolism	Energy supplying and amplifying mechanism
homeostasis and control activity	stabilizing mechanism
information gaining	information-gaining mechanism
replication	copying mechanism

Since a one-to-one mapping exists between the (sub)goals of subunits of an information-conserving device and those of living objects, living objects may be, at least in a certain approximation, considered as goal-directed systems functionally isomorphic with the information-conserving devices. In both cases the top goal is the conservation of their information contents. Accordingly, the organization of biological objects can be considered as functional and *modelled* by the information-conserving device.

When taking the above postulated top goal of living objects, we can also determine the degree of their functional organization with respect to it, although it is an immense complicated task even for the very primitive organism. Nevertheless, the fact that the biological organization may be formalized within the frame of general systems theory, makes it possible to understand quantitatively many biological processes being hitherto described only quantitatively. Especially, the *biological evolution* can be understood as a gradual increase of the functional organization of biological objects by a gradual increase of their internal inform-

ation contents.[20]

Let us make some comments regarding the term 'information content' of a signal. It is well-known that the information content of a signal can be seen from different points of view. The amount of the selective information of a signal is given by the content of Shannon's information carried by it. The selective information carried by a signal may be useful, neutral or just an impediment with respect to the goal of a system. The part of information of a signal which is useful for reaching the goal is called the *pragmatic* information. Two signals can carry equal amounts of Shannon's information but with different pragmatic values for the different goal-directed systems.

The tendency to increase the entropy of a signal as a physical object may change the character of the total information carried by it. This increase may change the ratio between the prag-matic and selective pieces of information of a signal by the conservation of whole of Shannon's information (effect of the first type) or it can more or less erase a part of information carried by sign-al (effect of the second type). This can be explained best by taking a printed text as an example. A printed text represents a string of letters. Due to effects of the first type, some letters can be randomly changed transforming the original text to a new one by letting it readable. The effects of the second type consist of a more or less comprehensive erasure of the text by making it unreadable. While the effects of the first type generally change the pragmatic content of information only, effects of the second type decrease the content of Shannon's information. Both processes occur in living objects well. The effects of the first type change the ratio between the pragmatic and selective pieces of information in living objects. The effects of the second type erase a certain part of the Shannon information in them. From this point of view, the biological evolution can be understood as a process in which the biological objects increase their pragmatic information by means of the natural selection.[21]

5.4. Language

Language is recognized as a codifier of thought and a means of communication. From the point of systems theory language can be seen as a (i) purely statistical system, (ii) a goal-directed system and (iii) a synergetic system (Haken 1989). Language as a statistical system is determined by the statistical distribution laws. The statistical distribution laws of its elements are generally well-known in statistical linguistics, e.g., the distribution of word frequencies follows the **Zipf-Mandelbrot** law for nearly all texts in any languages (Orlov, Boroda and Nadarejšvili, 1982). String lengths of words, phrases, sentences are known to follow certain distribution, too (Naranan, Balasubrahmanyan,1992). The frequency distributions of graphemes and phonemes have been determined for many languages. Therefore, there is enough statistical data for determining the degree of the **configurational organization** of languages. It is known that the average information content of graphemes employed for writing a text in a given language is defined as the entropy of the grapheme distribution and is expressed in bits. For all languages whose alphabet has a number of graphemes of between 17 (= 2^4+1) and

[20]It is, in principle, also possible that the biological goal in its whole generality cannot be found by means of the foregoing procedure. If the biological objects had no top goal, then their organization could not be functional in the sense we considered, and it could not be formalized within the framework of the existing general systems theory. Biological objects would, then, represent more complex and specific objects, which are not hitherto even described.

[21]There is an essential difference between the gaining of pragmatic information in physical and biological objects. In physics the only source of information is measurement and in biology the ultimate source of new information is natural selection.

32 (= 2^5) as e.g. English, French, German or Spanish, 5 bits[22] are approximately the maximal value of the information carried by a grapheme. If all graphemes of a language had the same frequency in text, the entropy would be, according the known definition $H_0 = \log_2 N$, where N is the number of different graphemes used in the language. H_0 represents the so-called zero order entropy. The real value H of the entropy is less than H_0 due to the fact that it is not allowed to follow arbitrary successions of graphemes. Further approximate values of H can be obtained by computing the effective frequency of the various graphemes in texts, that is calculating the expression

$$H_1 = -\sum_{n=1}^{N} p_n \log_2 p_n,$$

where p_i is the probability of occurrence for a given grapheme. The entropies of n-grams of higher order can be expressed by the formula

$$H_n = -\sum_{j,k,m,\ldots,s}^{N} p_{ijkm\ldots s} \log_2 p_{ijkm\ldots s}$$

where $p_{ijk\ldots n}$ is the probability of occurrence of n-grams in a language.

What has been said so far is connected with the configurational organization of language. Relatively little attention has, however, been paid to the functional organization of languages. To characterize language as a functionally organized goal-directed system we have to determine its goal, its basic set, the set of all functional connections and the organizational function. The goal of language is to transmit information. the basic set form the graphemes, words, sentences according to the hierarchic level we shall consider. The organizational function is given as the account of information transmitted in time (or space) unit.

The relation between the functional and configurational organization can be estimated by means of letter utility of a language. The individual elements of these systems can be functionally connected so that they form strings. Some functionally connected elements are cooperatively dependent with respect to the goal of language to carry the relevant information and form words, phrases and sentences, i.e. they represent the functional strings. The other elements are cooperatively independent and do not transmit information. The ratio of the number of the functional strings of the length n to all strings of length n determines the letter utility of a language system. So far, this ratio has not been determined for any language.

Language can be considered as a synergetic and open system. According to Haken (1989), synergetics is an interdisciplinary field of research that is concerned with the spontaneous formation of structure in system far from thermal equilibrium as well as in non-physical systems. The central concepts of synergetics are stability, order parameter and the slaving principle. The paradigms developed in synergetics can be applied also in human sciences (Haken 1989). These paradigms and concepts have been applied by Altmann and Köhler to language in the framework of synergetic linguistics (Altmann, Köhler 1986; Köhler 1990). This yields new view on language as an open system interacting with its performers by means of the slaving principle (Köhler 1990).

An other aspect of speech communication shall be here mentioned. A natural, or controversial, speech represents physically a stochastic sequence of a set of different complex sounds generated by the flow of air through the human vocal tracts. Due the nonstationary character of these sounds, the speech recognition represents a complicated physical and mental process by which the artificial intelligence research has to be employed. In spite of a

[22] Units of the Shannon entropy in the logarithm to the base two.

great effort, in this field, the results of this research are still rather poor because the majority of researchers consider language as purely statistical system organized configurationally. The speech recognition problem would be essentially progressed if one would consider language as a goal-directed system with functional organization.

5.5. Music as a functionally organized system

Musik ist Sprache, wo Sprachen enden
R. M. Rilke

A piece of music represents a sequence of musical elements with the goal to generate sensory-aesthetic perception in the music listener. The elements of classical music are the acoustical signals that are quantified as musical sounds, i.e., those sounds that are smooth, regular, pleasant and of the definite pitch. This is not a good classification of acoustical signals for the modern music where the physical parameters do not play the fundamental role any more. Here, all sounds which produce aesthetic sensation are considered musical. The set of all musical signals forms the basic set of music as a system. A basic problem in treating music as general system is the determination of its alphabet. Many scientific studies have been devoted to finding the psychological and physical sign systems of music. Physicists have studied the structure of the complex sounds. The psychologists have tried to quantify the sound perceptions. In spite of these studies the determination of an alphabet of music is still an open problem. The organization function is determined by the aesthetic evaluation of the pieces of music.

The functional connections of musical elements can be either melodic or harmonic. The melodic connections of musical elements form the melodic phrases of a piece of music, the harmonic connections of the musical elements form its harmonic structure. Taking the *redundance* as the measure of the configurational organization of music shows that different arts of music pieces have different degree of their configurational organization. Redundancy in Gregorian song is about 20 %, redundancy in Slovak folksongs is about 25%, that of German romantic music about 14%. Surprisingly, redundancy in atonal music composition (dodecaphony) is relative high, i.e. about 33%. This is due to strong composition rules of the dodecaphony (Majerník 1970). The letter utility, as a measure of the functional organization of music, has not been analyzed yet.

5.6 Biogenesis

Wer wenig weiß muß viel glauben.
Marie von Ebner-Eschenbach

We all are believers, a part of us believes in „biological Big-Bang" directed by God, the other part of us believes in the immesely improbable random origin of life.
J. Monod

The question of life's origins is one of the oldest and most difficult. Its answer, should it ever be known, will not be a single statement of fact but rather an extended chronology. Since life must have emerged from the inorganic matter the fundamental question arises as to how to rationalize this process, i.e., the question as to how the originally purely physical systems which had only configurational organization have finally led to systems having a very high functional organization? In other words, how could these primordial living systems have

originated in a functionally unorganized medium (Kuhn 1988)?[23] The theory of the biogenesis represents a typical multidisciplinary chain at the beginning of which there are the physico-chemical processes, foremost the processes far from equilibrium, then a chemical evolution and at the end the simplest organisms. Most bioscientists believe today that life emerged on Earth at the early stadium of its geophysical evolution. According to present view the first prebiotic molecules on Earth originated from constituents of the atmosphere, hydrosphere and lithosphere through physico-chemical processes that were propelled by lightning, heat, solar and cosmic radiations. These prebiotic molecules are thought to have interacted in condensation reactions and in processes of *self-organization* to yield macromolecules, supramolecular structure (for example, membranes) and, finally, early precursors of living cells (protobionts or protocells). These earliest and as yet completely unknown forms of life supposedly evolved into the postulated progenotes. The progenotes finally are considered to be the first regular cellular systems, from which three kingdoms of contemporary organisms (archaebacteria, eubacteria and eukaryotes) originated by Darwinian evolution.

Considerable disagreements exists among scientists about the detailed evolutionary steps (Dose 1988). The problem is that the principal evolutionary processes from prebiotic molecules to progenotes have not been proved by experimentation and the environmental conditions under which these processes occurred are not known. Moreover, we do not actually know where the genetic information of all living cells originates, how the first replicable polynucleotides (nucleic acids) evolved, or how the extremely complex *structure-function relationship* in modern cells came into existence.

Two types of biomolecules are responsible for the most biological processes, *nucleic acids and proteins* (Parry et al. 1984). The first carry information (and may be considered as the legislative of a cell) while the proteins perform the activity prescribed by the information in the nucleic acids (and may be considered as the executive of a cell) (Frauenfelder 1984). Both substances take part in the biosyntesis a process in which the nucleic acids store and transport information and direct the assembling of proteins. Proteins, assembled from amino acid, are the "machines" of life which can be often modelled by the Maxwell demon. The study of life origin is mostly concentrated on the chemism of nuclein acid and proteins in the early history of Earth. Both substances are important for the biogenesis, however the main point here is how these substances became functionally connected and formed the a highly functional system which is necessary for the origin of a cell. For the time being, nothing is known about any physical or chemical processes which caused this connection to come into being. This is why the only possibility of forming this connection is a certain incidental process with very tiny probability.

The necessary condition for the beginning of the biological evolution is the self-reproduction of the primordial living objects. Here, the main difficulty is that even the simplest form of the self-reproducing apparatus of a cell seems so sophisticated that it could have been formed only by a chain of extremely rare and improbable physico-chemical processes. This self-reproduction of living objects is an absolute prerequisite for the beginning of biological evolution.[24] An interesting problem linked with the functional organization of a goal-directed system is that of the interrelation between the value of functional organization and its physical entropy. It is well-known that the physical entropy of a system is low when the statistical dependencies between its elements are large. As we have shown so far, a large functional

[23]The problem of the life origin also has an ideological context. This can be shown by the fact that the famous and popular book of Nobel laureate J.Monod (1971) was not published in communist countries. The reason was that it contained some claims about the origin of life which do not comply with the Marxist doctrine.

[24]A remarkable property of any organism is its biochemical unity. By the unity of biochemistry we mean the fact that any organism, whether a pineapple or an elephant, has the same amino acids in its proteins.

organization often also generates the large statistical dependencies between its elements and, therefore, it is in a low-entropy state. On the other hand, the second law of thermodynamics tells us that a thermodynamic system tends to assume the state of its maximum entropy. This tendency necessarily also causes the decay of the statistical dependencies in the underlying physical system. Since many goal-directed systems are realized by means of physical systems, the tendency to increase their entropies leads to the decay of the cooperative dependencies between their elements needed to perform their functions. To resist this tendency any functionally organized physical system must have some entropy decreasing mechanism. In a living object this function is mainly seized by enzymes. However, despite the best endeavour of all the entropy-decreasing devices acting in an individual living object it is not generally possible to maintain its low-entropic state for a long time. This means that the statistical dependencies and, in consequence the functional dependencies as well, among the individual elements of a living object are gradually destroyed. The same holds also for other goal-directed systems, especially, for those whose basic activities are realized by the physico-chemical processes as it is the case in living objects.

6. The problem with the goal-directedness

"The most surprising of all is that world almost certainly has a meaning"
A. Einstein

"For my part, I believe that the aim of science should be to show that no feature of the Universe is accidental"
W. Sciama

"Now I put the question: is the creation of intelligent beings the ultimate goal of the cosmological and biological evolution? "
W. J. Brand

Philosophers are chewing continuously without having anything in their mouths
A. Einstein

There is a fundamental problem with the classification of systems, namely what are the criteria for recognizing which of the systems are goal-directed. In principle, some purely physical systems could be also considered goal-directed, e.g. a thermodynamic system, according to the second principle of thermodynamics, tends to the state of its maximum entropy. There-fore, one could consider even this system as a goal-directed with the goal to reach the maxi-mum entropy state. However, intuitively one feels that to consider a physical system as goal-directed does not correspond to the generally accepted concept of functional systems. Which of systems is really goal-directed is a very difficult and perplex issue closely linked to the so-called ’structure-function’ problem.

One of the oldest structure-function problems that man has been speculating about since the very beginning is: How is matter related to mind? Hitherto, this problem has mainly been investigated in philosophy. Its philosophical considerations have usually proved fruitless, in the sense that it does not, by itself, lead to any advances in science. In the modern times, due to the progress of the exact sciences, this problem has again moved from philosophy to the exact sciences. The physical laws were quantified earlier than the teleological ones. These laws formed certain paradigms which are fully formalized and can be expressed in an exact mathematical language. Therefore, some teleological scientists also attempt to apply these paradigms for the description of the goal-directed systems. This holds especially for biologists who are seeking the paradigms appropriate for the description of living objects. However, such a paradigm transfer may often be misleading because physical sciences investigate mat-

erial objects governed by natural laws while the teleological sciences are foremost interested in the description of how a goal-directed system fulfills its goal. The central concept in the physical sciences is orderliness, whereas in the teleological sciences it is function (Woodfield 1976). No goal-directed system can be fully described without taking into account its goal and function, and the physical laws here represent only certain constraints for behaviour of the system.

A most fundamental 'structure-function' problem is the nature and origin of the living objects. This is why all existing systems with functional organization are either biological objects or objects and entities created by man (machines, cognitive robots, computers, languages, cultures, etc.). The transition from the configurational organization to the functional one during biogenesis represents the *single* known case of an autonomous transformation of the purely physical system to the goal-directed one.

In previous chapters we tried to postulate a possible top goal for living objects and characterized the biological evolution as a functional process. The question whether the cosmical evolution follows a goal as well is nowadays intensively discussed. A new light on this problem was thrown by the so-called anthropic principle of cosmology (Gale 1981). This principle claims that the universe is physically 'constructed' in a such way that the intelligent beings could develop. This sounds like a metaphysical argument but the anthropic principle appears as a constructive instrument in cosmology which helps to find the physical structure of the universe. The appearance of intelligent beings in the universe is conditioned by many special physical properties and conditions of the universe. This is because the cosmological evolution is very sensitive to the values of certain natural numbers called physical constants. Already extremely small changes of their values would lead to the universes in which life could not develop.[25] The anthropic principle 'translated' into the language of general systems theory means that the universe represents, in a sense, a goal-directed system.[26] Since, according to the anthropic principle a universe should have such physical conditions that intelligent beings can develop, we have a criterion to distinguish the "functional" universes from the set of all possible universes. Here, it is a similar situation as e.g. in linguistics where only those strings of letters are functional which represent words of a language, i.e., those strings which have "meanings". Accordingly, the assumption that our universe is a goal-directed system is equi-valent with Einstein's claim that our universe has almost certainly a 'meaning'. In spite of the fact that the anthropic principle gives a criterion for which universe is functional we cannot determine the corresponding 'letter utility', as it is the case in a language, because we do not know the number of all functional universes as well as all possible ones.[27]

---

[25]The anthropic principle was introduced by Robert H. Dicke of Princeton University in 1961: he proposed it in the course of analysing work done by P. A. M. Dirac some 30 years before.

[26]Theory of organization threw a new light on the vague word "meaning". In the theory of the functional organization, one can speak of a "meaning" of an object or a process if it has a relation to a system with a goal. Whether an object or a process has a meaning is a problem recently intensively discussed. Einstein claimed that the universe has almost certainly a meaning. On the other side, Weinberg (1993) said that the more we understand the universe the less we find a meaning of it. Translated this in the language of the theory of organization, it means that the universe is well-organized with regard to its configurational organization but possesses no functional organization. Recently, according to the anthropic principle of cosmology, the universe is seen as a system organized such that an intelligent observer arose. In view of the above mentioned, it would have a meaning.

[27]There are cosmological theories which assume that at the big bang, except of our universe, a series of so-called baby universes was created. If we knew they number we could determine the 'letter utility' also for the set of possible universes.

7. Epilogue

Gott würfelt nicht.
A. Einstein

Gott würfelt **doch**.
N. Bohr

Nothing is living in the cell only the cell in its totality.
L. Cuénot

One of the most amply discussed problems in science is the relationship between the configurational and functional organization. This is also reflected in the fields of interest of physical and teleological sciences. An important task of the physical sciences is the unification of some seemingly unrelated natural phenomena. This unification represents a permanent endeavour to reduce the description of variety of the natural phenomena to a few fundamental entities, ideas or propositions which have a simple logical and mathematical structure like e.g. geometry, field, motion and randomness. This unification took place several times in the history of physics. Let us mention only one example, namely the unification of gravitation with the geometrical structure of space-time in Einstein's general relativity. In fact, each physical discipline has a specific and typical entity, framework or the underlying mathematical structure: mechanics-motion, thermodynamics-motion and randomness, electrodynamics-motion and field, theory of relativity-motion and geometry, quantum mechanics-motion, field and randomness. The most important piece of knowledge in the modern description of matter is that nature at its fundamental level is governed by **probabilistic laws**, and **randomness** becomes a central concept of nature.[28]

One of the problems of the teleological sciences, is that it consists in the *reduction* of the phenomena studied in them to those described in physics. Take, for example, biology. Any reduction of biological processes to physical and physico-chemical processes represents a progress in understanding of biological reality. While the important task of the physical sciences is the unification of the phenomena and processes of the material world, the task of biology is the reduction of biological processes to the physical ones. This reduction, when going too far, might lead also to the reduction of many functionally organized systems to those organized configurationally. The extreme reductionism frequently considers a system as a collective of individual elements disregarding its system character. This tendency is especially seen in biology (see e.g. Eigen and Winkler 1975) and even in psychology (see e.g. discussions in the book by Penrose 1997). Here the question arises whether a configurationally organized system can be transformed by itself to a system with function organization. It has been already pointed out that high functional organization of a system is generally connected with its high configurational organization. However, the high configurational organization of a system does not *automatically* imply that it has also the functional organization. This is why, e.g., the musical composition cannot be formalized, although there were many attempts to find laws for its formalization (Berger et al. 1997). The detailed investigation of the mutual relationship between the functional and configurational organization of the general systems appears as an important future work.

[28]Let us remark that humans have an intuitive resistance against the probabilistic description of nature. This can be explained by the fact that - from their biological evolution - they have inherited a simple scheme of reasoning which consists of the one-dimensional cause-effect causality for which the deterministic way of inference is most adequate. Any other way of reasoning seems uncommon and strange to them. This holds especially for a reasoning in which the simple cause-effect scheme must be replaced by the cause-many effect scheme typical of probabilistic inference.

We conclude this Chapter with the remark that the conceptualization of the configurational and functional organization, described in the previous Chapters, is only an attempt to formalize the term 'organization' in the general systems theory whereby we realize that this concept being considered from another point of view may also be formalized alternatively.

References

Aczél, J. and Daróczy, Z. (1975), *On measures of information and their characterization.* New York: Academic Press.

Berger, R., Riečan, B. (eds.) (1997). *Mathematika a hudba.* Bratislava: Veda..

Bertalanffy, L. (1953). *Physik des Fliessgleigewichts.* Braunschweig: Vieweg.

Boltzmann, L. (1896). *Vorlesungen über Gastheorie.* Leipzig: A. Barth.

Bunge, M. (1979). *Treatise on Basic Philosophy, Vol. 4, Ontology II: A World of Systems*. Dordrecht: Reidel.

Dose, K. (1988). The origin of life – More questions than answers. *Interdisciplinary Science Reviews 13*, 348-356.

Frauenfelder, H. (1984). From atoms to bomolecules. Helvetica Physica Acta *57, 165.*

Denbigh, K.G. (1975). A non-conserved function for organized systems. In: L. Kubát, J. Zeman (eds.), *Entropy and Information in Science and Philosophy: 75-84.* Prague: Academia.

Eigen, M., Winkler, R. (1975). *Das Spiel.* München: Piper.

Eckman, D.P., Mesarovic', M.D. (1961). On the basic concepts of the general systems theory. In: *Proceedings of the 3rd International Congress on Cybernetics*. Namur.

Faddejew, D. K. (1957). Der Begriff der Entropie in der Wahrscheinlichkeitstheorie. In: *Arbeiten zur Informationstheorie I.* Berlin: Deutscher Verlag der Wissenschaften.

Gale, G. (1981). The Anthropic Principle. *Scientific American 245, 154-166.*

Haken, H. (1983). *Advanced Synergetics.* Berlin, Heidelberg, New York: Springer.

Haken, H. (1989). Synergetics: an overview. *Reports on Progress in Physics 52, 515-525.*

Ke-Hsueh Li (1986). Physics of open systems. *Physics Reports 134, 1-30.*

Khinchin, A. I. (1957). *Mathematical foundation of information theory.* .New York: Dover.

Klir,G. J. (1970). *An approach to general systems theory*. New York: van Norstrand.

Köhler, R. (1990). Synergetik und sprachliche Dynamik. In: W.A. Koch (ed.), *Natürlichkeit der Sprache und der Kultur: 69-112.* Bochum: Brockmeyer.

Köhler, R., Altmann, G. (1986), Synergetische Aspekte der Linguistik. *Zeitschrift für Sprachwissenschaft 5, 253-265.*

Kolmogoroff, A. N. (1957). *Theorie der Nachrichtungübertragung.* Berlin*:* Deutscher Verlag der Wissenschaften.

Kolmogorov, A. N. (1965). *Problems of Information Transmission 1, 1-52.* New York: Interscience.

Kubo, R. (1974). *Statistical Mechanics.* Amsterdam: North-Holland.

Kuhn, H. (1988). Origin of life and physics: Diversified microstructure-inducement to form information-carrying and knowledge-accumulating systems. *IBM J. Research and Development 32,37-52.*

Majerník, V. (1985). Elementary approach to the theory of fucntional organization. *International J. of General Systems 11, 221- 231.*

Majerník, V. (1988). Biological objects as fucntionally organized systems. *International J. of General Systems 14, 19-31.*

Majerník, V. (1970). Informationstheoretische Parameter von einfachem Musik-Material. *Kybernetika 6, 333-342.*

Majerník, V. (1998), Systems-theoretical approach to the concept of organization. In: G. Altmann, W. Koch (eds.), *Systems. New paradigm of human sciences: 126-142.* Berlin: de Gruyter.

Messiah, A. (1961). *Quantum Mechanics.* New York: Interscience.

Monod, J. (1971). *Chance and Necessity*. New York: A. Knopf.

Naranan, S., Balasubrahmanyan, V. K. (1991). Information-theoretical models in statistical linguistic-Part I: A model for word frequencies. *Current Science 63, 261-269*.

Nicolis, G. and Prigogine, I. (1976). *Self-organization in non-equilibrium systems*. New York, A. Wiley.

Orlov, J.K., Boroda, M.G., Nasdarejšvili, I. Š. (1982). *Sprache, Text, Kunst. Quantitative Analysen*. Bochum: Brockmeyer.

Parry, D.A.D., Baker, N. (1984). Biopolymers. *Reports on Progress in Physics 47, 1-46*.

Penrose, R. (1997). *The large, the small and the human mind.* Cambridge: University Press.

Schödinger, E. (1965). *What is life and other scientific essays.* New York: Doubleday.

Vajda, I. (1982). *Theory of information and statistical decisions*. Bratislava: Alfa.

Watanabe, S. (1969). *Knowing and guessing.* New York: Wiley.

Weinberg, S. (1993). *Dreams of final theory*. New York: Pantheon Books.

Welsh, D.J.A. (1970). Probability theory and its applications. In: E.Roubine (ed.) *Mathematics Applied to Physics: 387-598*. New York: Springer.

Woodfield, A. (1976). *Teleology*. Cambridge: University Press.

Glottometrics 3, 2002,136-142 *To honor G.K. Zipf*

The distribution of rhythmic units in German short prose

Karl-Heinz Best, Göttingen[1]

Abstract. The study of length of rhythmic units was initiated by K. Marbe (1904). In the present article we try to embed the problem in the general theory of length of linguistic units. It is shown that in modern German prose the length of rhythmic units abides by the Hyperpoisson distribution.

Keywords: Rhythmic unit, Hyperpoisson distribution

1. Zipf´s "Law of abbreviation"

One of the best known hypotheses of G.K. Zipf is his "law of abbreviation" (Zipf 1935/1968: 38) establishing a relationship between the length of words and their frequency and resulting in the generalization "The law of abbreviation is by no means restricted in its scope to the length of words" (39).

In the meantime it has been ascertained that the above hypothesis has been corroborated by the behavior of different entities. If the occurrence of forms of different length of any linguistic entity in texts or in lexicons is examined, it can be seen every time that the more complex the units, the less frequent they are.

However, there are deviations from this principle. For example, in Latin texts it is not the shortest words that are the most frequent ones, but usually the frequency of monosyllables is lower than that of disyllables, occasionally even that of disyllables is lower than that of trisyllables, but somewhere the frequency of longer words begins to decrease (Röttger 1996; Röttger, Schweers 1997; Wilson, McEnery 1998). If we measure word length according to the number of letters, sounds or phonemes instead of number of syllables, then, apparently, the shortest words are never the most frequent ones. The same holds for morphs and syllables (Best 2000, 2001b,c; Cassier 2001). Evidently, it depends on both, the unit examined and the way of measurement as well as the language, i.e. on the degree of its syntheticity, etc. One can observe the same phenomenon in many languages. Nevertheless, it is uncontested that there is a relationship between length and frequency, but boundary conditions can lead to different results.

The theoretical background for the distributions of units of different length in texts combined with preliminary examinations has especially been formulated for sentences (Altmann 1988a,b) and words (Wimmer et al. 1994; Wimmer, Altmann 1996; Wimmer, Witkovský, Altmann 1999). It consists, in principle, of the assumption that if length is a variable, its different values occur in texts in certain proportions, i.e. they are lawfully organized. A very frequently observed model is the Hyperpoisson distribution. There are, of course, many other distributions expressing the dependence on language, unit, author, style, etc. A long series of tests shows that the theoretical considerations concerning sentences and words are evidently valid and apparently applicable to all other language units (Best 2001).

[1] Address correspondence to: Karl-Heinz Best, Im Siebigsfeld 17, D-37115 Duderstadt. E-mail: kbest@gwdg.de

2. The distribution of rhythmic units in texts

In this paper we are concerned with a unit which has not been taken much note of in the research up to now. It is the rhythmic unit examined by Marbe (1904) in text passages of Goethe´s *Sankt Rochusfest zu Bingen* und Heine´s *Harzreise*. Marbe tried to find out whether Goethe had a more uniform rhythm than Heine.

If one reads a prosaic text and especially marks stressed syllables in sentences, one can ascertain that there are differently long passages of unstressed syllabes between two stressed ones; these are the "rhythmic units": their length is given by the stressed syllables and the subsequent unstressed ones, i.e. length 0 does not exist. As ascertained by Marbe, texts differ concerning the sequence of these units, as expected.

Automatically the question arises whether in Marbe´s data the length of these units abides by the established laws as well. This could preliminarily be corroborated by quite good results: in the chosen passages they followed the Hyperpoisson distribution (Best 2001d). Three further German texts yielded conform results (Best 2001a). This is, at last, a corroboration of Zipf´s assumptions.

Since up to now few text passages have been examined concerning the distribution of rhythmic units there are good grounds to perform further tests. The main question is: Do these units really follow the theoretically founded distributions?

3. Excursus

After his own research Marbe inspired a number of examinations with partially the same aim. The data presented in these examinations have been also checked, as far as possible. Let us summarize the results.

Friedmann (1921/22) examined passages consisting of approx. 1000 words from Old Hebrew Bible texts (Song of Solomon, Genesis, Chronicle, Psalms); he presents his results in two ways: directly as distributions of these units in the given passages and converted in relation to text length of 1000 words. The latter will not be considered here; for the unchanged data mostly unsatisfatory results have been attained using different distributions.

Gropp (1915) examined text passages of 1000 words in August Ludwig Hülsen. 1800. *Natur-Betrachtungen auf einer Reise durch die Schweiz.* (Athenaeum, Dritten Bandes Erstes Stück, 34ff, cit. according to Gropp,16), in all 7 full and one (last) shorter passage. Our trial to find a distribution for these data failed almost totally.

A different result has been found in data prepared by Lipsky (1907: 9). He presented the values of rhythmic units for the first 1004 words of the novel *The Red River* by J.F. Cooper. Here the binomial distribution can be fitted with a very good result ($P = 0.92$) as seen in Table 1.

Table 1
Fitting the binomial distribution to data by Lipsky (1907)

x	1	2	3	4	5	6	7
n_x	33	120	153	116	57	15	3
NP_x	35.24	113.34	156.22	119.62	54.96	15.15	2.47
$n = 7, p = 0.3148, X^2 = 0.90, DF = 4, P = 0.92$							

Explanatory notes to this table can be found in the next chapter; *n* and *p* are the parameters of the binomial distribution. More on this and other models can be found in Wimmer, Altmann (1999: 21ff.).

Summarizingly, it must be ascertained that the attempt at fitting distributions to text data leads to partial success only. Different causes may be accounted for: the arbitrariness of text passages and the uncertainty of identifying stressed syllables especially in foreign language texts, to mention only the most important ones. Marbe (1904: 4) assures in fact: that "such an accentuation can be very easily and in general surely performed"; similarly Gropp (1915: 22) who ascertains that his own allocation of stresses and that of two colleagues display merely tiny differences. My own experience rather suggests that even the repeated processing of the same text leads to different results. This holds even more with foreign language texts.

There are both objective and subjective reasons of the fact that one can perform different allocations of stresses in the same text. The objective ones consist, among other things, of the fact that some words permit different allocations of stresses ("wéshalb" or "weshálb"); the subjective ones result from the fact that text passages can be interpreted differently and none of the interpretations has a compelling preference. Perhaps one should consider the actual allocation of stresses as snapshots of the text interpretation by the given performer. This view shows at the same time that Hřebíček´s (1997) distinction between the generated (given) text and the interpreted one must be performed not only at the semantic level, but "down" to the phonetic level.

The problem of ambiguous allocations of stresses can, of course, not be removed. The other problem, the forming of arbitrary passages, interfering with data homogeneity, can and should be avoided (Altmann 1992: 296), for example such that short closed texts are processed. This was the reason for the good results obtained after the examination of three fairy tales by Pestalozzi in Best (2001a). For the same reason we analyzed some short prose texts by Erwin Strittmatter in order to enlarge the data base. The choice of the texts was arbitrary: they were available and they had the wanted shortness.

4. The distribution of rhythmic units in the short prose of Strittmatter

The processing of the texts was performed according to Marbe´s (1904) description: the stresses were positioned during soft and slow reading; this was corrected, if necessary, during the second reading. Then the lengths of rhythmic units were ascertained and tabulated. Merely the running text of the *Kleingeschichten* without titles and other textual additions was evaluated. The boundaries of sentences, chapters etc. were not taken into account if they did not form the beginning or the end of the text. The unstressed syllables before the first and last stressed syllable were taken into account. All processed texts were taken from Erwin Strittmatter (1985). *¾ hundert Kleingeschichten.* Berlin/Weimar: Aufbau.

Just as in the former investigations (Best 2001a,d) the 1-displaced Hyperpoisson distribution

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

was fitted to all data by means of the Altmann-Fitter (1997) yielding a satisfactory model of the data. The results can be found in the following tables (see Table 2).

Legend to the tables:

- x: length of rhythmic units. x = 1 means that a stressed syllable follows another stressed syllable; x = 2 means that there is one unstressed syllbable between two stressed ones, etc.
- n_x: number of rhythmic unit with length x

NP_x: theoretical number of rhythmic units computed by means of the Hyperpoisson distribution
a,b: parameters (rounded to four decimal digits)
X^2: the chi-square value of the goodness-of-fit test
DF: degrees of freedom
P: the probability of the given or greater chi-square.

The result were considered satisfactory if $P \geq 0.05$. Underlined values show the pooling of length classes.

Table 2
Fitting of the Hyperpoisson distribution to selected texts

Text 1: Neue Nachrichten vom Eis (p. 10f.)

x	1	2	3	4	5	6	7	8	9	10
n_x	2	22	41	29	11	10				
NP_x	1.67	24.65	36.51	28.47	15.06	8.65				
$a = 1.6463$			$b = 0.1115$		$X^2 = 2.219$		$DF = 3$		$P = 0.53$	

Text 2: Frühlingsanstoß (p. 16f.)

x	1	2	3	4	5	6	7	8	9	10
n_x	5	37	41	29	23	9	5	3	1	
NP_x	9.99	29.43	39.40	34.14	21.86	11.11	4.67	<u>1.68</u>	<u>0.72</u>	
$a = 2.4553$			$b = 0.8337$		$X^2 = 6.831$		$DF = 5$		$P = 0.23$	

Text 3: Birken (p. 26f.)

x	1	2	3	4	5	6	7	8	9	10
n_x	6	31	42	29	21	9	5	3	0	2
NP_x	12.19	27.56	35.16	31.24	21.29	11.77	5.47	2.20	<u>0.78</u>	<u>0.34</u>
$a = 2.9267$			$b = 1.2942$		$X^2 = 6.762$		$DF = 6$		$P = 0.35$	

Text 4: Ein andres Fohlen kommt zur Welt (p 45f.)

x	1	2	3	4	5	6	7	8	9	10
n_x	6	31	42	32	14	6	6	2		
NP_x	5.44	28.11	40.04	33.07	19.24	8.63	3.16	1.31		
$a = 1.9663$			$b = 0.3806$		$X^2 = 5.644$		$DF = 5$		$P = 0.34$	

Text 5: Damals bei der Haferaussaat (p. 36f.)

x	1	2	3	4	5	6	7	8	9	10
n_x	5	35	47	32	17	6	2			
NP_x	5.09	35.60	46.01	32.75	16.09	6.03	2.42			
$a = 1.5850$			$b = 0.2264$		$X^2 = 0.175$		$DF = 4$		$P = 0.99$	

Text 6: Hasenhaare (p. 47f.)

x	1	2	3	4	5	6	7	8	9	10
n_x	13	33	42	21	15	12	4	1	0	1
NP_x	16.21	30.74	34.74	27.96	17.47	8.93	3.85	<u>1.44</u>	<u>0.48</u>	<u>0.19</u>
$a = 2.7962$			$b = 1.4742$		$X^2 = 5.471$		$DF = 5$		$P = 0.36$	

Text 7: Ein Leckermaul (p. 50f.)

x	1	2	3	4	5	6	7	8	9	10
n_x	3	22	32	31	22	10	1	1		
NP_x	2.93	21.52	34.21	30.50	18.90	8.97	3.45	1.52		
$a = 2.0300$		$b = 0.2768$		$X^2 = 2.712$		$DF = 5$		$P = 0.74$		

Text 8: Schwertlilien (p. 67f.)

x	1	2	3	4	5	6	7	8	9	10
n_x	6	45	35	31	30	11	9	4	1	
NP_x	16.00	31.24	38.55	34.78	24.73	14.50	7.24	3.15	1.81	
$a = 3.3536$		$b = 1.7175$		$X^2 = 2.271$		$DF = 2$		$P = 0.32$		

Text 9: Grüne Lauben (p. 72f.)

x	1	2	3	4	5	6	7	8	9	10
n_x	9	33	36	29	15	9	3	1		
NP_x	9.90	30.19	37.53	29.30	16.68	7.47	2.76	1.17		
$a = 2.0992$		$b = 0.6886$		$X^2 = 0.940$		$DF = 5$		$P = 0.97$		

Text 10: Der Kauz (p. 81f.)

x	1	2	3	4	5	6	7	8	9	10
n_x	6	32	41	18	16	4	2	1		
NP_x	5.57	30.22	37.63	26.52	13.04	4.92	1.51	0.49		
$a = 1.6242$		$b = 0.3045$		$X^2 = 4.508$		$DF = 4$		$P = 0.34$		

Text 11: Das Eichhorn (p. 84f.)

x	1	2	3	4	5	6	7	8	9	10
n_x	4	20	25	25	20	8	6			
NP_x	3.57	17.87	28.13	26.28	17.46	9.00	5.70			
$a = 2.2980$		$b = 0.4596$		$X^2 = 1.213$		$DF = 4$		$P = 0.88$		

Text 12: Werkstattschwalben (p. 87f.)

x	1	2	3	4	5	6	7	8	9	10
n_x	5	20	30	16	16	9	3	0	1	1
NP_x	4.61	18.43	26.78	23.77	15.19	7.59	3.11	1.08	0.33	0.11
$a = 2.2817$		$b = 0.5704$		$X^2 = 3.560$		$DF = 5$		$P = 0.61$		

Text 13: Der Pony-Igel (p. 90f.)

x	1	2	3	4	5	6	7	8	9	10
n_x	6	33	54	31	18	10	4	1	1	
NP_x	6.05	33.30	46.63	37.41	21.03	9.10	3.20	0.95	0.31	
$a = 1.8786$		$b = 0.3416$		$X^2 = 3.428$		$DF = 5$		$P = 0.63$		

Text 14: Schlechte Laune (p. 113f.)

x	1	2	3	4	5	6	7	8	9	10
n_x	7	32	33	36	16	6	3	1		
NP_x	6.83	29.70	39.25	30.59	16.90	7.23	2.53	0.99		
$a = 1.8986$		$b = 0.4364$		$X^2 = 2.462$		$DF = 4$		$P = 0.65$		

Text 15: Schneefrühling (p. 115f.)

x	1	2	3	4	5	6	7	8	9	10
n_x	5	34	49	23	13	11	5			
NP_x	4.51	30.68	42.81	33.29	17.94	7.39	3.37			
$a = 1.7556$ $b = 0.2582$ $X^2 = 8.384$ $DF = 4$ $P = 0.08$										

Text 16: Eine Freude ist eine Freude (p. 128f.)

x	1	2	3	4	5	6	7	8	9	10
n_x	6	24	50	28	20	12	4			
NP_x	4.51	28.33	41.84	35.02	20.44	9.16	4.70			
$a = 1.9308$ $b = 0.3072$ $X^2 = 5.144$ $DF = 4$ $P = 0.27$										

5. The result

The results show that - just as it is the case with Pestalozzi´s fairy tales (Best 2001a: 5f) and Strittmatter´s *Kleingeschichten* the Hyperpoisson distribution seems to be an adequate model for the distribution of rhythmic units in German texts. This result corroborates the numerous investigations concerning other language units with comparable success. However, since no scientific model is a final one and the number of fully processed texts is very small, this model is just a preliminary one. Besides, similar examinations should be performed by different researchers in order to reduce the problem of differently allocating stresses to the same text. The processing of texts in other languages would be welcomed, too.

Let us touch still another problem. Deußing (1927/1969: 129ff) pointed out that there is a tendency with children to use shorter rhythmic units, perhaps because children possess a vocabulary with a high number of monosyllabic words. In view of our aim it could be examined whether statements or texts of children abide by the same or different distributions observed up to now. Laass´ examination of word length in reading books (1996) has shown that this possibility should be taken into account.

References

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

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

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

Best, Karl-Heinz (2000). Morphlängen in Fabeln von Pestalozzi. *Göttinger Beiträge zur Sprachwissenschaft 3, 19-30*.

Best, Karl-Heinz (2001). Kommentierte Bibliographie zum Göttinger Projekt. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 284-310.* Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2001a). Probability distributions of language entities. *Journal of Quantitative Linguistics 8: 1-11.*

Best, Karl-Heinz (2001b). Zur Länge von Morphen in deutschen Texten. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 1-14*. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2001c). Silbenlängen in Meldungen der Tagespresse. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 15-32*. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2001d). Zur Verteilung rhythmischer Einheiten in deutscher Prosa. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 162-166*. Göttingen: Peust & Gutschmidt.

Cassier, Falk-Uwe (2001). Silbenlängen in Meldungen der deutschen Tagespresse. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 33-42*. Göttingen: Peust & Gutschmidt.

Deußing, Hans (1927/ 1969). Der sprachliche Ausdruck des Schulkindes. In: Helmers, Hermann (Hrsg.), *Zur Sprache des Kindes: 60-131.* Darmstadt: Wissenschaftliche Buchgesellschaft.

Friedmann, Martin (1921/22). *Der Prosarhythmus des Hebräischen im alten Testament.* Würzburg, diss.phil.

Gropp, Friedrich (1915). *Zur Ästhetik und Statistik des Prosarhythmus.* Würzburg, diss.phil.

Hřebíček, Luděk (1997). *Lectures on text theory.* Prague: Oriental Institute**.**

Laass, Françoise (1996). Zur Verteilung der Wortlänge in deutschen Lesebuchtexten. In: Schmidt, Peter (Hrsg.), *Glottometrika 15, 181-194*. Trier: Wissenschaftlicher Verlag Trier.

Lipsky, Abram (1907). *Rhythm as a distinguishing characteristic of prose style.* New York: The Science Press.

Marbe, Karl (1904). *Über den Rhythmus der Prosa*. Giessen: J. Ricker'sche Verlagsbuchhandlung.

Röttger, Winfred (1996). The distribution of word length in Ciceronian Letters. *Journal of Quantitative Linguistics 3, 68-72.*

Röttger, Winfred / Schweers, Anja (1997). Wortlängenhäufigkeiten in Plinius-Briefen. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16, 121-126*. Trier: Wissenschaftlicher Verlag Trier.

Wilson, Andrew / McEnery, Tony (1998). Word length distributions in Biblical and Medieval Latin. *The Prague Bulletin of Mathematical Linguistics 70, 5-21.*

Wimmer, Gejza / Altmann, Gabriel (1996). The Theory of word length distribution: some results and generalizations. In: Schmidt, Peter (Hrsg.), *Glottometrika 15, 112-133.* Trier: Wissenschaftlicher Verlag Trier.

Wimmer, Gejza / Altmann, Gabriel (1999). *Thesaurus of univariate discrete probability distributions.* Essen: Stamm.

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.*

Zipf, George Kingsley (1935/1968). *The psychobiology of language: An introduction to dynamic philology.* Cambrige, Mass. The M.I.T. Press

Software

Altmann-Fitter (1997). *Iterative Fitting of Probability Distributions.* Lüdenscheid: RAM-Verlag.

Glottometrics 3, 2002,143-150 *To honor G.K. Zipf*

Zipf's law and the Internet

Lada A. Adamic[1]
Bernardo A. Huberman

Abstract. Zipf's law governs many features of the Internet. Observations of Zipf distributions, while interesting in and of themselves, have strong implications for the design and function of the Internet. The connectivity of Internet routers influences the robustness of the network while the distribution in the number of email contacts affects the spread of email viruses. Even web caching strategies are formulated to account for a Zipf distribution in the number of requests for webpages.

Keywords: Zipf's law, caching, networks

Introduction

The wide adoption of the Internet has fundamentally altered the ways in which we communicate, gather information, conduct businesses and make purchases. As the use of the World Wide Web and email skyrocketed, computer scientists and physicists rushed to characterize this new phenomenon. While initially they were surprised by the tremendous variety the Internet demonstrated in the size of its features, they soon discovered a widespread pattern in their measurements: there are many small elements contained within the Web, but few large ones. A few sites consist of millions of pages, but millions of sites only contain a handful of pages. Few sites contain millions of links, but many sites have one or two. Millions of users flock to a few select sites, giving little attention to millions of others.

This pattern has of course long been familiar to those studying distributions in income (Pareto 1896), word frequencies in text (Zipf 1932), and city sizes (Zipf 1949). It can be expressed in mathematical fashion as a power law, meaning that the probability of attaining a certain size x is proportional to $x^{-\tau}$, where τ is greater than or equal to 1. Unlike the more familiar Gaussian distribution, a power law distribution has no 'typical' scale and is hence frequently called 'scale-free'. A power law also gives a finite probability to very large elements, whereas the exponential tail in a Gaussian distribution makes elements much larger than the mean extremely unlikely. For example, city sizes, which are governed by a power law distribution, include a few mega cities that are orders of magnitude larger than the mean city size. On the other hand, a Gaussian, which describes for example the distribution of heights in humans, does not allow for a person who is several times taller than the average. Figure 1 shows a series of scale free distributions in the sizes of websites in terms of the number of pages they include, the number of links given to or received from other sites and the number of unique users visiting the site.

[1] Address correspondence to: Lada A. Adamic, HP Laboratories, 1501 Page Mill Road, ms 1139, Palo Alto, CA 94304, USA. E-mail: ladamic@exch.hpl.hp.com.

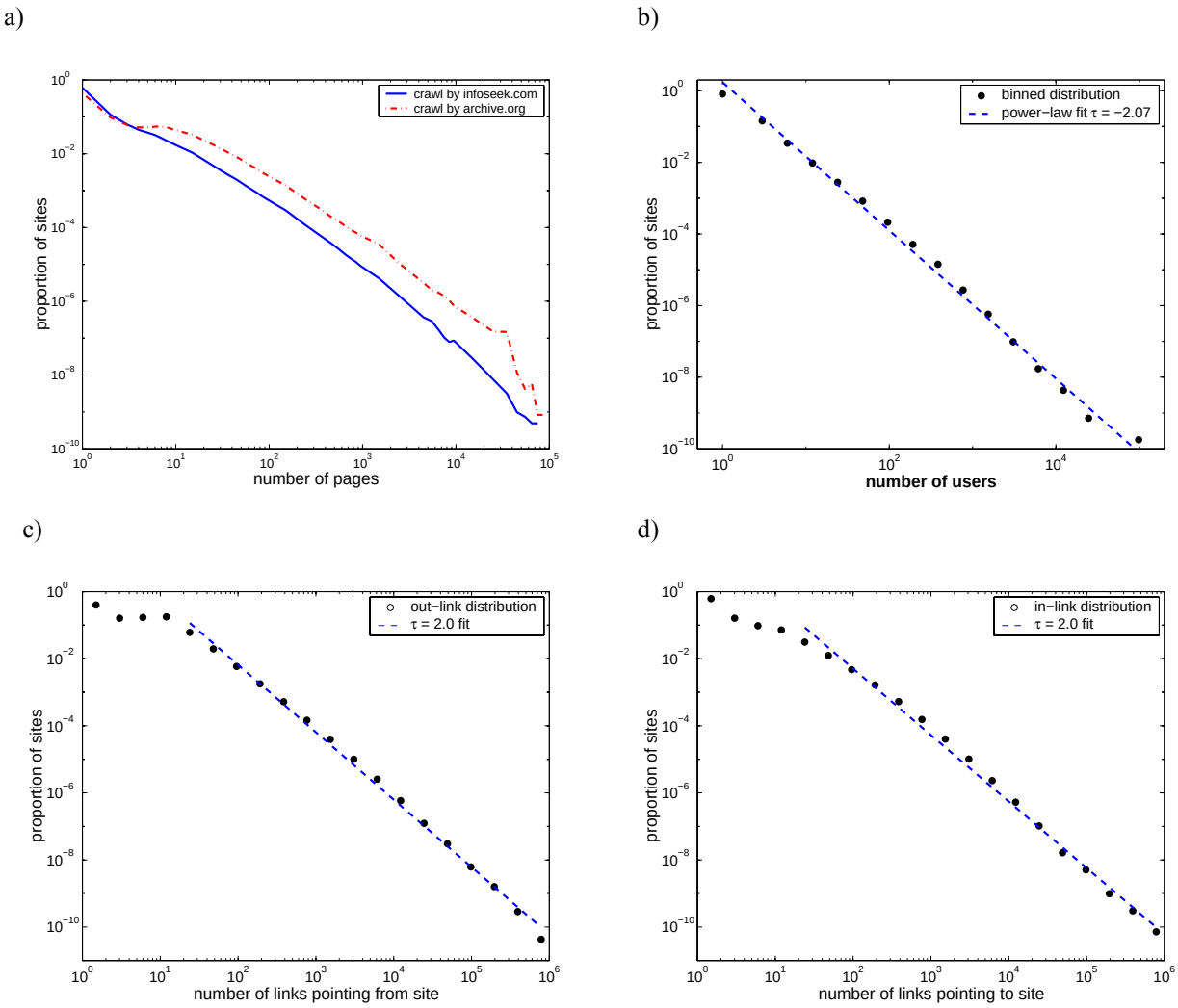


Figure 1. Fitted power law distributions of the number of site a) pages, b) visitors, c) out links, and d) in links, measured in 1997.

Although the distributions plotted above are given in terms of the probability density function (PDF), they can also be easily recast in terms of Zipf's ranked distribution. In fact, any purely power-law probability density function will yield a Zipf ranked distribution as follows:

Let the PDF be $p(x) = Cx^{-\tau}$. Then the probability that a website is of size y or lager $P(x > y) = \sum_{y}^{\infty} Cx^{-\tau} \approx Ay^{-\tau+1}$, C and A are constants. If there are N websites total, the expected number of sites greater than N is given by $r = NAy^{\tau-1}$. Solving for y, we find that the size of the r^{th} ranked variable is proportional to $r^{-\frac{1}{\tau-1}} = r^{-\alpha}$, α being the Zipf rank exponent. While the PDF emphasizes the count of small elements, the ranked distribution emphasizes the size of the largest ones.

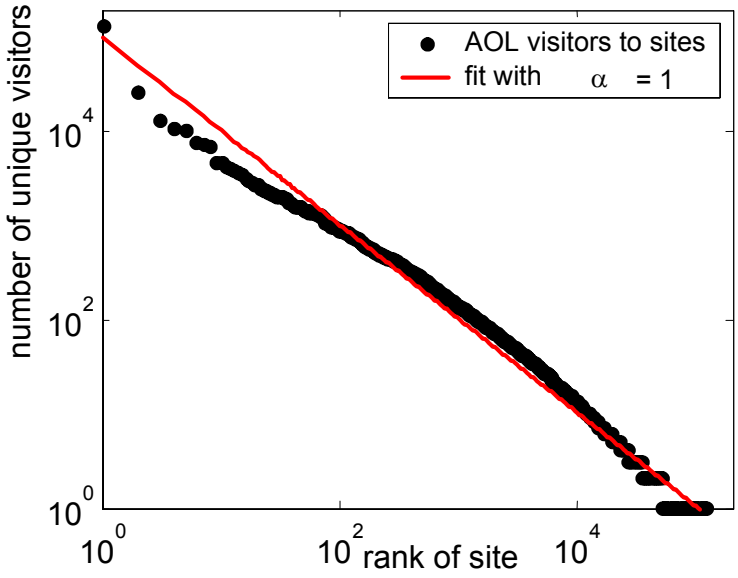


Figure 2. Sites ranked by the number of unique AOL visitors they received Dec. 1, 1997. AOL (America Online) is the largest Internet service provider in the United States. The fit is a Zipf distribution $n_r \sim r^{-1}$

Figure 2, for example, shows the ranked distribution of visitors to web sites corresponding to the PDF plotted in Figure 1b. The distribution shows mild concavity and a ranked exponent of 1: Zipf's law. As Table 1 shows, a small number of sites such as Yahoo are extremely popular and capture a disproportionate amount of traffic.

Table 1
Distribution of user volume among sites in general, adult sites, and .edu domain sites, as determined by counting the number of unique AOL visitors on Dec. 1, 1997.

%sites	%user volume
0.1	32.36
1	55.63
5	74.81
10	82.26
50	94.92

A Growth Model

The pervasiveness of Zipf distributions on the Internet can be explained by an intuitive growth model (Huberman 1999) that incorporates three simple assumptions. Let us formulate the argu-ment in terms of the number of web pages hosted on a website. Similar arguments can be applied just as easily to the number of visitors or links. The first assumption is that of proportional growth or preferential attachment, i.e. the number of pages added to or removed from the site is proportional to the number of pages already present. For example, a site with a million pages might have a whole team of webmasters or generate its content automatically. It could easily gain or shed a several thousand pages on any given day. On the other hand, it would be surprising, but not impossible, for a site with only a handful of pages to suddenly add a thousand more.

This multiplicative stochastic growth process yields a lognormal distribution in the number of pages at a site after a fixed period of time. However, the World Wide Web is anything but fixed. Its first decade was a period of rapid growth, with sites appearing at an exponential rate. It so happens that when one computes an exponentially weighted mixture of lognormals one obtains a power-law distribution exactly!

While the exponential growth in the number of websites and their stochastic addition of pages alone can produce power law distributions, a key ingredient is still missing. For in spite of the random nature of the growth, if one were taking a mixture of lognormals depending only on a time variable, one would expect that the sites established early would have grown to greater sizes than recently founded ones. However, studies have found only weak correlation between the size of a site and its age (equivalently some very popular sites were founded more recently, while sites present at the very start of the Internet boom did not necessarily acquire a wide audience). The missing assumption is that sites can grow at different rates, depending on the type of content and interest that they generate. Incorporating variability in growth rates again yields power law distributions with varying exponents. The greater the difference in growth rates among sites, the lower the exponent τ, which means that the inequality in site sizes increases. In summary, a very simple assumption of stochastic multiplicative growth, combined with the fact that sites appear at different times and/or grow at different rates, leads to an explanation for the scale free behavior so prevalent on the Web (Huberman 2001).

Caching

Computer scientists have gone beyond observations and explanations of Zipf's law to apply it to the design of content delivery on the Internet. A problem Internet service providers (ISP's) face is devising ways to support rapidly growing web traffic while maintaining quality of service in the form of fast response time for file requests. In order to quickly satisfy users' request for web content, ISP's utilize caching, whereby frequently used files are copied and stored "near" to users on the network. It is important to note, however, that the effectiveness of caching relies heavily on the existence of Zipf's law.

Let's say that there is a web server in the United States serving a page that is extremely popular in a town in Europe. In the absence of caching, every time someone in that town requests the page, their request travels across the Atlantic, reaches the US server, which in turn sends the page back across the Atlantic to the requester in Europe.

To avoid sending unnecessary cross-Atlantic requests, the Internet service provider serving the European town can place a proxy server near the town. The proxy server's role is to accept requests from the users and forward them on their behalf. Now, when the first user requests the document, the request goes to the proxy. If the proxy cache does not contain the document, it makes a request to the US server, which replies to the proxy. The proxy then sends the file to the requesting user, and stores the file locally in a cache. When additional users send their requests for the file to the proxy, the proxy can serve them the file directly from its cache, without having to contact the webserver in the US. Of course, files that are updated frequently, such as the front page of a news site, have an expiration time after which the file is considered 'stale'. The cache uses the expiration time to determine when to request a new version from the origin server.

Caching has two advantages. First, since the requests are served immediately from the cache, the response time can be significantly faster than contacting the origin server. Second, caching conserves bandwidth by avoiding redundant transfers along remote internet links. The benefits of caching are confirmed by its wide use by ISPs. They benefit because they are able

to reduce the amount of inter-ISP traffic that they have to pay for. Caching by proxies benefits not only the ISPs and the users, but also the websites holding the original content. Their content reaches the users more quickly and they avoid being overloaded themselves by two many direct requests.

However, since any cache has a finite size, it is impossible for the cache to store all of the files users are requesting. Here Zipf's law comes into play. Several studies (Cunha 1995, Breslau 1999) have found the popularity of files requested follows a Zipf distribution. Hence, the cache need only store the most frequently requested files in order to satisfy a large fraction of users requests.

Networks

The Internet is comprised of networks on many levels, and some of the most exciting consequences of Zipf's law have been discovered in this area. The World Wide Web is a network of interconnected webpages and the Internet backbone is a physical network used to transmit data, including web pages, in the form of packets, from one location to another. Measurements on both the World Wide Web (Adamic 1999, Jeong 1999) and the Internet backbone (Faloutsos 1999, Albert 2000) have shown that they differ significantly from the classic Erdös-Rényi model of random graphs (Erdös 1960). While the traditional Erdös-Rényi model has a Poisson node degree distribution, with most nodes having a characteristic number of links, these networks approximately follow a Zipf or scale-free degree distribution $p(k) \sim k^{-\tau}$, where k is the node degree, and τ is the scale-free exponent. To account for these observations, new random graph growth models have been developed that rely on the above mentioned idea of preferential attachment (Albert 2002).

The scale free degree distribution of the Internet backbone, shown in Figure 3, implies that some nodes in the network maintain a large number of connections (proportional to the total size of the network), while for the most part nodes have just one or two connections. This is a two edged sword when it comes to resilience of the network. It means that if a node fails at random, it is most likely one with very few connections, and its failure won't affect the performance of the network overall. However, if one were to specifically target just a few of the high degree nodes, the network could be adversely affected. Because many routes pass through the high degree nodes, their removal would require rerouting through longer and less optimal paths. Once a sufficient number of high degree nodes are removed, the network itself can become fragmented, without a way to communicate from one location to another.

On a different level, one of the recent developments in the use of the Internet has been the emergence of peer-to-peer (P2P) networks. These networks are used by millions of users daily to exchange a variety of files directly with one another. Examples of P2P networks include Napster, Gnutella, and Kazaa. Although Napster was immensely popular, it was forced to shut down by the recording industry over concerns that users were trading copyrighted music files. Part of the reason Napster could so easily be shut down is that it operated with a central server. The users would report which files they were sharing to the central server, and when they looked for additional files, they would query the central server to locate other users who had those files.

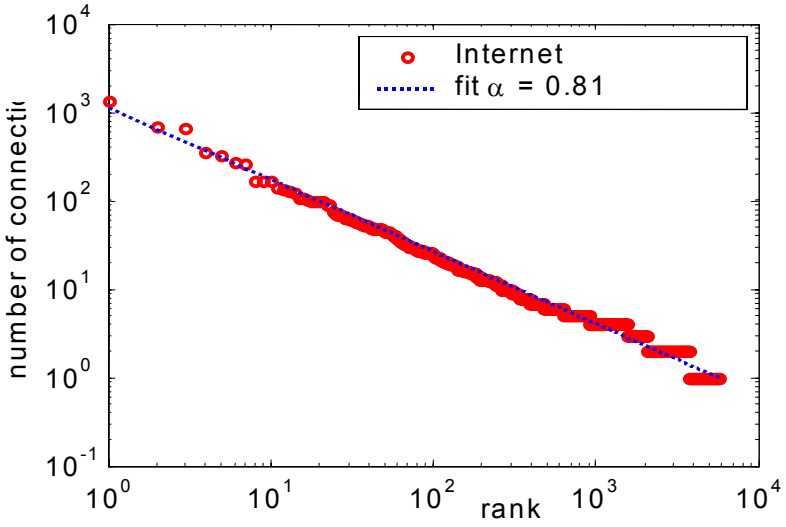


Figure 3. The connectivity of the internet backbone at the autonomous system (AS level). Each AS is itself a network corresponding to a single ISP, business entity or educational institution.

Having learned from Napster's troubles, current peer-to-peer networks tend to be decentralized. That is, nodes connect directly to one another rather than to a central server. The distribution in the number of computers a computer has connections to is a Zipf distribution (recently it has shifted into a two-sided Zipf distribution, with a shallower exponent for the high degree nodes and a steeper exponent for the low degree ones) (Ripeanu 2002). The presence of Zipf's law has implications for the search strategies used in P2P networks. Currently, most P2P networks use a broadcast method of locating files. Because there is no central server that queries can be sent to, each node broadcasts the query to all of its neighbors who in turn broadcast to all of their neighbors, out to some fixed distance from the originating node. As one can well imagine, the network can become quite congested with broadcasted queries. Recent research has shown, however, that routing queries to the high degree nodes may provide a degree of congestion relief, while maintaining a short response time (Adamic 2001). Again, knowledge of Zipf's law in the connectivity distribution has offered a solution to an Internet communication problem.

Finally, it has been shown that scale-free networks are more susceptible to viruses than networks with a more even degree distribution. Namely, a virus spreading in a random network needs to surpass a threshold of infectiousness in order not to die out. However, if the network has a Zipf degree distribution, the virus can persist in the network indefinitely, no matter what level of its infectiousness (Pastor-Satarros 2002).

Both email (Ebel 2002) and instant messaging networks (Smith 2002) have been shown to be scale free. Some individuals have a large number of email contacts but most individuals would keep only a few addresses in their contact lists. This wide variance in the connectivity of electronic communication reflects the different degrees of communicativeness in people and their different roles at work and in society overall. Over the past few years, email viruses have plagued the Internet, no doubt facilitated by hubs, or individuals with large contact lists. An email virus can be passed on as an attachment in email messages. Once the attachment is opened, the virus can activate and cause the email program to send numerous infected emails to email addresses from the person's contact list. The "I love you" email virus alone infected over 500,000 individual systems in May of 2000[2]. Sometimes the sheer quantity of viral

[2]Source: CERT® Advisory, http://www.cert.org/advisories/CA-2000-04.html

email can affect the Internet's performance. But just as hubs (individuals or computers with many contacts) can facilitate the spreading of a virus, they can also aid in preventing their spread. Carefully im-munizing the hubs could stop the virus in its tracks.

Conclusions

On the Internet, Zipf's law appears to be the rule rather than the exception. It is present at the level of routers transmitting data from one geographic location to another and in the content of the World Wide Web. It is also present at the social and economic level, in how individuals select the websites they visit and form peer-to-peer communities. The ubiquitous nature of Zipf's law in cyberspace has lead to a deeper understanding of Internet phenomena, and has consequently influenced the way in which it has evolved.

Acknowledgements
We would like to thank T.J. Giuli, Eytan Adar and Rajan Lukose for their comments and suggestions.

References

Adamic, L.A. (1999). The Small World Web, Proceedings of ECDL'99. *Lecture Notes in Computer Science 1696, 443-452*. Berlin: Springer.

Adamic, L.A., Lukose, R.M., Puniyani, A.R., and Huberman, B.A. (2001). Search in Power-Law Networks. *Physical Review E 64: 046135*.

Albert, R., Jeong, H., and Barabasi, A.-L. (2000). Attack and error tolerance of complex networks. *Nature 406, 378*.

Albert R. and Barabasi A.-L. (2002). Statistical mechanics of complex networks. *Review of Modern Physics 74, 47-94*.

Breslau L. et al. (1999). Web Caching and Zipf-like Distributions: Evidence and Implications. *Proceedings of INFOCOM '99, 126-134*.

Cunha, C.R., Bestavros A., and Crovella, M.E. (1995). Characteristics of WWW Client-based Traces". *Technical Report TR-95-010*. Boston University Computer Science Department.

Ebel, H, Mielsch, L.-I., and Bornholdt, S. (2002). Scale-free topology of e-mail networks. *cond-mat/0201476*.

Erdös, P. and Rényi, A. (1960). On the Evolution of Random Graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5, 17-61*.

Faloutsos M., Faloutsos P. and Faloutsos C. (1999). On Power-Law Relationships of the Internet Topology. *Proceedings of ACM SIGCOMM '99, 251-262*.

Jeong, H., Albert R. and Barabasi, A.-L (1999). Diameter of the World Wide Web. *Nature 401, 130*.

Huberman, B. and Adamic, L. (1999). Growth Dynamics of the World Wide Web. *Nature 401, 131*.

Huberman, B. A. (2001). *The Laws of the Web.* The MIT Press.

Pareto, V. (1896). *Cours d'Economie Politique*. Genève: Droz.

Pastor-Satarros, R. and Vespignani, A. (2001). Epidemic spreading in Scale Free Networks. *Physical Review Letters 86, 3200*.

Ripeanu M., Foster, I. and Iamnitchi A. (2002). Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design. *IEEE Internet Computing Journal special issue on peer-to-peer networking, vol. 6(1), 50-57.*

Smith, R.D. (2002). Instant Messaging as a Scale-Free Network. *cond-mat/0206378*.

Zipf, G.K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA.: Harvard University Press.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort.* Cambridge, MA: Addison-Wesley.