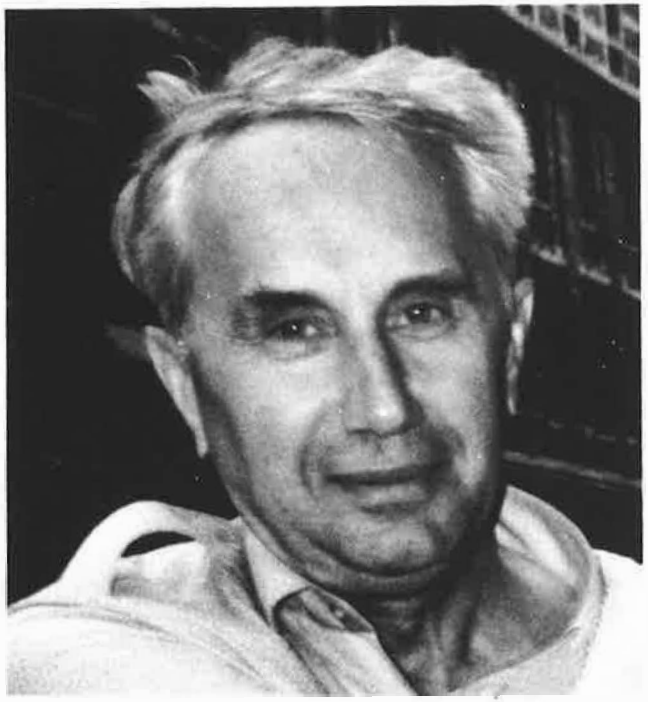
QUANTITATIVE LINGUISTICS

Volume 60



Text as a Linguistic Paradigm: Levels, Constituents, Constructs

Festschrift in honour of Luděk Hřebíček

edited by

Ludmila Uhlířová, Gejza Wimmer
Gabriel Altmann, Reinhard Köhler

WVT Wissenschaftlicher Verlag Trier

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Text as a Linguistic Pradigm: Levels, Constituents, Constructs. Festschrift in Honour of Luděk Hřebíček
Ed. by Ludmila Uhlířová, Gejza Wimmer,
Gabriel Altmann, Reinhard Köhler. -
WVT Wissenschaftlicher Verlag Trier, 2001
ISBN 3-88476-398-9

Umschlag: Brigitta Disseldorf

ISBN 3-88476-398-9
ISSN 0932-7991

WVT Wissenschaftlicher Verlag Trier
Bergstraße 27, 54295 Trier
Postfach 4005, 54230 Trier
Tel. 0651-41503, Fax 41504
Internet: www.wvttrier.de
Email: wvt@wvttrier.de

Contents

Preface

Luděk Hřebíček used to say that the best way to celebrate one's jubilee, and, actually, the only way he really likes and appreciates, is to do it by work. The present volume contains papers written by 34 linguists, mathematicians, psychologists and information scientists from 8 countries situated both near the Greenwich meridian and far eastward from it, from countries situated near the equator as well as from others near to the north pole - altogether, a large collection of intellectual scientific work has been gathered here. It is dedicated to Hřebíček on the occasion of his sixty-fifth birthday with the contributors' most cordial and most sincere wishes for future years: Good health and great success in linguistic *science*.

The word science is highlighted in the preceding sentence intentionally. One of the dominant features of Hřebíček's scientific thinking and reasoning is a high sensitivity to what is going on, not only within the traditional framework of linguistics, but also outside its boundaries, in science in general. His quantitative text theory is grounded in a broad interdisciplinary context. It is closely connected with the disciplines usually called the exact sciences in contrast to the humanities. Hřebíček is one of those who have made a decisive step in quantitative linguistics over the last decades, changing it from a mere descriptive discipline into a standard scientific discipline formulating and testing linguistic hypotheses. The leading idea of present-day quantitative linguistics, as presented in Hřebíček's papers and books, may be characterised in the following way: Today it is far from sufficient to propose a statistical hypothesis concerning, for example a probability distribution, as derived from an empirical observation of linguistic data. As a starting point, theoretical distributions should be deduced from linguistic presumptions (conjectures, hypotheses, models, theories). Only then should the probability theory be applied and the empirical variables tested. As Hřebíček claims, "the construction of testable theories seems to be the only way to the formulation of questions and answers comparable with questions and answers occurring in other contemporary sciences" (1997:3). And this is precisely what Hřebíček aims to do when seeking hidden structures and general principles of order in language.

Quantitative text theory is the discipline which has become Hřebíček's life interest - and life success. The tested linguistic hypothesis, which he considers crucial for text linguistics, is Menzerath-Altmann's law. Hřebíček has shown

the immense consequences which the law has for text analysis. He started from a hypothesis that text is an entity which is segmentable in accordance with Menzerath-Altmann's law ("the longer a language construct the shorter its components"). He demonstrated that not only constructs and their con-structional elements (constituents) at different language levels obey Menzerath-Altmann's law. Moreover, if a text is segmented by Menzerath-Altmann's law, new levels of its structure may be revealed, and, the relations between language levels (both newly revealed and already known) are identical with those between constructs and constituents. General principles on which text as a linguistic paradigm, or as a construct in the sense of Menzerath-Altmann law (1995:99) is constructed, are explicated by Hřebíček exactly, discussed systematically, in a deep philosophical context, so typical for the author's style of exposition. For the time being, there exists no alternative quantitative text theory elaborated so consistently and in such detail, which could be compared to Hřebíček's. There seem to be no opponents so far, but a lot of positive response; we can say that Hřebíček is a leading scholar in his field..

Hřebíček was born in Prague in 1934. He studied Persian and Turkish at the Charles University, Prague (he graduated in 1958) and has spent most of his professional career in the Oriental Institute, Academy of Sciences of the Czech republic; at present, he is the head of the Africa and Near East department and the chairman of the Scientific Council of the Institute. He took his Ph.D. degree in 1963 and D.Sc. degree 1992.

At the very beginning of his carreer he admired the magnificent Persian literature most of all. Also his Ph.D. thesis was oriented to literature. It dealt with the analysis of poems by a classic Kazakh poet Abay Kunanbayev. Hřebíček analysed the rhythmic and metric structure of Abay's verse and published several papers on different aspects of it between the years 1964 and 1966 (see the bibliography below). Let us notice that his versologic and stylistic thesis already had the main characteristic features of his favourite methodology: He presented a detailed stylometric analysis of the poems, using quantitative methods. The most valuable results concerned the phenomenon of euphonia - Hřebíček developed some ideas of the Prague structuralist Jan Mukařovský.

As time went on, linguistic interests gradually prevailed. Hřebíček himself repeatedly mentioned how much he was fascinated by the typological principle of agglutination, so different from all those typological principles with which he became familiar when studying English, German, French and Latin as early as in grammar school. In the 1970s he started various linguistic topics and published papers and monographs. A book on Turkish syntax (1971) and a monograph on social communication (1986) were among them.

During the last two decades, which he himself evaluates as the most versatile, he began experimenting with texts. He chose the field of language study, which is very difficult and very challenging. He always claimed that experiment

is a specific way to observe the reality under exactly defined conditions, and it is the strictness in the defining of conditions that is so characteristic of his work. His textual experiments, mostly on Turkish data, were published in three monographs (1992), (1995), and (1997) and in a great number of articles and studies in journals and various volumes, mostly in English (see the bibliography below).

Stochasticity as a property of the language system, its subsystems, levels, and constructs provoked Hřebíček to ask principal theoretical questions. For example: How many language levels can be recognized in language, if one starts from a presupposition that it is possible between any two levels (defined in terms of the Menzerath-Altmann law) to find another level, defined again with the help of the Menzerath-Altmann law? Hřebíček defined a text level of aggregates in terms of the Menzerath-Altmann law. An aggregate is a set of sentences of a text, in all of which the same lexical unit occurs. For example, let us have a text "...the old quarrel between the British and the Continental schools of philosophy..." (1995:30); an aggregate of this text consists of all sentences in which "quarrel" occurs, another aggregate is formed by all sentences in which "school" occurs, still another aggregate consists of all sentences containing "British" etc. Obviously, the number of aggregates of a text equals the number of different lexical units. Hřebíček demonstrated that - according to the Menzerath-Altmann law - the longer an aggregate (in number of sentences) the shorter the mean length of its sentences (in number of words). Hřebíček not only claimed that text is a construct of aggregates, but even more generally, that "level is, in fact, a consequence of the MA law" (1995:19).

Hřebíček is a man who has brilliantly succeeded in his scientific career. When he was young, he was lucky in choosing a research field which became his life's love, a field to which he has been fully devoted, a field in which he has become an excellent representative of the best Czech quantitative traditions laid down by the Czech structuralists as early as the 1930s, and in which he has gained an international reputation. He was lucky in meeting good university teachers, colleagues and friends, with whom he had an understanding in those years in the past when the climate on the Czech political scene was not happy for the development of oriental studies or for quantitative linguistics. A deep personal friendship and a close scientific and philosophical relationship with Germany enriched his life and helped him to overcome difficult moments.

For his whole life, Hřebíček has been deeply absorbed and deeply engaged in his linguistic activities. He finds real happiness and life sense in doing scientific research - with no desire of fame. It was not easy to persuade him to agree to a festschrift as a tribute to him. The editors express thanks to him for placing his personal and bibliographic data at their disposal.

The editors

Selected Bibliography

LUDĚK HŘEBÍČEK

ArOr - *Archív orientální.* Quaterly journal of African and Asian studies published in English, German and French.

NO - *Nový Orient.* Monthly journal for popularization of the Asian and African cultures. Published in Czech.

JQL - *Journal of Quantitative Linguistics.* Official journal of the International Quantitative Linguistics Association. Published by Swets & Zeitlinger.

Monographs, Textbook, Editorship

Turečtina. [Turkish textbook.] Praha, Academia, 1969, 56 pp.

Turkish grammar as a graph. (Dissertationes Orientales 31) Prague, Academia, 1971, 148 pp.

Quantities of social communication. With general applications to Islam and social morphogenesis. (Dissertationes Orientales 43) Praha, Orientální ústav ČSAV, 1986, v+197 pp.

Text in communication: supra-sentence structures. (Quantitative Linguistics, Vol. 48) Bochum, Brockmeyer, 1992, 115 pp.

L. Hřebíček & G. Altmann (eds.), *Quantitative text analysis.* Trier, Wissenschaftlicher Verlag Trier, 1993, 307 pp.

A. Křikavová & L. Hřebíček (eds.), *Ex Oriente. Collected papers in honour of Jiří Bečka.* Prague, Oriental Institute, 1995, 209 pp.

Text levels. Language constructs, constituents and Menzerath-Altmann law. (Quantitative Linguistics, Vol. 56) Trier, Wissenschaftlicher Verlag Trier, 1995, 162 pp. [Reviewed by G. Altmann, *JQL,* 1996, Vol. 3, No. 2, 169-171.]

Lectures on text theory. Prague, Oriental institute, 1997.

[Co-authorship:]
Slovník spisovatelů národů SSSR. [The vocabulary of Soviet writers. - Kazakh and Kirgiz authors.] Praha, Svět sovětů, 1966.

Slovník spisovatelů. Asie a Afrika. [The vocabulary of writers. Asia and Africa. - Turkish writers.] Praha, Odeon, 1967.

Slovník spisovatelů. Sovětský svaz. [The vocabulary of writers. Soviet Union. - Kazakh and Kirgiz authors.] Praha, Odeon 1977.

Articles

Staroturecká a literatura střední periody. [The Old-Turkic literature and the literature of the middle period.] In: *Z dějin literatur Asie a Afriky III.* [On the history of the Asian and African literatures.] Praha, Státní pedagogické nakladatelství, 1963, 114-119.

Kazašská literatura. [Kazakh literature.] In: *Z dějin literatur Asie a Afriky III.* [On the history of Asian and African literatures.] Praha, Státní pedagogické nakladatelství, 1963, 129-139.

Kirgizská literatura. [Kirgiz literature] In: *Z dějin literatur Asie a Afriky III.* [From the history of Asian and African literatures.] Praha, Státní pedagogické nakladatelství, 1963, 140-144.

Aesthetic function of vocal harmony in the poetry of Abay Kunanbayef. *ArOr* 32, 1964, 100-103.

[L. Gržebiček:] O nekotorych količestvennych svojstvach leksiki Abaja. *Vestnik Akademii nauk Kazachskoj SSR,* 4, Altma-Ata 1964, 69-75.

Euphony in Abay Kunanbayev's poetry. *Asian and African studies,* 1, 1965, 123-130.

Alliterations in Abay Kunanbayev's poetry, *ArOr 33,* 1965, 67-72.

An attempt at quantitative analysis of rhymes (in Abay Kunanbayef's poetry). *Prague Studies in Mathematical Linguistics 1,* 1966, 105-112.

Some features of the Turkish generative system. *Linguistics 28,* December 1966, 74-81.

Russian borrowings in Kazakh. *ArOr 34,* 1966, 67-72.

Metrics of Abay Kunanbayef's poetry. *Ural-Altaische Jahrbücher,* Vol. 38, 1966, 9-12.

Are the Old-Turkic inscriptions written in verses? *ArOr 35,* 1967, 477-482.

The structure of the Uzbek stem-morpheme. *ArOr 35,* 1967, 452-462.

The phonological structure of the Turkish words. *Asian and African Studies III,* 1967, 50-60.

[L. Gržebiček:] Opyt primenenija količestvennogo metoda pri izučenii dviženija v jazyke. In: *Issledovanija po tjurkologii.* Alma-Ata, 113-120. [Quoted from: A. Ch. Džubanov, *Kvantitativnaja struktura kazachskogo teksta.* Alma-Ata 1987, 135.]

Several Turkish homonymous constructions and their generative description. *ArOr 39,* 1971, 146-154.

The Turkic first syllable and its correlation analysis. *ArOr 41,* 4, 1973, 340-349.

A method of semantic analysis of the Turkish text. With an application to a newspaper text. (Preliminary study). In: *Asian and African languages in social context.* Collected papers by V. Černý, Z. Heřmanová-Novotná, L. Hřebíček, V. Miltner, O. Švarný, P. Zima. (Dissertationes Orientales 34). Prague, Oriental Institute, 1974, 187-209.

The phonemic structure of the first syllable in several Turkic languages. In: *Researches in Altaic languages.* Budapest, Akadémiai Kiadó, 1975, 79-82.

The Turkish language reform and contemporary texts. *ArOr 43,* 3, 1975, 223-231.

The Turkish language reform and contemporary lexicon (A contribution to the description of differences between the spoken and written language.) *ArOr 45,* 2, 1977, 132-139.

The Turkish language reform and contemporary grammar. (The difference between the spoken and written texts on the level of grammatical morphemes). *ArOr 46,* 4, 1978, 334-337.

The phonological adaptations of borrowings: methodological problems. In: *Rapports, co-rapport, communications tchécoslovaques pour le IV^e congrès de Associations internationale d'études de sudest européen.* Prague 1979, 371-378.

A. Křikavová & L. Hřebíček: The educational reform in Iran. *ArOr 49,* 3, 1981, 221-239.

A. Křikavová & L. Hřebíček: Some remarks on typological classifications in social sciences. *ArOr 51,* 1983, 5-13.

Stability, morphogenesis and the developing system. (An attempt at a theoretical approach to observations). *ArOr 52,* 2, 1984, 127-141.

The class differentiation and its estimate. *Studies and documents 1,* Oriental institute, Prague 1984, 18 pp.

Text as a unit and co-references. In: Thomas T Ballmer (ed.), *Linguistic dynamics. Discourses, Procedures and evolution.* Berlin-New York, de Gruyter, 1985, 190-198.

Cohesion in Ottoman poetic texts. *ArOr 54,* 3, 1986, 252-256.

[L. Gržebiček:] Fonologičeskaja struktura tureckogo slova. In: *Novoe v zarubežnoj lingvistike. Vypusk XIX. Problemy sovremennoj tjurkologii.* Perevody s anglijskogo i nemeckogo jazykov. Obščaja redakcija akademika A.N. Kononova, sostavlenie A.N. Barulina. Moskva, Progress, 1987, 48-58. [This article was originally published in *Asian and African Studies III,* 1967, 50-60.]

A syntactic variable on the text level. In: Rolf Hammerl (ed.), *Glottometrika 10,* Bochum, Brockmeyer, 1989, 205-218.

Syntactic homogeneity of the Turkish text. In: *Rapports, co-rapport, communications tchécoslovaques pour le VI^e congrès des Associations internationales d'études de sud-est européen.* Prague, L'Institut de l'histoire, 1989, 227-236.

Quantitative studies. In: György Hazai (ed.), *Handbuch der türkischen Sprachwissenschaft,* Teil I. Budapest Akadémiai Kiadó, 1990, 371-387.

Menzerath-Altman's law on the semantic level. In: L. Hřebíček (ed.), *Glottometrika 11,* Bochum, Brockmeyer, 1990, 47-56.

The constants of Menzerath-Altmann law. In: Rolf Hammerl (ed.), *Glottometrika 12,* Bochum, Brockmeyer, 1990, 61-71.

[L. Gržebiček:] Ėstetičeskie funkcii garmonii glasnykh v poezii Abaja Kunanbaeva. In: *Qazaq teksining statistikasy.* II šygharyluy. Almaty, Ghylym, 1990, 49-53. [This article was originally published in *ArOr 32,* 1964, 100-103.]

The relation "to consist of". In: Jiří Prosecký (ed.), *Ex Pede Pontis. Papers presented on the occasion of the 70th aniversary of the foundation of the Oriental institute.* Prague, Oriental Institute, 1992, 80-85.

Predication in Turkish and the segmentation of text. *ArOr 61,* 4, 1993, 412-418.

Text as a construct of aggregations. In: R. Köhler & B. B. Rieger (eds.), *Contributions to quantitative linguistics.* Dordrecht, Kluwer, 1993, 33-39.

L. Hřebíček & G. Altmann: Prospects of text linguistics. In: L. Hřebíček & G. Altmann (eds.), *Quantitative text analysis.* Trier, Wissenschaftlicher Verlag Trier, 1993, 1-28.

Text as a strategic process. In: L. Hřebíček & G. Altmann, *Quantitative text analysis.* Trier, Wissenschaftlicher Verlag Trier, 1993, 136-150.

Fractals in language. *JQL,* 1, 1994, 82-86.

The stairway of subsystems. In: A.A. Polikarpov (ed.), *QUALICO 94. Abstracts of papers.* Moscow 1994, 210-222.

Interpretation and equilibrium of a text. In: A. Křikavová & L. Hřebíček (eds.), *Ex Oriente. Collected papers in honour of Jiří Bečka.* Prague, Oriental Institute, 1995, 63-73.

Phase transition in texts. *ZeT - Zeitschrift für empirische Textforschung 2,* 1995, 52-58.

L. Hřebíček & G. Altmann: The levels of order in language. In: P. Schmidt (ed.), *Glottometrika 15.* Trier, Wissenschaftlicher Verlag Trier, 1996, 38-61.

Word associations and text. In: P. Schmidt (ed.), *Glottometrika 15.* Trier, Wissenschaftlicher Verlag Trier, 1996, 96-101.

G. Altmann, E. Erat & L. Hřebíček: Word length distribution in Turkish texts. In: P. Schmidt (ed.), *Glottometrika 15.* Trier, Wissenschaftlicher Verlag Trier, 1996, 195-204.

Word frequency and word location in a text. *ArOr 64,* 3 1996, 339-347.

Text and interpretation. In: Petr Zemánek (ed.), *Studies in Near Eastern languages and literatures. Memorial volume of Karel Petráček.* Prague, Oriental Institute, 1996, 279-286.

Persistence and other aspects of sentence-length series. *JQL,* 1997, Vol. 4, No. 1-3 (*Festschrift in honour of Juh. Tuldava.* Eds.: G. Altmann, J. Mikk, P. Saukkonen, G. Wimmer), 103-109.

Book Reviews

Kniha o Menzerathově-Altmannově zákonu. [Review of: G. Altmann, M.H. Schwibbe et al., *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim-Zürich-New York, Olms, 1989, 132 pp.] - *Slovo a slovesnost,* 4, LII - 1991, 310-311.

Marie Těšitelová, *Quantitative linguistics.* Praha, Academia, 1992, 253 pp. - G. Altmann (ed.), *Glottometrika 14,* Trier, Wissenschaftlicher Verlag Trier, 1993, 213. - *ArOr 61,* 3, 1993, 333.

M.G. Boroda, *Musikometrika 4.* (Quantitative Linguistics, Vol. 50), Bochum, Brockmeyer, 1992, 205 pp. - *ArOr 61,* 3, 1993, 333.

Gunnar Jarring, *Cultural clash in Central Asia. Islamic views on Chinese theatre.* Stockholm 1991, 40 pp. - *ArOr 61,* 3, 1993, 333-334.

Gunnar Jarring, *Garments from top to toe.* Stockholm 1992, 93 pp. - *ArOr 61,* 3, 1993, 333.

Juhan Tuldava, *Methods in quantitative linguistics.* Trier, Wissenschaftlicher Verlag Trier, 1995, viii+187 pp. - *JQL,* Vol. 2, No. 2, 168-171.

Reinhard Köhler, with the assistence of Cristiane Hoffmann, *Bibliography of quantitative linguistics.* Amsterdam - Philadelphia, Benjamins, 1995, LII+750 pp. - In : P. Schmidt (ed.), *Glottometrika 15.* Trier, Wissenschaftlicher Verlag Trier, 1996, 214-217.

Gabriel Altmann, *Statistik für Linguisten.* Trier, Wissenschaftlicher Verlag Trier, 1995, 243 pp. - *ArOr 64,* 1996, 290-291.

Altmannova učebnice statistiky pro linguisty. [Review of: G. Altmann, *Statistik für Linguisten.* Wissenschaftlicher Verlag Trier, 1995.]

Popular Articles

Pohled skulinou na tureckou literaturu. [Peeping at the Turkish literature.] *Literární noviny* roč. 7, čís. 4, 25.1.1958, 8.

The forefather of the Turks. *New Orient Bimonthly,* 2/6, 1961, 179.

How the Turkish nomads measured time. *New Orient Bimonthly*, Vol. 2, No. 2, April 1961, 11.

Slon v porcelánu aneb matematika ve filologii. [An elephant in porcelain or Mathematics in philology.] NO 8/1964, 252-253.

Z Ašchabádu do Taškentu. [From Askhabad to Tashkent.] NO 8/1967, 253-254.

Turecko a Západ. [Turkey and the West.] *Světová literatura,* 1, 1968, 149-151.

Drobty z tureckého jazykového sendviče. [Crumbs from the Turkish language sandwich.] NO 9/1971, 266-268.

O tajích fonetiky. [The secrets of phonetics. Interview with dr. O. Švarný.] NO 5/1972, 156-159.

Co soudíte o transformační gramatice. [What do you think about transformational grammar.] NO 1/1972, 2.

Významy a lidé. [Meanings and people.] NO 3/1972, 92-93.

Turecký karikaturista Turhan Selçuk. [The Turkish caricaturist T. S.] NO 4/1972, 107.

Dialog o měření básně. [Dialogue on measuring a poem.] NO 4/1973, 108-110.

Jazyk a politika. [Language and politics.] NO 3/1974, 85-87.

Dialog o měření významu. [Dialogue on the measurement of meaning.] NO 1/1975, 19-20.

Dialog o měření stylu. [Dialogue on the measurement of style.] NO 4/1975, 122-123.

Metafora jako hádanka - hádanka jako metafora. [Metaphor as riddle and riddle as metaphor.] NO 9/1975, 271-272.

Chvála turecké řeči. [Eulogy of the Turkish language.] NO 10/1975, 301-303.

Uzbecká rodina. [The Uzbek family.] NO 2/1975, 44-46.142.

Necati Cumali a téma Anatolie v moderní turecké literatuře. [About the contemporary Turkish novel 'Rains and Earth' by Necati Cumali.] *Světová literatura,* 4 1975, 217-234.

Karagöz neboli Černoočko. [The Turkish folk theater Karagöz.] NO 9/1976, 277-279.

Kazašský režisér v pražském Národním divadle. [Kazakh director in the National Theater of Prague.] NO 4/1976, 112-115.

Kirgizská hudba. [Kigiz music. About the epic of Manas.] Czech Broadcasting, Vltava, March 1976.

Ortaoyunu neboli "hra na place". [The Turkish folk theatre Ortaoyunu.] NO 5/1977, 144-147.

Meddah neboli vypravěč. [The Turkish folk narrator Meddah.] NO 6/1977, 177-179.

Jméno Gunnar Jarring. [About the distinguished Turkologist and diplomatist G. J.]. NO 9/1977, 279.

Podobnost a věda. [Similarity and science.] NO 10/1980, 308-310.

Systémy a jejich řeč. [Systems and their language.] NO 6/1981, 179-180.

Turecko 1980: boj o spisovný jazyk pokračuje. [Turkey 1980: the struggle for the literary language goes on.] NO 8/1981, 231-232.

O jednom teorému a jeho důsledcích. [About a theoreme and its consequences.] NO 3/1983, 77-78.

Data se špatnou pověstí. [Data with ill fame.] NO 9/1983, 275-277.

Nad ujgurskými příslovími. [Uyghur proverbs.] NO 4/1986, 106-108.

Vědecké zákony a společenské vědy. [Scientific laws and the humanities.] NO 9/1986, 277-279.

Kazach, Kazak, Qazak. In: *Studna v poušti. Prózy spisovatelů Kazachstánu.* [A well in the desert. Tales by Kazakh writers. - Epilogue to the book.] Praha, Mladá fronta, 1986, 229-234.

Metodologie a morálka. [Methodology and moral.] NO 8/1990, 243-244.

Racionální iracionalita. Na okraj veršů Yunusa Emreho. [Rational irrationality. In margine of the verses by Y. E.] NO 9/1990, 277-278.

Aus dem Notizbuch eines Linguisten, der an der glücklichen Zukunft der Menschheit baute. In: R. Grotjahn, S. Kempgen, R. Köhler, W. Lehfeldt (eds.), *Viribus Unitis. Festschrift für Gabriel Altmann zum 60. Geburtstag.* Trier, Wissenschaftlicher Verlag Trier, 1991, 103-107.

Gabriel Altmann čili Hledání teorie jazyka. [G. A. or The search for the theory of language.] NO 4/1991, 102-103.

Jak překládat gazel? [How to translate a ghazal?] NO 5/1991, 153-154.

Několik poznámek k pošetilostem vědy. [Some remarks on the foolishness of science.] *Universum 12,* 1993, 30-32.

Jazyk v zrcadle vědeckého poznání. Předneseno u příležitosti udělení zlaté medaile Dobrovského Gabrielu Altmannovi Čs. akademií věd. [Language in the mirror of scientific knowledge. On occasion of the award of Dobrovsky's Golden Medal to Gabriel Altmann by the Academy of Sciences of the Czech Republic.] NO 2/1993, 35-37.

Yunus Emre v pojetí Annemarie Schimmelové. K mystickým tématům včerejška i dneška. [Y. E. in the conception of A. S. On the mystic themes of yesterday and today.] NO 6/1995, 227-230.

Písmo v pohybu. [Writing in movement.] NO 5/1996, 170-173.

Řád textu. [The order of text.] NO 8/1996, 317-319.

Význam v poetice, mystice a jinde. [Meaning in poetics, mysticism and elsewhere.] NO 9/1997, 352-354.

Translated novels and works for the stage

Abaj Kunanbajev, *Čtyřicet rozjímání o životě a lidech.* [Forty meditations about life and humans. Translation from Kazakh.] Praha, Svět sovětů, 1959.

Orhan Veli, *Čeho se nemohu zříci.* [What I cannot give up. Translation of Turkish poems.] Praha, Státní nakladatelství krásné literatury a umění, 1964, 130 pp.

Melih Cevdet Anday, *Jak se hraje mikádo.* [How to play Mikado. Translation of a Turkish drama.] Praha, Dilia, 1970, 82 pp.

Krev a pot. [Blood and sweat. Dramatization of the novel by Ä. Nurpeisov. Translation from Kazakh.] National Theater of Prague, first night on March 1, 1976.

Yasar Kemal, *Jeřábi se zlatými pery.* [The Turkish novel „Ortadirek".] Praha, Odeon, 1979, 312 pp.

Aziz Nesin, *Ulice Istanbulu.* [The streets of Istanbul. Memoirs of a Turkish writer.] Praha, Odeon, 1983, 405 pp.

Note: The complete bibliography contains 107 book reviews, 74 popularizing articles and 101 contributions for different Czech journals which mostly are translations from different Turkic languages.

Theory Building in Text Science

Gabriel Altmann

1. Theory

The aim of this article is to elaborate some principles that can be used as research guides in setting up and developing text theories.

Text, whether written or spoken, is something real. Thus in agreement with the ontological principle "every thing abides by laws" (cf. Bunge, 1977:16) we can assume that texts also abide by laws. If we differentiate between objective laws representing real mechanisms generating observable phenomena and scientific laws represented by statements, we can accept without hesitation that both with text generation and text reception not only rules/conventions hold but also laws are active. These capture mental, social, physiological, communicative, linguistic, aesthetic, etc. mechanisms controlling the generation and processing of texts. The aim of any future text theory is formulating statements about these mechanisms and joining them in a consistent system. Statements of this kind must be very general (i.e. holding for all texts), they must be testable or already well corroborated and they must be won deductively, i.e. derived from some few basic assumptions (in progressed cases from axioms) (cf. Bunge, 1967).

On the way to this aim we are at once confronted with the following circumstances:

(i) No empirical theory contains only deductive statements. Many of them are merely empirical generalizations - even in physics - thus in empirical sciences we strive rather for inductive-deductive theories. But every theory should contain at least one law (cf. Galtung, 1967).

(ii) Mature theories are axiomatized, i.e. they contain some non-derived statements from which all the others follow. The greatest difficulty in text theory is the problem of a starting point. But once found, it is not difficult to find its consequences.

(iii) Since every theory encompasses merely a part of reality, merely a restricted aspect of the object under investigation, it is possible to set up as many theories of the given object as there are aspects we are able to conceive. This makes the decision in point (ii) easier: one can begin with whatever aspect. All sciences developed in this way. There is no "central" aspect and there are no "essential" parts of things.

(iv) In texts some rules are temporarily deterministic but the background mechanisms are stochastic, thus a future text theory will consist of probabilistic statements, even if we use for representation, description or explication qualitative (even mathematical) concepts.

2. Levels and units

Since every thing is a system, text can be considered as such, too. At the present state of science this is a relatively sure starting point. Bunge (1983:267-270) recommends directly to "*Study every entity as a system or a component of such*" which is a principle accepted not only in synergetic linguistics but in all advanced empirical sciences. Of course, text is a very complex system that can function only if it is organised hierarchically, i.e. it has different levels and at each level there are elements endowed with properties and linked by different relations. Some levels and their elements (called units) have been established in linguistics by different, mostly *ad hoc*, criteria that changed from language to language. However, there is a possibility of beginning here quasi-axiomatically using a result following from Hřebíček's works (1995, 1997). We shall call it *Hřebíček's conjecture*:

> *Let there be some (hypothetical) text constructs composed of some (hypothetical) components. If the size of the components is a function of the size of the constructs, according to Menzerath's law, then both the constructs and the components are textual units and they lie on two different levels.*

In practice, this means that the size of an immediate component (y) is a power function of the size of the construct (x), i.e. $y = Ax^{-b}$. This conjecture has been corroborated on all levels in many languages (cf. Menzerath, 1964; Gerlach, 1982; Köhler, 1982; Altmann & Schwibbe, 1989; Hřebíček, 1997) and led Hřebíček to the discovery of the referential level, lying between sentence and text. The pertinent units of the referential level consist of sentences and are called *hrebs*. The above conjecture can be used as a criterion for the identification of units - perhaps the only lawlike one besides many conventional ones. We realise that there is a difference between (real) language entities and conceptually constructed textual units used in our analyses. If there are competing decisions about the units on the same level, conventional criteria do not allow clear decisions to be made since such criteria can be replaced by other conventional ones. However, the above conjecture enables us to decide unambiguously: that of two segmentations, that one is "better" (i.e. more fruitful for a theory) which better follows Menzerath's law. A construct can, of course, have components of different kinds, e.g. the components of a word can be syllables, morphemes,

phonemes, but they can be accepted all side by side. Ambiguity on the same level exists merely between different segmentations of the component units, e.g. diphthongs or affricates considered as one or two units. For theoretical purposes units must be prolific in the sense that it is possible to set up laws for their structure, behaviour and links. For descriptive purposes, e.g. for grammar, quite other units can be prolific and necessary.

3. Properties

Textual units are endowed with real properties. But what we assign to these units, are our conceptual creations arising through our interaction with these entities. Thus, as a matter of fact, each unit can have an infinite number of properties. We are even the originators of many real properties, without us no word would have semantic, emotional, aesthetic etc. properties. But whatever the origin of the properties, it holds that

All properties are in principle measurable

and if not, then according to Galilei they must be made measurable. A measure, however, is a very abstract concept, our creation, and not an inherent property of things. The measurement of properties is important not only for exact characterization or classification, but above all for discovering the behavior and interrelations between properties, testing hypotheses and giving the theory a solid background.

There are some properties shared by many units but different units can have their particular properties. Some of them have been studied very thoroughly. Thus each unit has a physical size. Beginning with the morpheme all higher units have, in addition, meaning whose extent can be defined as complexity. The higher the unit the more properties it seems to have, or better, the more properties we are able to discern. Thus words have not only size, semantic complexity and grammatical properties but also emotionality, polyanna, connotative complexity, their use refers to psychological properties, attitudes, intelligence etc. of the author (cf. Schwibbe, 1984). Each higher unit has all the properties of the lower ones and, in addition, some further properties generated by the interaction of lower units, as is usual in hierarchical systems, or by operations on them. Some properties are individual, i.e. measurable directly on the respective units, others are collective, measurable on sets, e.g. average, entropy.

As a matter of fact, what we work with are not real properties but attributes i.e. our conceptual constructs. This fact can clearly be seen on properties like average (e.g. average length) which is not associated to any real entity, or on numerical concepts we *ascribe* to properties. Now, since textual units can have an infinite (countable) number of properties/attributes, our only problem is to

decide which of them are worth of being measured, analysed and described. The following principle can be of help:

Those properties are scientifically prolific for which we can find laws.

Each aspect of a textual entity is associated with some laws, i.e. mechanisms that generate it, but there are aspects that are more easily available than other ones, e.g. those that are assessed from responses of test persons and not from direct text analysis. Because of a long tradition in analysing some aspects, e.g. grammar, our search for laws became blind, the majority of researchers looks for grammatical rules. But it can be shown that even here laws rule (cf. Köhler, 1999). The following fundamental hypotheses concern all properties:

(i) *All properties of text units are random variables following in text a "proper" distribution.*
(ii) *Since text is a linear formation, each property of text units forms some linear patterns, whether deterministic, stochastic or chaotic.*
(iii) *No property of text units is isolated. It is linked in a lawlike manner with at least one other property (colateral relations).*
(iv) *Each property of text units gives rise (or contributes) to a special quality of the given text (hierarchical relations).*
(v) *In the history of a language no text property is constant, each of them develops.*

These hypotheses can be subsumed under a more general statement claiming that all units, their properties, behaviours, links and changes create some kind of order (cf. Hřebíček & Altmann, 1996, 1993).

Since observations corresponding to these hypotheses are measured entities, it follows in turn that:

Lawlike statements in text theory should have a mathematical form.

Hypotheses (i) and (ii) consider the behaviour of entities, (iii) and (iv) their systemic character, and (v) their evolution. Let us consider them separately.

4. Distribution

Units occur in text repeatedly. Of course, there are units that occur merely once, e.g. sentences, some words, etc. This is, however, no restriction to the concepts of repetition and frequency because the number one also belongs to the domain of the given variable. Units can be assigned to different classes that in turn can be ranked according to a chosen criterion. In that case we obtain a rank-order

distribution. Usually, these distributions can be transformed into the well-known frequency (spectrum) distributions (cf. Haight & Jones, 1974; Baayen, 1989; Chitashvili & Baayen, 1993; Zörnig & Boroda, 1992). Both kinds can be modelled mathematically and are candidates for law statements (for more details see Tuldava, 1998). Of course, frequency spectra can be modelled independently of rank-order distributions and if set up deductively, they bring to light at least a part of the mechanism effecting the repetition (cf. Wimmer et al., 1994; Wimmer & Altmann, 1996). The parameters of the distributions represent "forces" operating in text (cf. Altmann & Köhler, 1996). In addition, rank-order distributions can be used as criteria of "correct" classification of entities, a fact that can be expressed in a tentative principle:

If some entities of text are ordered in classes then the classification is theoretically prolific only if it follows a "proper" ranking distribution.

It is not yet clear which ranking distributions are adequate - there are too many models - we merely know that they must be monotone decreasing (and in no case discrete uniform).

The shape of some distributions is characteristic not only of the given entity but can tell us something about a property of text according to hypothesis (iv). Consider Hřebíček's hrebs consisting of sentences containing the same word or the same concept or its reference. In a highly coherent text the sentences are linked to long hrebs and there is a small number of hrebs with few sentences (short hrebs). Since the distribution of hrebs is always monotone decreasing, the more coherent a text the "flatter" is the distribution of hreb lengths, i.e. the smaller its excess. Thus text coherence (cf. Halliday & Hasan, 1976) can be expressed as a property of the distribution of hrebs. Further, since word and sentence length are the basic factors influencing text readability (cf. Tuldava, 1993), a characteristic or parameter of word and sentence length distributions indicates the measure of text readability.

5. Sequential patterning

Text entities are placed in a linear sequence ordered (a) on the surface by rules, e.g. grammatical and metrical, many of them being deterministic or almost deterministic; (b) in the deep strata by laws having a stochastic character. They form time series, periodic or chaotic oscillations, runs, patterned sequences, gaps, etc. (for references see Hřebíček, 1995, 1997; Altmann, 1988). In music they differentiate historical epochs, styles, national music, etc. Their detection can nowadays be mechanised - as far as they are known. But their detection by computer and modelling is merely a first step in theory building.

The second step is to discover their connection with the overall character of text and/or that of the personality of the author.

Different patterns can evoke the same textual effect and one special pattern can be associated with different effects. The research in this direction is still embryonic. The best known cases are the linear patterns in poetry and music lending the whole a special rhythmic character. It can easily be seen that euphonic effects are evoked by a special linear patterning of phonemes or their combinations. Again, the linear ordering within Hřebíček's hrebs is responsible for a special aspect of text coherence. In general, hrebs are discontinuous units with gaps between individual sentences. The more the writer appears to be concentrating on the constituting concepts of a given hreb the smaller the gap between the sentences of the hreb will be. This is a kind of semantic alternative to Skinner's principle of formal reinforcement. Particular forms of linear patterning are symptoms of psychic states or diseases, of age, sex, intelligence, character features, associations etc.

6. Systemic character

In part 4 and 5 we have shown several cases of links between patterns of units and the character of text. Sometimes the patterns seem to "cause" something, to "give rise" to a special property of the text, in other cases it is "evident" that they are merely symptoms of something, but in all cases they seem to "evoke" an effect in the reader. It would not be appropriate to speak here about "causality" though the rather lax use of "to cause" is not ruled out. It is not simple to point to the "direct" cause. Which of the four Aristotelian causes (materialis, formalis, efficiens, finalis) is involved in the actual case? Which cause is necessary and which is sufficient? Multiple causes, multiple effects, causal chains, etc. belong to this jargon. For theoretical purposes in this domain it is more appropriate to speak about influence, association, relation, etc. (on causality see Bunge, 1963, 1979, 1983a:310ff, 1983b:25ff; Riedl, 1982). We know that the author controls the text but at each moment there is feedback, felt by all writers; we know that the author influences the reader but there is always feedback: the author tries to fulfil the reader's expectation; we know that there are interactions between text levels, etc. Thus one should comply with the rule:

Search for interactions within levels, between levels, between author, text and environment.

In practice, this means to set up control cycles, to reconstruct the nomological net whose parts are already known in linguistics and text science (cf. e.g. Nöth, 1974; Köhler & Altmann, 1983; Köhler, 1986; Hammerl, 1991; Krott, 1994). Finally, we should obtain a net whose main parts are shown in Fig. 1

where the loops indicate self-stimulation, distribution, sequencing of equal units, and the straight lines indicate influences, associations, collateral, hierarchical and environmental relations.

In order to arrive at deep text theories we do not search for direct causes but:

Try to find the requirements of language users that influence both the properties and their links,

i.e. that lead to the rise of text mechanisms. For a set of such requirements see Köhler (1986, 1987, 1989, 1990, 1991). This all is merely the concretisation of Bunge's recommendation in part 2. Any text law is merely part of the complex nomological net that is active in text generation. Only their consistent systematisation yields a text theory, but every measurement, every ascertainment of a single arrow in Fig. 1 is an important contribution to it.

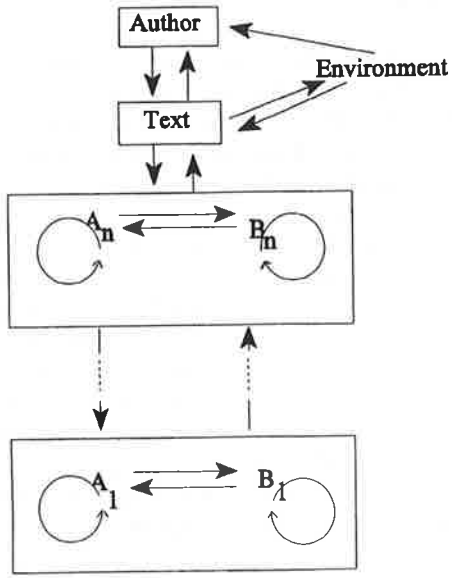


Fig. 1: The nomological net of text theory. A's and B's denote properties.

7. Change

The fact that every thing changes is an ontological principle of scientific research (cf. Bunge, 1977:16) known since Heraclitus. In order to capture a change in texts, we must process a number of texts in historical succession. But how can this be reconciled with the fact that laws are invariants? There are four ways to proceed:

(i) We admit that there are historical laws,

(ii) we admit that something in the background mechanism has changed, i.e. perhaps the objective law itself has changed,

(iii) unknown subsidiary conditions force us to admit variations in our models, i.e. perhaps the scientific law must be changed,

(iv) changes in texts are necessary due to self-organisation and self-regulation.

Let us consider briefly these possibilities in application to texts. It is not always clear which case is relevant, usually more than one is involved.

The existence of historical laws is admitted both in natural and social sciences. Radioactive decay, growth and death processes, glottochronology, increase of aspects of written communication, progress in science, etc. are lawlike processes. A part of them can be accounted for by the omnipresent self-regulation which allows a change of properties holding the relations invariant. The development of literary genres, dialects, language "types" are cases of different solutions to self-organisation problems. The relations hold but the new steady states can be fixed at different points. In Figure 2 property A decreases, property B increases in agreement with relation R which remains constant, but some aspects of properties, e.g. their distributions, have changed. We can say that the distributions etc. of properties have found a new attractor. In simpler cases our models must be slightly modified (cf. for word length in Czech texts see Uhlířová, 1996) but in cases where some components of systems are lost or new ones are added (cf. Csányi, 1989; Kampis, 1991) we must assume that the generating mechanism has changed. Thus change is always present but its background can be captured only within a systems theoretical view.

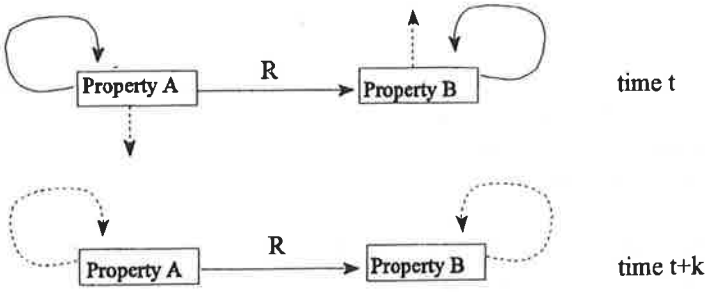


Fig. 2: Change (---) of properties at constant relations (R)

8. Summary

In order to build a text theory (i) define units and measure their properties, (ii) set up models for their different aspects (distributions, sequences, etc.) (iii) link them in a nomological net (cf. Schwegler, 1998), i.e. construct systems and (iv) explain states and changes by recourse to laws of interactions within text, between text and author and between the environment and text system.

References

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

Altmann, G., & Köhler, R. (1996). "Language Forces" and Synergetic Modelling of Language Phenomena. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 62-76), Trier: WVT.

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

Baayen, R.H. (1989). *A corpus-based Approach to morphological Productivity.* Amsterdam: Centrum voor Wiskunde en Informatica.

Bunge, M. (1963). *The Myth of Simplicity.* Englewood Cliffs: Prentice-Hall.

Bunge, M. (1967). *Scientific Research I-II.* Berlin: Springer.

Bunge, M. (1977). *The Furniture of the World.* Dordrecht: Reidel.

Bunge, M. (1979a). *The World of Systems.* Dordrecht: Reidel.

Bunge, M. (1979b). *Causality in modern Science.* New York: Dover.

Bunge, M. (1983a). *Exploring the World.* Dordrecht: Reidel.

Bunge, M. (1983b). *Understanding the World.* Dordrecht: Reidel.

Chitashvili, R.J., & Baayen, R.H. (1993). Word Frequency Distributions of Texts and Corpora as large Number of rare Event Distributions. In L. Hřebíček & G. Altmann (Eds.), *Quantitative Text Analysis* (pp. 54-135), Trier: WVT.

Csányi, V. (1989). *Evolutionary Systems: A general Theory.* Durkham: Duke University Press.

Galtung, J. (1967). *Theory and Methods in social Research.* Oslo. Universitetsforlaget.

Gerlach, R. (1982). Zur Überprüfung des Menzerathschen Gesetzes im Bereich der Morphologie. In W. Lehfeldt, & U. Strauss (Eds.), *Glottometrika 4* (pp. 95-10), Bochum: Brockmeyer.

Haight, F.A., & Jones, R.B. (1974). A probabilistic Treatment of qualitative Data with special Reference to Word Association Tests. *Journal of Mathematical Psychology,* 11, 237-244.

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

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.* (Quantitative Linguistics, Vol. 56) Trier: WVT.

Hřebíček, L. (1997). *Lectures on Text Theory.* Prague: Oriental Institute.

Hřebíček, L., & Altmann, G. (1996). The Levels of Order in Language. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 38-61), Trier: WVT.

Hřebíček, L., & Altmann, G. (1993). Prospects of Text Linguistics. In L. Hřebíček & G. Altmann (Eds.), *Quantitative Text Analysis* (S. 38-61), Trier: WVT.

Kampis, G. (1991). *Self-modifying systems: A new Framework for Dynamics, Information and Complexity.* Oxford: Pergamon Press.

Köhler, R. (1982). Das Menzerathsche Gesetz auf der Satzebene. In W. Lehfeldt & U. Strauss, (Eds.), *Glottometrika 4* (pp. 103-113), Bochum: Brockmeyer.

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

Köhler, R. (1989). Linguistische Analyseebenen, Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation. In L. Hřebíček (Ed.), *Glottometrika 11* (pp. 1-18), Bochum: Brockmeyer.

Köhler, R. (1990). Elemente der synergetischen Linguistik. In R. Hammerl (Ed.), *Glottometrika 12* (pp. 179-187), Bochum: Brockmeyer.

Köhler, R. (1991). Diversification of coding Methods in Grammar. In U. Rothe (Ed.), *Diversification Processes in Language: Grammar* (pp. 47-55), Hagen: Rottmann.

Köhler, R. (1999). Syntactic Structures: Properties and Interrelations. *Journal of Quantitative Linguistics,* 6,1, 46-57.
Köhler, R., & Altmann, G. (1983). Systemtheorie und Semiotik. *Zeitschrift für Semiotik,* 5, 424-431.
Krott, A. (1994). *Ein funktionalanalytisches Modell der Wortbildung.* Trier: unveröffentlichte Magisterarbeit.
Menzerath, P. (1964). *Die Architektonik des deutschen Wortschatzes.* Bonn: Dümmler.
Nöth, W. (1974). Kybernetische Regelkreise in Linguistik und Textwissenschaft. *Grundlagen der Kybernetik und Geisteswissenschaften,* 15, 75-86.
Riedl, R. (1982). *Evolution und Erkenntnis.* München: Piper.
Schwegler, H. (1998). The Plurality of Systems, and the Unity of the World. In G. Altmann & W.A. Koch (Eds.), *Systems. New Paradigms for the human Sciences* (pp. 165-179), Berlin: W. de Gruyter.
Schwibbe, G. (1984). *Intelligenz und Sprache.* Bochum: Brockmeyer.
Tuldava, J. (1993). The statistical Structure of a Text and its Readability. In L. Hřebíček & G. Altmann (Eds.), *Quantitative Text Analysis* (pp. 215-227), Trier: WVT.
Tuldava, J. (1998). *Probleme und Methoden der quantitativ-systemischen Lexikologie.* Trier: WVT.
Uhlířová, L. (1996). How long are Words in Czech? In P. Schmidt (Eds.), *Glottometrika 15* (pp. 134-146), Trier: WVT.
Wimmer, G., & Altmann, G. (1996). The Theory of Word Length: Some Results and Generalizations. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 112-133), Trier: WVT.
Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics,* 1, 98-106.
Zörnig, P., & Boroda, M. (1992). The Zipf-Mandelbrot Law and the Interdependencies between Frequency Structure and Frequency Distribution in coherent Texts. In B. Rieger (Ed.), *Glottometrika 13* (pp. 205-218), Bochum: Brockmeyer.

Wortlängen in Gedichten A. T. Twardowskis

Karl-Heinz Best, Svetlana Zinenko

0. Die quantitative Linguistik hat in jüngster Zeit bedeutsame Fortschritte erzielt, für die hier beispielhaft zwei Werke genannt seien, die für wesentliche Forschungskonzepte stehen mögen. Eines davon ist Bohn (1998), in dem in aller Kürze ein statistisches Portrait einer Sprache, hier des Chinesischen, entwickelt wird. Hier wird das Chinesische, bezogen auf seine Sprachebenen, sowohl hierarchisch-vertikal als auch horizontal charakterisiert. Die horizontale Betrachtung besteht darin, daß Bohn die Häufigkeitsverteilungen unterschiedlicher Einheiten hinsichtlich ihrer Komplexität modelliert, so die der Schriftzeichen, der Wortlängen, der Teilsatzlängen, etc. Die vertikale Charakterisierung entwickelt Bohn damit, daß er das sog. Menzerathsche Gesetz (auch: Menzerath-Altmannsches Gesetz, s.u.), das ja immer den Zusammenhang zwischen Konstrukt und Konstituente, also zwischen Einheiten unterschiedlicher hierarchischer Ordnung, auf beliebigen Sprachebenen untersucht (Altmann, 1980:3), und zwar bei Schriftzeichen und deren Komponenten, Wörtern, Teilsätzen und Sätzen. Bohn überprüft eine ganze Reihe spezifischer Hypothesen an chinesischem Material und kommt fast immer zu guten Ergebnissen.

Die zweite Arbeit, auf die hingewiesen sei, ist Hřebíček (1997), mit dem der Jubilar wie schon in früheren Untersuchungen (Hřebíček, 1992; 1996) seine Arbeit an der Entwicklung der Texttheorie auf der Basis des Menzerath-Altmannschen Gesetzes fortsetzt. Es handelt sich damit im obigen Sinne um eine wesentlich vertikal ausgerichtete Betrachtungsweise, bei der Hřebíček eine Vielfalt von Hypothesen entwickelt und an türkischen literarischen Texten Demir Özlüs mit Erfolg überprüft.

In einem Exkurs seines Buches (Hřebíček, 1997:13-14) stellt er die linguistische Synergetik unter Bezug auf Köhler (1986) und Altmann & Köhler (1996) kurz dar und bestimmt sie als „another thorough text model" (Hřebíček, 1997:13) neben seinem eigenen. In diesen Kontext reiht er auch die vielfältigen Untersuchungen zur Häufigkeitsverteilung von Wortlängen und anderen Spracheinheiten ein und kommt zu der Einschätzung:

„For the time being, it can be assumed that both these models are parallel or complementary. In the future, it can be expected either that one of them will appear to be more general, or that they will be united into a more adequate and explicable theory" (Hřebíček, 1997:14).

Es ist die Absicht der Verfasser dieser Arbeit, den Jubilar mit einem weiteren Beitrag zu dem von ihm als komplementär verstandenen Modell zu ehren. Als Gegenstand wurden hierzu die Wortlängenverteilungen in einem kleinen russischen Korpus, 20 Gedichten A.T. Twardowskis, ausgewählt.

1. Die Untersuchung von Wortlängenverteilungen in Texten slawischer Sprachen verdient es, auf eine breitere Basis gestellt zu werden; das gilt auch für das Russische:

Bisher sind nur wenige Arbeiten zu Wortlängenverteilungen russischer Texte verfertigt worden; sie haben gezeigt, daß die erweiterte positive Binomialverteilung an Briefe Puschkins (Stitz, 1994) und Majakowkis (Culp, 1994) sowie an verschiedenartige literarische Texte unterschiedlicher Autoren (31 Gedichte und 7 Kurzerzählungen: Girzig, 1997) fast immer mit Erfolg angepaßt werden kann. In diesen drei Arbeiten wurden die Textdateien einschließlich der nullsilbigen Wörter betrachtet.

Bei einer weiteren derartigen Untersuchung, die den Wortlängenverteilungen in Briefen des russischen Dichters A.T. Twardowski gewidmet war, konnte gezeigt werden, daß die Hyperpoisson-Verteilung an diese Texte angepaßt werden kann, während die erweiterte positive Binomialverteilung in 2 von 20 Fällen versagte (Best & Zinenko, 1998c). Dieses Ergebnis ließ sich mit nur geringfügigen Unterschieden sowohl dann erzielen, wenn die nullsilbigen Wörter berücksichtigt wurden, als auch dann, wenn sie in den Dateien der Texte nicht enthalten waren. In Ergänzung zu dieser Untersuchung sollen hier die Anpassungen der 1-verschobenen Hyperpoisson-Verteilung an die Dateien von 20 Gedichten Twardowskis präsentiert werden, wobei die Gedichte ohne die nullsilbigen Wörter bearbeitet werden, da so bessere Ergebnisse erzielt wurden als mit Berücksichtigung der nullsilbigen.

Die Gedichte Twardowskis wurden nach den gleichen Prinzipien wie schon seine Briefe bearbeitet (Best & Zinenko, 1998): Nur der laufende Text ohne die Überschriften wird erfaßt. Als „Wort" wird das orthographische Wort bestimmt; seine Länge wird nach der Zahl der in ihm enthaltenen Silben festgelegt, wobei die Zahl der Silben der Zahl der Vokale des Wortes entspricht.

2. Es folgen die Anpassungen der Hyperpoisson-Verteilung an die Dateien der Gedichte.

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

${}_1F_1(1; b; a) = 1 + a/b + a^2/b(b+1) + ...$ und $b^{(x-1)} = b(b+1)(b+2)...(b+x-2)$.

Die Symbole in den Tabellen bedeuten: x - Zahl der Silben pro Wort; n_x - beobachtete Zahl der Wörter mit x Silben im jeweiligen Gedicht; NP_x - theoretische Zahl der Wörter mit x Silben, berechnet aufgrund der jeweiligen Form der Hyperpoisson-Verteilung; a, b - Parameter der Verteilung; FG - Freiheitsgrade.

Als Prüfkriterium dafür, ob die Anpassung gelungen ist oder nicht, dient der Chiquadrattest (X^2); die Anpassung wird als erfolgreich betrachtet, wenn die Wahrscheinlichkeit P für das betreffende X^2 einen Wert von $P \geq 0.05$ erreicht; Anpassungen mit $0.01 \leq P < 0.05$ gelten noch als akzeptabel. Falls P mangels Freiheitsgraden nicht bestimmt werden kann, wird der *Diskrepanzkoeffizient* $C = X^2/N$ als Kriterium verwendet, der mit $C \leq 0.01$ eine gute Anpassung signalisiert.

3. Die Ergebnisse der Anpassung der 1-verschobenen Hyperpoisson-Verteilung an die Gedichte Twardowskis:

	Gedicht 1		Gedicht 2		Gedicht 3		Gedicht 4	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	92	88.26	21	20.71	33	33.97	38	36.37
2	86	96.97	29	28.60	33	33.87	32	29.93
3	62	48.94	16	16.99	26	18.14l	10	15.13
4	15	20.83	9	8.70	3	9.02l	9	7.57
a =	0.9336		1.0419		1.1575		1.3101	
b =	0.8497		0.7545		1.1610		1.5916	
X^2 =	6.504		0.078		0.178		2.234	
FG =	1		1		0		1	
P =	0.01		0.78				0.14	
C =					0.0019			

Gedicht 1: *O rodine* (S. 7)
Gedicht 2: *V storožke na dače...* (S. 21)
Gedicht 3: *Synu pogibšego voina* (S. 50)
Gedicht 4: *Est' čto-to v dolgoletje* (S. 73)

Die senkrechten Striche in der Tabelle des Gedichts 3 zeigen eine Zusammenfassung der betreffenden Längenklassen an; dies gilt auch in den folgenden Tabellen entsprechend.

Das Ergebnis der Anpassung der 1-verschobenen Hyperpoisson-Verteilung an die Datei des Gedichts 1 ist schwach, aber noch nicht so schlecht, daß man sie verwerfen müßte. Sie läßt sich verbessern, wenn man eine Klasse der 5-silbigen Wörter mit 0 Beobachtungen hinzufügt; man erhält dann ein $P = 0.11$.

Bei Gedicht 3 wurden die 3- und 4-silbigen Wörter zusammengefaßt; mangels Freiheitsgraden kann in diesem Fall kein P bestimmt werden; der Diskrepanzkoeffizient C zeigt aber eine gute Anpassung an.

	Gedicht 5		Gedicht 6		Gedicht 7		Gedicht 8	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	44	44.21	44	43.89	74	73.46	45	45.77
2	48	48.23	38	38.46	91	90.34	35	33.43
3	18	17.23	10	9.22	36	38.29	16	16.76
4	4	4.33	1	1.43	12	9.80	6	6.40
5					0	1.79|	2	1.97|
6					1	0.32|	0	0.51|
7							1	0.16|
a =	0.5313		0.3299		0.6468		1.5991	
b =	0.4870		0.3765		0.5260		2.1892	
X^2 =	0.059		0.194		1.205		0.201	
FG =	1		1		2		2	
P =	0.81		0.66		0.55		0.90	

Gedicht 5: *Ne mnogo nadobno truda...* (S. 82)
Gedicht 6: *Spasibo, moja rodnaja...* (S. 83)
Gedicht 7: *Porog padun* (S. 87)
Gedicht 8: *Ta krov', čto prolita nedarom...* (S. 91)

	Gedicht 9		Gedicht 10		Gedicht 11		Gedicht 12	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	79	78.98	42	41.97	34	33.70	33	33.50
2	50	48.87	31	30.97	29	30.30	43	43.66
3	11	12.75	14	14.01	13	11.01	25	23.00
4	3	2.40	6	6.05	1	2.50|	9	9.84
5					1	0.49|		
a =	0.4512		1.1693		0.6099		0.8843	
b =	0.7292		1.5842		0.6784		0.6786	
X^2 =	0.423		0.000		0.742		0.261	
FG =	1		1		1		1	
P =	0.52		0.99		0.39		0.61	

Gedicht 9: *Ty i ja* (S. 99)
Gedicht 10: *O suščem* (S. 101)
Gedicht 11: *Ustalost'* (S. 113)
Gedicht 12: *Svidetelstvo* (S. 114)

	Gedicht 13		Gedicht 14		Gedicht 15		Gedicht 16	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	44	43.08	91	90.20	49	47.97	41	41.33
2	47	49.67	67	73.27	40	43.79	43	43.34
3	30	26.32	44	36.16	28	24.75	23	22.30
4	7	9.05	13	12.82	11	10.13	8	7.60
5	3	2.88	2	4.55	3	4.36	2	2.43
a =	0.9808		1.2578		1.4839		1.0104	
b =	0.8507		1.5484		1.6256		0.9634	
X^2 =	1.150		3.659		1.267		0.118	
FG =	2		2		2		2	
P =	0.56		0.16		0.53		0.94	

Gedicht 13: *Gornye tropy* (S. 138)
Gedicht 14: *Kosmonavtu* (S. 141)
Gedicht 15: *Est' knigi...* (S. 147)
Gedicht 16: *Na novostrojkach v...* (S. 153)

	Gedicht 17		Gedicht 18		Gedicht 19		Gedicht 20	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	69	68.61	73	70.79	28	25.76	17	16.60
2	54	53.70	72	75.64	34	40.37	36	35.16
3	28	28.76	49	46.84	30	20.71	18	19.91
4	16	11.70	20	20.42	6	7.78	7	6.51
5	0	3.84	9	9.31	1	2.38	1	1.49
6	1	1.39					1	0.33
a =	1.6964		1.4731		0.8778		0.7734	
b =	2.1676		1.3786		0.5602		0.3652	
X^2 =	5.542		0.361		4.728		0.272	
FG =	3		2		2		2	
P =	0.14		0.83		0.09		0.87	

Gedicht 17: *A ty samich poslušaj...* (S. 154)
Gedicht 18: *Bereza* (S. 180)
Gedicht 19: *Takoju otmečen ja dolej...* (S. 182)
Gedicht 20: *Čut' zacvetet ivan-čaj...* (S. 193)

4. Man kann nun feststellen, daß die 1-verschobene Hyperpoisson-Verteilung in allen Fällen an die Dateien der Gedichte ohne die nullsilbigen Wörter angepaßt werden kann. Das Ergebnis ist im Fall des ersten Gedichts gerade noch akzeptabel; in den übrigen Fällen entspricht es den angegebenen Kriterien.

Die gleichen Gedichte wurden noch auf andere Weise untersucht, ohne daß dies hier im Einzelnen dokumentiert wird. Es ging dabei um folgende Aspekte: Die Anpassung der erweiterten positiven Binomialverteilung an die Dateien ohne Berücksichtigung der nullsilbigen Wörter lieferte einige inakzeptable Ergebnisse. Bei der Untersuchung der Dateien der Gedichte einschließlich der nullsilbigen Wörter konnten sowohl mit der erweiterten positiven Binomialverteilung als auch mit der Hyperpoisson-Verteilung in allen Fällen gute Anpassungen erzielt werden.

Die Hyperpoisson-Verteilung scheint zumindest bei den Briefen (Best & Zinenko, 1998c) und Gedichten Twardowskis gegenüber der sonst bei russischen Texten mehrmals bewährten erweiterten positiven Binomialverteilung (Girzig, 1997) das bessere Modell zu sein. Dies gilt für die Anpassung an Dateien mit und ohne nullsilbige Wörter. Die erweiterte positive Binomialverteilung hat, verglichen mit der Hyperpoisson-Verteilung, den Nachteil, daß sie einen Parameter mehr aufweist und damit an Texte mit nur wenig Längenklassen, zu denen auch die Gedichte Twardowskis gehören, weniger gut anzupassen ist.

Die Untersuchungen zum Russischen haben gezeigt, daß an alle bisher untersuchten Texte eine der von Wimmer u.a. (1994) und Wimmer & Altmann (1996) entwickelten Verteilungen angepaßt werden können. Das Russische macht dabei insofern einen recht einheitlichen Eindruck, als anscheinend nur zwei der vielen theoretisch begründeten Modelle für so unterschiedliche Textgattungen wie Briefe, Erzählungen und Gedichte benötigt werden (zur Modellierung vgl. auch Altmann, 1991). Die hier vorgestellten Ergebnisse zu den Gedichten (und Briefen) Twardowskis stimmen mit denen entsprechender Untersuchungen zum Ukrainischen überein; auch in diesem Fall konnte die Hyperpoisson-Verteilung mit Erfolg angepaßt werden (Best & Zinenko, 1998a,b). Zu den andern slawischen Sprachen vergleiche man die Hinweise bei Girzig (1997:152, Fußnote 1).

Als generelles Ergebnis läßt sich feststellen: Die Hypothese, daß die Häufigkeitsverteilung von Wörtern verschiedener Länge in Texten gesetzmäßig geregelt ist, hat sich auch in diesem Fall - wie schon bei vielen andern Sprachen (Best & Altmann 1996; Best, 1998) - vollauf bewährt. Es gibt außerdem eindeutige Hinweise darauf, daß dies nicht nur für Wortlängen, sondern auch für Satzlängen (Altmann, 1988: 57ff.; Niehaus, 1997) und weitere Spracheinheiten (Best, 1998) gilt.

Quelle

Twardowski = Tvardovskij, Alexandr Trifonovič. 1978. Sobranie sočinenij v 6 tomach. Tom 3. Stichi (1946-1970). Moskva: „Chudožestvennaja literatura".

Literatur

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

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

Altmann, G. (1991). Modelling Diversification Phenomena in Language. In U. Rothe (Hrsg.), *Diversification Processes in Language: Grammar* (S. 33-46), Hagen: Margit Rottmann Medienverlag.

Altmann, G., & Köhler, R. (1995). „Language Forces" and Synergetic Modelling of Language Phenomena. In P. Schmidt (Hrsg.), *Glottometrika 15* (S. 62-76), Trier: WVT.

Best, K.-H. (1998). Results and Perspectives of the Göttingen Project on Quantitative Linguistics. *Journal of Quantitative Linguistics,* 5,3, 155-162.

Best, K.-H., & Altmann, G. (1996). Project Report. *Journal of Quantitative Linguistics,* 3, 85-88.

Best, K.-H., & Zinenko, S. (1998a). Wortkomplexität im Ukrainischen und ihre linguistische Bedeutung. *Zeitschrift für slavische Philologie* (im Druck).

Best, K.-H., & Zinenko, S. (1998b). Wortlängen in Gedichten des ukrainischen Autors Ivan Franko. In *Festschrift für Viktor Krupa* (im Druck).
Best, K.-H., & Zinenko, S. (1998c). Wortlängenverteilungen in Briefen A.T. Twardowskis. *Göttinger Beiträge zur Sprachwissenschaft*, 1, 7-19.
Bohn, H. (1998). *Quantitative Untersuchungen der modernen chinesischen Sprache und Schrift*. Hamburg: Kovač.
Culp, C. (1995). *Untersuchung zur Häufigkeit von Wortlängen in ausgewählten Briefen Majakovskijs*. Seminararbeit. Göttingen.
Girzig, P. (1997). Untersuchung zur Häufigkeit von Wortlängen in russischen Texten. In K.-H. Best (Hrsg.), *Glottometrika 16* (S. 152-162), Trier: WVT.
Hřebíček, L. (1992). *Text in Communication: Supra-Sentence Structures*. Bochum: Brockmeyer.
Hřebíček, L. (1996). *Text Levels*. 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.
Niehaus, B. (1997). Untersuchung zur Satzlängenhäufigkeit im Deutschen. In K.-H. Best (Hrsg.), *Glottometrika 16* (S. 213-275), Trier: WVT.
Stitz, K. (1994). *Untersuchung zu den Wortlängen in deutschen und russischen Briefen des 19. Jahrhunderts*. Staatsexamensarbeit. Göttingen.
Wimmer, G., & Altmann, G. (1996). The Theory of Word Length Distribution: Some Results and Generalizations. In P. Schmidt (Hrsg.), *Glottometrika 15* (S. 112-133), Trier: WVT.
Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics, 1*, 98-106.

Software

Altmann-FITTER (1994). Lüdenscheid: RAM-Verlag.

Internetadresse

Zum aktuellen Stand der Arbeiten vgl. die entsprechenden Angaben unter der Internetadresse: http://www.gwdg.de/~kbest/projekt.htm.
E-Mail: kbest@gwdg.de.

Quantitative Constraints on Case Assignment in Bilingual Discourse

Svitlana Budzhak-Jones

In recent years, a wide range of scholars – sociolinguists, psycholinguists, grammarians, and others – have investigated the simultaneous use of two or more languages in one discourse (see, e.g., Milroy & Muysken, 1995). Despite such broad, multi-disciplinary attention, however, the categorization of other-language items, especially those consisting of single words, in the discourse of another still remains one of the thorniest issues in bilingual research (Poplack & Meechan, 1998; Budzhak-Jones, 1998a). In this paper I would like to suggest one possible avenue for determining which grammar produced such items, by employing a quantitative approach.

I will concentrate on one aspect of bilingual grammar. I will examine the mechanisms of case assignment in bilingual discourse involving two typologically different languages with distinct case systems, Ukrainian and English. Making use of language specific features with respect to case, I will categorize ambiguous utterances as native or non-native, by comparing them to their monolingual counterparts in every language involved. Based on the assumption that loanwords are fully syntactically, morphologically and (sometimes) phonologically assimilated into the host language (Poplack, 1993), I will expect that borrowings will obey the rules of case assignment in exactly the same way as their native counterparts, whereas nouns which are code-switched will retain their original grammar, and will not submit to the same rules of case assignment in the same manner as host language nouns.

Theoretical background

Case is used to express grammatical relations between nouns in any language (Blake, 1994). Within Government and Binding theory (Chomsky, 1981), it has been argued that some of these dependencies are structurally determined by a Universal grammar, while others are language specific, i.e. inherent. Structural case is assigned to a noun phrase according to its position in a structural configu-

ration under government and can be overtly or covertly realized. Inherent case is peculiar to a particular language and has to be specified in the lexicon (Chomsky & Lasnik, 1991).

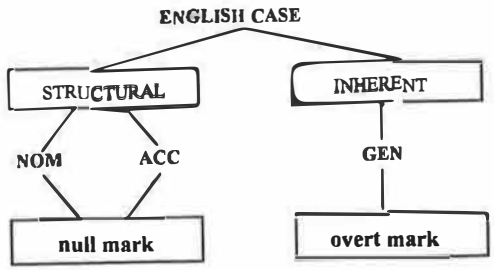


Fig. 1. English case system

In English there are three cases (Quirk et al.,1980), as shown in Figure 1. Two of them, Nominative and Accusative, are assigned structurally and remain morphologically unmarked (with the exception of some pronouns, which are overtly marked). The third one, Genitive (or possessive) is inherent and morphologically marked. In Ukrainian, one of the Eastern Slavic languages, there are seven cases (see Figure 2). Like English, Nominative and Accusative are structural cases, and they are usually morphologically unmarked. Unlike English, however, Accusative may be overtly marked, depending on a noun's gender. The other five cases in Ukrainian are inherent. Genitive and Vocative may be both morphologically marked or unmarked, whereas Dative, Instrumental and Locative are obligatorily overtly marked (Ditel',1993; Pljušč, 1994).

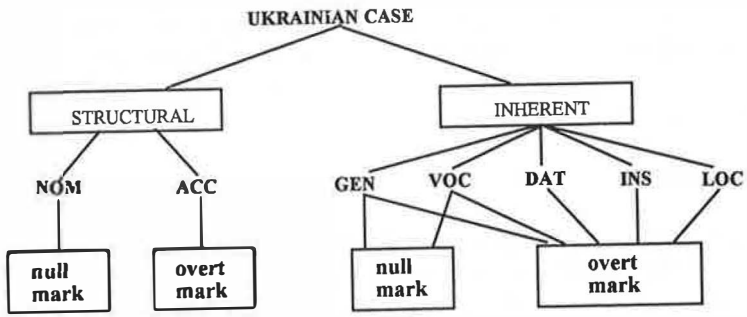


Fig. 2. Ukrainian case system.

With respect to case assigners, English and Ukrainian have some similarities, but also some notable differences (see Table 1). In both languages Nominative and Accusative are assigned similarly. The former is assigned by INFL[1] and the latter – by verbs and prepositions. In Ukrainian, however, verbs and prepositions may also assign Genitive, Dative and Instrumental. Prepositions also assign Locative. Furthermore, contrary to English, where Genitive is limited to nouns only, Genitive in Ukrainian may be assigned by a wide variety of elements, including quantifiers, cardinal numerals, interrogative pronoun *skil'ky* 'how many/much' and an empty category (in case of adjuncts). The latter also assigns Instrumental and Vocative. None of these elements can assign case in English (cf. Haegeman, 1992). Finally, unlike English, every noun in Ukrainian must establish a case concord with its modifiers. All these case relationships will be exemplified later, in the discussion of coding.

Table 1
Case assigners in English and Ukrainian.

	NOM	*ACC*	*GEN*	*DAT*	*INST*	*LOC*	*VOC*
INFL	■						
Verb		■	▒	▒	▒		
Preposition		■	▒	▒	▒	▒	
Noun			■				
Quantifier			▒				
Cardinal Numeral			▒				
Interrogative Pronoun *skil'ky*			▒				
Empty Category			▒		▒		▒

English & Ukrainian ■ Ukrainian only ▒

Data

This research is based on the data collected by the author in the Ukrainian-English bilingual community in Lehighton, Pennsylvania (USA). It comprises 36 hours of natural tape-recorded sociolinguistic interviews with 25 bilingual speakers.[2] For this project two corpora are employed: 1) monolingual Ukrainian (1951 tokens) and 2) English-origin nouns, used in an otherwise Ukrainian context (1637 tokens). Monolingual English corpus was not included in this research,

[1] INFL is an abstract element within the verb phrase (see, e.g., Chomsky, 1981, Chomsky & Lasnik, 1991).

[2] See Budzhak-Jones (1998a) for detailed discussion of the data.

since there were only two ambiguous tokens of possessive/plural marking; all the other cases were structural, and hence null marked. All nouns were extracted from the same interviews of the same informants.

Coding

All tokens in both corpora were coded for a number of factors relevant to case assignment in both languages. First, nouns were coded for their case type, i.e. structural or inherent. For example, **elevatery** in (1) and **trailju** in (2) were coded as being in a position where structural case is assigned, i.e. Nominative or Accusative. Inherent case is illustrated by **elevatera** in (1), which is assigned Genitive by a preposition *do* 'to'.

(1) *Toj* ***ukrajinec'*** *vyviv nas znajete do*
That-M.Nom Ukrainian-man-M.Nom led-out-M us you-know-pl to

takoho do ***elevatera,*** *znajete* ***elevatery*** *velyki šče*
such-M.Gen to elevator-M.Gen you-know-pl elevators-Nom big-pl.Nom still
v sudi znajete tak. (16/093)[3]
in court-house-M.Loc you-know-pl yes
'That Ukrainian man led us out to such an elevator, you know, big elevators, especially in the court house, you know, yes'.

(2) *Ja rišyv kupyty* ***trailju*** *taku na kolesach.* (24/438)
I decided-M to-buy trailer-F.Acc such-F.Acc on wheels-Loc
'I decided to buy such a trailer on wheels'.

Second, all tokens were coded for the type of a case assigners, i.e. 1) verb, 2) preposition, and 3) other. Nouns that received their case from a verb are demonstrated by **trailju** in (2). This group also includes Subjects which receive their case from INFL within a verb phrase. Nouns receiving case from a preposition are illustrated by **elevatera** in (1).[4] Tokens which were assigned case by any other elements (as discussed earlier), are demonstrated in (3) and (4). In (3), the noun **apartmentiv** is assigned Genitive by a quantifier *bahato* 'many'. And in

[3] Each example is identified by cassette number and count number. All examples are glossed with the corresponding English lexical item with grammatical labels of a Ukrainian noun, unless the grammatical information is conveyed by the translation itself. Grammatical markers are coded in the following way: F = feminine, M = masculine, N = neuter; sg = singular, pl = plural; Nom = nominative, Acc = accusative, Gen = genitive, Dat = dative, Ins = instrumental, Loc = locative, 1, 2, 3 = person, Ø = missing overt inflection. Since grammatical gender is distinguished in most cases in singular, gender marks also imply singular, unless otherwise specified.

[4] These are also shown in (5) and (6).

(4), the noun **petroliju** receives its case from another noun *zbirnykach* 'storage-tanks'.

(3) *Tam duže bahato je* ***apartmentiv*** *dorohych.* (22/387)
There very many is/are apartments-Gen expensive-pl.Gen
'There are very many expensive apartments there'.

(4) *Čolovik pracjuvav pry tych zbirnykach* **petroliju.**
Husband-M.Nom worked-M at those-Loc storage-tanks-M.Loc petroleum-M.Gen
'My husband worked at those petroleum storage tanks'. (36/067)

Third, I coded all nouns for their case assigners' ability to assign morphologically marked cases. Case assigners which can only assign morphologically null marked cases were inferred as having a covert feature, and are illustrated by **elevatery** in (1).[5] Case assigners with exclusively overt case marking are demonstrated by **elevatera** in (1), **apartmentiv** in (3) and **petroliju** in (4). All these nouns received their case from elements which can only assign morphologically marked cases. Case assigners with a 'double' feature, i.e. covert and overt morphology, are illustrated in (5) and (6). In (5) the noun **junkach** received its overt case from a preposition *na* 'on', whereas the same preposition assigned a null marked case (i.e. Accusative) to the noun **public-school** in (6).

(5) *To ja des' znajšov na* ***junkach*** [laughs] *taku staru*
It I somewhere found-M on junks-Loc such-F.Acc old-F.Acc

vannu.
bath-tub-F.Acc
'I found such old bath-tub somewhere in a junk-yard'. (25/202)

(6) *Nu to čoho ja budu ditjam posylaty na* ***public-school?***
Well then why Iwill-be-1sg children-Dat to-send to public school-M.Acc
'Well, then why would I send children to a public school?' (12/318)

Fourth, all nouns were coded for their participation in case agreement. Nouns occurring in constructions with overt modifiers, like **trailju** in (2) and **apartmentiv** in (3), were referred to as requiring case concord. Nouns without any modifiers, like **petroliju** in (4), **junkach** in (5) or **public-school** in (6), constituted the group of nouns for which case agreement was not required.

Finally, every token in both the English-origin and Ukrainian corpora was analyzed as to whether it was marked according to prescriptive rules of standard

[5] Its case assigner, INFL, may only assign structural case, i.e. Nominative.

Ukrainian, as inferred from Ukrainian grammars.[6] Nouns, like **junkach** in (5), were coded as being *standardly* marked. Nouns like *ditjam* 'children'[7] in (6) and **living-roomu**[8] in (7), were referred to as *non-standardly* marked.

(7) *A v* ***living-roomu*** *to lyšyly vse.* (30/093)
And in lining-room-M.Dat-? well left-pl all
'And in the living room we left everything.'

Analysis

The data was analyzed by the variable rule analysis, GOLDVARB 2.0 for Macintosh (Rand & Sankoff, 1990). This is a multiple regression procedure which extracts regularities from naturally occurring frequencies in the corpus-based data. It makes an assessment of the influence of different factors on a particular choice, and retains the most statistically significant factors which increase the likelihood of a dependent variant to occur. It is performed in two steps. The step-up procedure tries to find a single statistically significant factor-group, and then gradually adds other factor-groups to measure their significance. The step-down solution is based on the reversed procedure, where the likelihood of the occurrence of the dependent variable is calculated first and then factor groups are eliminated one-by-one, starting from the least significant. Finally, both steps retain the most significant factors influencing a given choice. If the factors considered in the analysis are not entirely independent, a less accurate, one level calculation can be executed, which analyzes the input of all groups simultaneously.[9]

In this research, non-standard marking of nouns across corpora was considered a dependent variable. All the relevant factors discussed above, were tested to determine their statistical significance in the occurrence of this variable. I anticipated that the same factors would be selected for both lone English-origin nouns and their monolingual Ukrainian counterparts, if the former were borrowed. Moreover, I expected that these factors would not only be the same, but that they would influence marking variability to the same extent in both corpora, irrespective of the nouns' origin. If English-origin and Ukrainian nouns were not produced by the same grammar, the former would not replicate the results of the latter. Taking into account both similarities and dissimilarities in Ukrainian and English case marking, the same factors might, or might not be selected significant

[6] See, for example, *Ukrajins'kyj pravopys* (Ditel', 1993), *Sučasna ukrajins'ka literaturna mova* (Pljušč, 1994), etc.

[7] The noun ***ditjam*** 'children'-Dat in (6) should have been prescriptively marked as *ditej* 'children'-Acc.

[8] The noun **living-roomu** in (7) should have been prescriptively marked for Locative, i.e. **living-roomi**.

[9] See Sankoff (1988) for details.

for each corpus. Likewise, the hierarchies of effect within significant factor-groups might, or might not coincide. Essentially, the more dissimilar the features between the two languages, the more dissimilar the hierarchies of effect for code-switched and native Ukrainian nouns.

Results

The results of the variable rule analysis are shown in Table 2.

Table 2
Variable rule analysis of the contribution of factors selected as significant to non-standard case marking in Ukrainian context across corpora.

	Ukrainian monolingual		*English-origin in Ukrainian*	
CORRECTED MEAN:	**.055**		**.107**	
TOTAL N:	**1951**		**1637**	
	Probability	N	Probability	N
Case type				
Structural	**.310**	(1049)	**.337**	(1141)
Inherent	**.717**	(902)	**.825**	(496)
Case assigner (by type)				
Other	**.339**	(197)		
Preposition	**.408**	(675)		
Verb	**.619**	(1079)		
Case assigner (by feature)				
Covert	**.262**	(467)	**.400**	(453)
Overt	**.432**	(503)	**.481**	(245)
Both covert and overt	**.653**	(981)	**.554**	(939)
FACTORS NOT SELECTED				
Case assigner (by type)			**X**	
Case agreement	**X**		**X**	

Non-standard case marking in monolingual Ukrainian nouns highly depends on three factor-groups: case type, case assigner's type, and case assigner's feature. English-origin nouns are influenced only by two of them, i.e. case type and case assigner's feature. Note, however, that the hierarchy of effect across both significant factor-groups is the same in both corpora. In the factor-group of case type, nouns are most likely to receive a non-standard mark when the inherent case is required (.717 in Ukrainian and .825 in English-origin), and are less so if the structural case is assigned (.310 in Ukrainian and .337 in English-origin). This is

not surprising since structural case is usually morphologically null marked in either language.

In the second significant factor-group for both corpora the probability of non-standard marking is also the lowest when the case assigner can only assign a covert case (.262 in Ukrainian, and .400 in English-origin). When the case assigner has the property to assign both covert and overt cases, the probability of non-standard marking is the highest (.653 in Ukrainian and .554 in English-origin). This result is unexpected: it would be anticipated that nouns with case assigners categorically requiring overt case marking, would show most non-standard case marks. Instead, these are nouns with case assigners which do allow some null marking. This suggests that the speakers have more marking variation when choice is offered by case assigners, rather than when their options are restricted.

Case agreement did not have any significant effect in both corpora. Case assigner's type was selected significant only for one corpus, i.e. monolingual Ukrainian. The latter may indicate that the English-origin nouns were not conditioned by exactly the same factors as the native nouns. Recall, however, that the factors which were selected significant had identical hierarchies of effect across corpora. These results are puzzling and contradictory. To resolve this dilemma, I followed Budzhak-Jones (1998a), who argues that some single word other-language incorporations may be borrowed, while the others may be code-switched. Moreover, she shows that other-language tokens with *overt* host-language morphology can only be borrowed (Budzhak-Jones, 1998b). I, therefore, separated my tokens into two subcorpora (overt and null marked), and examined the influence of different factors on marking variation in each subcorpus. The division into overtly and null marked groups, however, produced considerable interaction between the factors considered above. It prevented me from performing a two-level analysis. Hence, a less accurate, one-level procedure was executed.

Table 3 shows that all factor-groups were selected significant for marking variation of nouns with overt morphology in both corpora. Moreover, the hierarchies of effect are parallel across all factor-groups irrespective of a noun's origin, with the exception of case agreement. The difference in relative weight between the factors in this factor-group, however, is the smallest. Recall also that 1) this is a less accurate procedure, and 2) this factor-group was not selected significant when considered simultaneously with all other factor-groups for the entire corpora, as shown earlier in Table 2. Very close similarities in the behavior of overtly marked nouns in both corpora suggest that they are created by the same grammar, at least with respect to case marking.

Table 3[10]

Variable rule analysis of the contribution of factors selected as significant to non-standard case marking of overtly inflected nouns across corpora.

	Ukrainian monolingual			*English-origin in Ukrainian*		
CORRECTED MEAN:	**.059**			**.088**		
TOTAL N:	**1640**			**803**		
	Weight	Input & weight	N	Weight	Input & weight	N
Case type						
Structural	.349	**.03**	(787)	.431	**.07**	(455)
Inherent	.640	**.10**	(853)	.590	**.12**	(348)
Case assigner (by type)						
Verb	.593	**.08**	(870)	.560	**.11**	(478)
Preposition	.397	**.04**	(608)	.431	**.07**	(285)
Other	.391	**.04**	(162)	.287	**.04**	(40)
Case assigner (by feature)						
Covert	.224	**.02**	(337)	.314	**.04**	(150)
Overt	.449	**.05**	(472)	.467	**.08**	(176)
Both covert and overt	.650	**.10**	(831)	.573	**.12**	(477)
Case agreement						
Agreement required	.555	**.07**	(604)	.455	**.10**	(287)
Agreement free	.468	**.05**	(1036)	.525	**.07**	(516)

The results for null marked nouns are quite different (see Table 4). Only with respect to case type are the two corpora conditioned by the same factors in the same manner. With respect to all other factor-groups the two corpora differ: not only is there a considerable difference in the probability of occurrence of non-standard marking, but also the hierarchy is different within each factor-group. For instance, with respect to case assigner's type, the probability of non-standard marking is the highest for Ukrainian nouns when case is assigned by verbs and equals only .04. For null marked English-origin nouns, however, the probability of non-standard marking is the lowest when case is assigned by a verb (.14), and it is the highest when case is assigned by other elements (.83).

[10] Note that in Table 2 the term 'probability' was used to show the statistical likelihood of the variable to occur by inclusion/exclusion of one non-variable group at a time. In the less accurate 1 level procedure the estimated maximum likelihood of a variable is measured by all non-variable groups weights simultaneously. Teh Term 'weight' is therefore used to differentiate between two different types of results (i.e. Table 2 versus Table 3 and 4).

The results in Table 4 are important evidence that null marked English-origin nouns are not assigned case in the same way as their native counterparts. However, since there is no evidence of the behavior of monolingual English nouns with respect to the same factors of case assignment, I cannot conclude that the null marked English-origin nouns (either all or some of them) are produced by the English grammar.

Table 4
Variable rule analysis of the contribution of factors selected as significant to non-standard case marking of nouns with null morphology across corpora.

	Ukrainian monolingual			*English-origin in Ukrainian*		
CORRECTED MEAN:	**.016**			**.159**		
TOTAL N:	**311**			**834**		
	Weight	Input & weight	N	Weight	Input & weight	N
Case type						
Structural	.239	**.00**	(262)	.195	**.04**	(686)
Inherent	.998	**.89**	(49)	.999	**.99**	(148)
Case assigner (by type)						
Verb	.727	**.04**	(209)	.461	**.14**	(651)
Preposition	.099	**.00**	(67)	.567	**.20**	(165)
Other	.165	**.00**	(37)	.962	**.83**	(18)
Case assigner (by feature)						
Covert	.380	**.01**	(130)	.513	**.17**	(303)
Overt	.138	**.00**	(31)	.424	**.12**	(69)
Both covert and overt	.690	**.03**	(150)	.503	**.16**	(462)
Case agreement						
Agreement required	.468	**.01**	(136)	.509	**.16**	(284)
Agreement free	.525	**.02**	(175)	.496	**.16**	(550)

Note, however, that in Table 4 nouns in both corpora are almost categorically non-standard in the position where inherent case is required (.99 for English-origin, and .89 for Ukrainian).[11] I, therefore, excluded case type from consideration. This eliminated the interaction within each subcorpus, and allowed me to execute a more accurate, binomial procedure for the entire corpora. The results are shown in Table 5.

[11] This is not surprising since inherent cases in both languages usually disallow null marking (see Figures 1 & 2).

Table 5
Variable rule analysis of the contribution of factors selected as significant to non-standard case marking of nouns across corpora, excluding case type.

	Ukrainian monolingual		*English-origin in Ukrainian*	
	Overt	Null	Overt	Null
CORRECTED MEAN:	**.064**	**.043**	**.091**	**.169**
Case assigner (by type)				
Verb		**.697**		**.423**
Preposition		**.142**		**.738**
Other		**.177**		**.844**
Case assigner (by feature)				
Covert	**.181**	**.131**	**.293**	**216**
Overt	**.508**	**.962**	**.498**	**.994**
Both covert and overt	**.644**	**.725**	**.569**	**.523**
FACTORS NOT SELECTED				
Case assigner (by type)	**X**		**X**	
Case agreement	**X**	**X**	**X**	**X**

Overtly marked nouns in both corpora are significantly conditioned by one and the same factor-group, i.e. case assigner's ability to assign overt and/or covert cases, and in exactly the same way. If a case assigner can assign a covert case exclusively, the probability of non-standard marking is the lowest (.181 for Ukrainian, and .293 for English-origin). The probability of non-standard marking is the highest when a case assigner can assign both overt and covert cases (.644 for Ukrainian, and .569 for English-origin). Neither case assigner's type nor case agreement were significant in influencing marking variation of overtly inflected nouns. These results again demonstrate that nouns with overt Ukrainian morphology appearing in an otherwise Ukrainian context, are assigned case in exactly the same manner, irrespective of their origin. I, therefore, conclude that overtly marked English-origin nouns are produced by Ukrainian grammar, at least with respect to case.[12]

Null marked English-origin nouns differ somewhat from their overtly marked counterparts (see Table 5). Although both English-origin and Ukrainian nouns with null Ukrainian morphology are conditioned by the same two factor-groups, only one of them, the factor-group of case assigner's feature, shows exactly the same pattern across the two corpora. The probability of non-standard marking is the lowest when a case assigner can only assign covert case (.131 for Ukrainian,

[12] This is in line with Budzhak-Jones's claims that other-language tokens appearing with overt host-language morphology are borrowings (Budzhak-Jones, 1998b).

and .216 for English-origin). And it is the highest when a case assigner can only assign morphologically overt case (.962 for Ukrainian, and .994 for English-origin). With respect to case assigner's type, the two corpora differ considerably. In the Ukrainian corpus the probability of non-standard marking is the highest when case is assigned by a verb (.697), whereas in the English-origin corpus it is so when case is assigned by other elements (.844). For Ukrainian nouns the probability of non-standard marking is the lowest if case is assigned by a preposition (.142), closely followed by other (.177). For English-origin nouns, however, it is the lowest when case is assigned by a verb (.423). These results demonstrate that null marked English-origin nouns do not behave exactly in the same manner as their Ukrainian counterparts do, supporting the code-switching hypothesis. However, the lack of evidence from monolingual English tokens, and the identical hierarchy of effect across the two corpora with respect to case assigner's feature, which is very different in English and Ukrainian, does not let me conclude that these ambiguous null marked tokens are code-switched. Such similarities and differences in the behaviour of the two corpora suggest that some null-marked tokens of English origin may be produced by Ukrainian whereas others by English grammar.[13]

Summary

I have demonstrated that quantitative methodology can be an important tool for determining which language created the utterances in question. Using multivariate rule analysis I have shown that in bilingual discourse nouns exhibit the properties of the grammar by which they are generated. In Ukrainian-English bilingual discourse all overtly marked nouns irrespective of their origin, exhibited the same behavior. Their marking variability with respect to case was significantly conditioned by the same factors and in exactly the same manner. This can be taken as evidence that both English-origin and Ukrainian nouns with overt morphology were assigned case by the same grammar, i.e. Ukrainian. Hence, they were most likely borrowed.

Null marked English-origin nouns did not replicate every detail of the behaviour established by their monolingual Ukrainian counterparts. Although the same factor-groups were selected significant for marking variability in both corpora, their hierarchies of effect did not always coincide. These similarities and differences in the behavior of null marked English-origin and monolingual Ukrainian nouns suggest that the former are not monolithic by nature. It appears that with respect to case assignment some of these nouns may have retained their

[13] Using a number of morpho-syntactic and discourse criteria, Budzhak-Jones (1998a) demonstrated that indeed some null marked English-origin tokens used in otherwise Ukrainian discourse, were borrowed whereas others were code-switched.

English grammar, i.e. were code-switched; whereas others may have been generated by Ukrainian grammar, i.e. were borrowed.

References

Blake, B.J. (1994). *Case.* Cambridge: University Press.

Budzhak-Jones, S. (1998a). *Single-Word Incorporations in Ukrainian-English bilingual Discourse: Little Things mean a lot.* Ph.D. dissertation. Ottawa: University of Ottawa.

Budzhak-Jones, S. (1998b). Against word-internal Code-Switching: Evidence from Ukrainian-English Bilingualism. *International Journal of Bilingualism,* 2 (2), 161-182.

Chomsky, N. (1981). *Lectures on Government and Binding.* Dordrecht: Foris.

Chomsky, N., & Lasnik, H. (1991). Principles and Parameters Theory. In J. Jacobs, A. von Stechow, W. Sternefeld & T. Vennemann (Eds.), *Syntax: An International Handbook of Contemporary Research.* Berlin: Walter de Gruyter.

Ditel', O.A. (Ed.) (1993). *Ukrajins'kyj pravopys.* Kyjiv: Naukova dumka.

Haegeman, L. (1992). *Introduction to government & binding Theory.* Oxford, UK & Cambridge, USA: Blackwell Publishers.

Milroy, L., & Muysken, P. (Eds.) (1995). *One Speaker, two Languages: Cross-disciplinary Perspectives on Code-Switching.* Cambridge: University Press.

Pljušč, M.J. (Ed.) (1994). *Sučasna ukrajins'ka literaturna mova.* Kyjiv: Vyšča škola.

Poplack, S. (1993). Variation Theory and Language Contact. In D Preston (Ed.), *American Dialect Research: An Anthology Celebrating the 100th Anniversary of the American Dialect Society* (pp. 251-286), Amsterdam/Philadelphia: John Benjamins.

Poplack, S., & Meechan, M. (1998). Introduction: How Languages fit together in Codemixing. *International Journal of Bilingualism,* 2 (2), 127-138.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive Grammar of the English Language.* New York: Longman.

Rand, D., & Sankoff, D. (1990). *GoldVarb. A variable Rule Application for the Macintosh.* Version 2. Montreal, Canada: Centre de recherches mathématiques, Université de Montréal.

Sankoff, D. (1988). Variable Rules. In U. Ammon, N. Dittmar & K.J. Mattheier (Eds.), *Variable Rules. Sociolinguistics: An International Handbook of the Science of Language and Society* (pp. 140-61), Berlin: Walter de Gruyter.

Quantitative Analysis of Text: An Indian Experience

Amitav Choudhry, Sukesh Debnath

1. Introduction

According to Shende and Prabhu-Ajgaonkar (1989), statistical investigation of texts and text styles is more directly concerned with the description and explanation of the features inherent in the text, their organisation and variability. Ross (1950) explains the role of statistics in linguistic studies by stating that probability theory and statistics should provide the instruments or the mathematical models for testing and verifying any conclusion in linguistics which is susceptible to numerical treatment, and thus provide an auxiliary tool for linguistic research. In many quantitative studies we cannot investigate every possible example of the phenomenon we are interested in. In some cases exhaustive investigation is theoretically impossible. Therefore, whenever we wish to collect quantitative data on language we need to pay careful attention to the design of our study, and to the selection of appropriate statistical methods of summarising the data, and of testing hypotheses concerning differences between sets of data.

In the present study, based on data from a complete wordcount of Rabindranath Tagore's short stories, "Galpaguccha" (Parts I to IV)[1], the hypothesis of vocabulary balance was tested. We obviously do not know whether there is in fact such a thing as vocabulary balance between our hypothetical forces of Unification and Diversification[2] since we do not yet know whether human beings invariably economise with the expenditure of their effort; for that, after all, is what we are trying to prove. Zipf (1949:22) enumerates for the sake of clarity certain vital points:

1. We assume explicitly that human beings do invariably economise with their effort.
2. The logic of a vocabulary balance between the two forces is sound.
3. We can test the validity of our explicit assumption of an economy of effort by appealing directly to the objective facts of some samples of actual speech that have served satisfactorily in communication.
4. We may find there evidence of a vocabulary balance of some sort in respect of our two forces, and,
5. We shall ipso facto seek a confirmation of our assumption of (1) an economy of effort.

Therefore much depends on our ability to show some demonstrable cases of vocabulary balance in some actual samples of speech that have served satisfactorily in communication.[3]

2. Parameters of vocabulary balance

According to Zipf (1949) if a condition of vocabulary balance does exist in a given sample of speech we shall have little difficulty in detecting it because of the very nature and direction of the two forces involved. Along one dimension, the force of unification will act in the direction of decreasing the number of different words to one, while increasing the frequency of that one to 100%. Conversely the force of diversification will act in the opposite direction of increasing the number of different words, while decreasing their average frequency of occurrence towards one. Therefore 'number' and 'frequency' will be the parameters of vocabulary balance. It may be mentioned here that according to Bhattacharya (1965) the rank-frequency relation for words is among the most famous findings of quantitative linguistics. Briefly the finding is this: If a word-count is carried out on a sufficiently long text and the frequencies of different words occurring in

[1] „Galpaguccha" is a collection of short stories written by Rabindranath Tagore. The short stories were written over a span of approximately 56 years from 1877-1933. There are approximately 94 short stories and some of them were published posthumously. The Linguistic Research Unit of the Indian Statistical Institute has done a complete and comprehensive word frequency count of these works and stylistic and statistical analysis is being done in a phased manner. Research in this area was initiated by Professor Mahalanobis in the late 1940s.

[2] According to Zipf (1949:22) "... *we shall consistently capitalise the terms Force of Unification and Diversification, in order to remind ourselves that the Forces do not represent forces as physicists traditionally understand the term, but only the natural consequences of our assumed underlying economy of effort. Moreover our term balance will include what are technically known as steady states and the equilibria of the physicists and of the economist.*"

[3] Similar studies were carried out by Kostić (1981), and also statistical methodology to analyse word frequency counts has been discussed in detail by Butler (1985). Our present study gained much help from these works.

the text determined, the frequencies of different words are found to follow an approximate harmonic progression. If f_r is the frequency of the 'r-th' commonest word, then $f_r = c/r$ approximately, (r = 1,2,3.....). If f_r is plotted against 'r' on a double logarithmic scale, the relationship is approximately linear with a slope of minus 1. This relation is approximately equivalent to a Pareto distribution for the variate word frequency - f.

3. Empirical evidence of vocabulary balance

Before we actually analyse our data and look for evidence of vocabulary balance, it is necessary to seek relevant empirical information about the number and frequency of occurrences of words in some actual samples of speech. For this we have Zipf's (1949) analysis of James Joyce's novel 'Ulysses' which has been indexed with exemplary methods by Hanley (1937) and Joos (1937)[4]. The appendix to the same published index contains all the quantitative information relevant to Zipf's (1949) analysis of the data. According to Joos (1937) there are 29,899 different words in the 260,430 running words; he also ranks those words in the decreasing order of their frequency of occurrence and tells us the actual frequency, 'f' with which the different ranks, 'r' occur. Zipf (1949), referring to Hanley's (1937) data also states that the 10th most frequent word (r = 10) occurs, 2,653 times (f = 2,653); or that the 100th word, (r = 100) occurs 265 times (f = 265). From this the actual frequency of occurrence 'f' of any rank, 'r', from r = 1 to r = 29,899 which is the terminal rank of the list, since 'Ulysses' contains only that number of different words.

Turning to the quantitative data of the Hanley (1937) index we can see from the arbitrarily selected ranks and frequencies in Table I, by its corresponding frequency 'f', in Column II, we obtain a product, 'C' in Column III, which is approximately the same size for all the different ranks and which as we see in Column IV represents approximately 1/10 of the 260,430 running words which constitute the total length of James Joyce's 'Ulysses'. Therefore Zipf (1949) concludes that there is a clearcut correlation between the number of the different words in 'Ulysses' and the frequency of their usage, in the sense that they approximate the simple equation of an equilateral hyperbola $r \times f = C$ in which 'r' refers to the word's rank in 'Ulysses' and 'f' to its frequency of occurrence (as we ignore for the present the size of 'C'.) In the following Table 1 we have the arbitrary ranks with frequencies in James Joyce's 'Ulysses'. (Hanley Index).

[4] Joos (1937) refers to his "statistical tabulation" in Hanley (1937).

Table 1
Arbitrary ranks with frequencies in James Joyce's "Ulysses" (Hanley Index)

I Rank (r)	II Frequency (f)	III Product of I & II ($r \times f = C$)	IV Theoretical length ($C \times 10$)
10	2653	26530	265300
20	1311	26220	262200
30	926	27780	277800
40	717	28680	286800
50	556	27800	278000
100	265	26500	265000
200	133	26600	266000
300	84	25200	252000
400	62	24800	248000
500	50	25000	250000
1000	26	26000	260000
2000	12	24000	240000
3000	8	24000	240000
4000	6	24000	240000
5000	5	25000	250000
10000	2	20000	200000
20000	1	20000	200000
29899	1	29899	298990

Incorporated from Zipf (1949:24: Table 2.1) Arbitrary Ranks with Frequencies in James Joyce's "Ulysses" (Hanley Index).

4. Evidence of vocabulary balance in Tagore's "Galpaguccha"

In "Galpaguccha" there are 39,145 different words in the 315,850 running words. The words are ranked in the decreasing order of their frequency of occurrence and their actual frequency f with which the different ranks r occurs, is also manifest. In "Galpaguccha" the 10th most frequent word (r = 10) occurs, 1,854 times (f = 1,854); or the 100th word (r = 100) occurs 355 times (f = 355). From this the actual frequency of occurrence f of any rank r, from (r = 1) to (r = 39,145), which is the ultimate rank of the list since "Galpaguccha" contains only 39,145 different words. In the following Table 2 we have the arbitrary ranks with frequencies in Tagore's "Galpaguccha". (Roy's Index).

Table 2
Arbitrary ranks with frequencies in Tagore's "Galpaguccha" (Roys Index)

I	II	III	IV
Rank (r)	Frequency (f)	Product of I & II ($r \times f = C$)	Theoretical length ($C \times 10$)
10	1854	18540	185400
20	1190	23800	238000
30	995	29850	298500
40	889	35560	355600
50	731	36550	365500
100	355	35500	355000
200	194	38800	388000
300	133	39900	399000
400	99	39600	396000
500	84	42000	420000
1000	43	43000	430000
2000	21	42000	420000
3000	14	42000	420000
4000	10	40000	400000
5000	7	35000	350000
10000	3	30000	300000
20000	1	20000	200000
30000	1	30000	300000
39145	1	39145	391450

Graphic representation of vocabulary balance

It is obvious that Table 2 contains only a few selected items out of a possible 39,145; hence the question is legitimate as to the rank frequency relationship for all those different words, nevertheless we can present in a tabular form the rank frequency relationships for all those different words and also present them quite conveniently on a graph, because we know that the equation ($r \times f = C$) will appear on a doubly logarithmic chart paper as a succession of points descending in a straight line from left to right at an angle of 45 degrees. If the ranks and frequencies of the 39,145 different words are plotted on a doubly logarithmic chart paper and if the points fall in a straight line descending from left to right at an angle of 45 degrees, we may argue that the rank-frequency distribution of the entire vocabulary of "Galpaguccha" follows the equation ($r \times f = C$) and suggests the presence of vocabulary balance to a large extent.

Following Zipf (1949) successive ranks from one through to 39,145 were plotted horizontally on the x–axis or abscissa. Then, in measuring frequency on the y–axis or ordinate, for each rank a 'dot' which corresponds to the actual frequency of occurrence of the word was given.

With reference to the graphic representation, Fig. 1 shows the rank frequency distribution of words from "Galpaguccha".

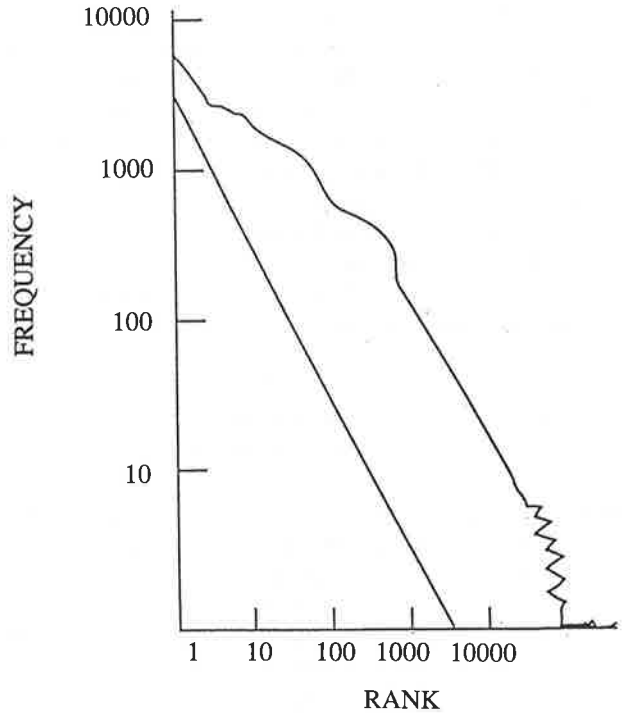


Fig. 1: Examination of the rank frequency relation (Zipf Law) for all words based on a complete count of "Galpaguccha" (Part I-IV) 39,145 distinct words and 315,850 occurrences.

After plotting the graph of the actual frequencies of the 39,145 ranked words, the dots were connected with a continuous line, to observe whether the line is straight and whether it descends from left to right at the expected angle of 45 degrees. A straight line with a slope (-1) has been placed on the figure to help in visual judgement. A closer look at Fig.1 shows how closely the curve approximates to a straight line line at an angle of 45 degrees. It may be added that the curve looks nearly straight with the slope showing some concavity. As the line approaches the bottom of the graph, the emergence of steps of progressive increase in size are seen.

In accordance with Zipf's (1949) theory since we are ranking the words from left to right in descending order of frequency, it is obvious that the line

that connects the succession of the dots does not at any point bend upwards as this would presuppose an incorrect ranking of the data according to decreasing frequencies. Simultaneously the line will proceed horizontally whenever adjacent ranks have precisely the same frequencies, and the apparently vertical lines of the 'steps' in Fig. 1 are not truly vertical as they do in fact connect adjacent dots. Referring to the quantitative data of "Galpaguccha" we can see from the arbitrarily selected ranks and frequencies in Table 2, that the relationship between *'r'* and *'f'* in the vocabulary in "Galpaguccha" is to a large extent uniform except for the 10th and 20th rank and tends to regularise after the 30th rank. The reason for this is likely to be that the percentage of common words between the 10th and the 30th rank is not very high and the different words are more evenly distributed over the other ranks, as is evident from the following Table III which contains the rank distribution in percentages of the 100 most common words in "Galpaguccha" Parts I – IV.

Table III
Rank distribution in percentages of the 100 most common words in Tagore's "Galpaguccha" Parts I-IV

Word rank in Galpaguccha	Frequency of occurrence of words	Percentage of total words
1-10	30319	9.6
11-20	14276	4.5
21-30	10847	3.4
30-40	9320	3.0
41-50	8004	2.5
51-60	6561	2.1
61-70	5303	1.7
71-80	4492	1.4
81-90	4109	1.3
91-100	3644	1.2
Total	96875	30.7

Total No. of distinct words = 39,145
Total No. of word occurrences = 315,850

It is evident from the data on "Galpaguccha" that the relationship between the various ranks *r* of these words and their respective frequencies *f* is potentially quite instructive about the entire matter of vocabulary balance, not only because it involves the frequencies with which the different words occur but also because the terminal rank of the list tells us the number of different words in the sample. It may be emphasised here that both the frequencies of occurrence and the number of different words will always be important factors in the counter-balancing of the Forces of Unification and Diversification in any sample of speech. In conclusion we may say that Tagore in "Galpaguccha" manifests a trend that whenever a person uses words to convey meanings he will automatically try to get his ideas across most efficiently by seeking a balance between the economy of a small wildly vocabulary of more general reference on the one hand and the economy of a larger one of more precise reference on the other, and in Zipf's (1949:22) words "with the result that will represent a vocabulary balance between our theoretical forces of Unification and Diversification."

Acknowledgement

We are grateful to Shri Arunendranath Roy for his assistance while going through the data.

References

Bhattacharya, N. (1965). *Some Statistical Studies on Languages.* Unpublished Ph.D. dissertation.

Butler, Ch. (1985). *Statistics in Linguistics.* Oxford: Basil Blackwell.

Hanley, M.L. (1937). *Word Index to James Joyce's "Ulysses".* Madison, Wisconsin: University of Wisconsin Pub.

Joos, M. (1936). Review of Zipf, Psycho-Biology of Language. *Language,* 12, 196-210.

Kostiæ, D. (1981). *The Lexicon of Branko Radiceviæ.* Calcutta: Indian Statistical Institute.

Mosteller, F., & Adavid, L.W. (1964). *Inference and Disputed Authorship: The Federalist.* Wesley Publishing Co. Inc.

Ross, A.S.C. (1950). Philological Probability Problems. *Journal of the Royal Statistical Society, Series-B,* 12, 19-59. (with discussion).

Roy, A. (1986). *Word Frequency Count of the Works of Bankimchandra Chattopadhyay and Rabindranath Tagore.* Unpublished monograph. Calcutta: Linguistic Research Unit, Indian Statistical Institute.

Roy, J. (1970). *Word Count of Galpaguccha.* Unpublished monograph. Calcutta: Linguistic Research Unit, Indian Statistical Institute.

Shende, P.S., & Prabhu-Ajgaonkar, S.G. (1989). On a Statistical Measure of Style. *Indian Linguistics,* 49, 1-10.

Zipf, G.K. (1945). The Repetition of Words, Time Perspective and Semantic Balance. *General Psychology,* 32, 127-48.

Zipf, G.K. (1949). *Human Behaviour and the Principle of Least Effort.* Cambridge: A.W. Press Inc.

Usus, quem penes arbitrium est, et ius, et norma loquendi.
Horace, *Ars poetica*

Game Theoretic Models of Text Construction

Johannes Gordesch, Peter Kunsmann

Introduction

'Text' is viewed here not as a physical product but as a dynamic process of expression and interpretation, whose construction can be investigated using mathematical models interpreted either as abstract structure or as psycholinguistic or sociolinguistic process. This approach overlaps considerably with discourse analysis, in particular when texts are thought of as dialogues rather than monologues.

Text construction rarely encounters situations in which a single decision maker chooses an optimal decision without reference to the effect that the decision has on other decision makers, and without reference to the effect that the decisions of others have on him or her. As a rule, however, a writer (speaker) and a reader (hearer) simultaneously choose an action, and the action chosen by each person ('player') affects the actions and reactions of the other person. For example, the use of taboo words might produce strong reactions of the recipient, and the expected or real reaction will influence the choice of words of the writer or speaker.

Game theory is useful for making decisions in cases where two or more decision makers have conflicting interests. Most of the paper deals with situations where there are only two decision makers without common interests (single persons, whole groups, abstract systems of standards etc.). The study of n-person (where n is greater than two) games involving conflict as well as cooperation, however, also makes sense.

Characteristics of Text Construction

1. There are two actors ('players') called 'sender' ('addressor', 'source', 'encoder') and 'recipient' or 'receiver' ('addressee', 'destination', 'decoder').
2. The sender must choose one of n strategies. Simultaneously, the recipient must choose one of m strategies.
3. If the sender chooses his or her i-th strategy and the recipient chooses his or her j-th strategy, the sender receives a 'reward' ('pay-off', 'gain' or 'loss', 'profit', 'utility', 'acceptance', 'reaction') of a_{ij} and the recipient loses an amount (gets a negative reward) a_{ij}. Thus we may think of the recipient's reward of a_{ij} as coming from the sender.

Such a set of actions is called a two-person zero-sum game, which is represented by the matrix in Table 1 (a game's reward or pay-off matrix). In other words, a_{ij} is the sender's reward and the recipient's loss if the sender chooses his or her i-th row strategy and the recipient chooses his or her j-th column strategy.

Table 1

	Recipient's strategy		
Sender's strategy	r_1	r_2	r_3
s_1	a_{11}	a_{12}	a_{13}
s_2	a_{21}	a_{22}	a_{23}
s_3	a_{31}	a_{32}	a_{33}
s_4	a_{41}	a_{42}	a_{43}

Example: A Game Theoretic Model of Adverbial Usage

Characteristics of Choice of Utterance

Every linguistic form, be it a specific sound, a word or a complete sentence, is dependent on the context in which it is uttered. This context is determined by the linguistic system of which the utterance is a part, by social factors and the setting in which the utterance is used. Generally, on the basis of these linguistic, social and situational factors the speech community agrees on the form of utterances. As a result, utterances are considered *correct* or *incorrect*.

Not all members of the speech community use correct forms in an actual speech event. For purposes of informality, for identification with specific subgroups of the community and for other reasons deviations from the agreed upon utterances occur. These may either be accepted or rejected by the speech community at large. The differential acceptance of such utterances allows their categorisation. In the literature these categories range from acceptable to unacceptable usage. They include utterances that are considered *divided usages*, that is usages in which correctness judgements of the members of the speech community vary considerably. Two or more divergent forms must be in use to qualify as divided usage. Example (1) shows four different ways to provide a *tag* to the statement *I am going to the store now*.

(1) I am going to the store now, *aren't I*?
I am going to the store now, *am I not*?
I am going to the store now, *ain't I*?
I am going to the store now, *amn't I*?

Utterances of divided usage can be analysed using game theoretical models of text construction. In order to collect the data, a questionnaire was developed that asked native speakers to mark items of divided usage for four degrees of acceptability. 1200 questionnaires were sent to postal addresses in six different cities on the East Coast and the Midwest of the United States. 207 questionnaires were returned.

Definition of usage types

The items that were selected for inclusion were those utterances of a set of divided usage forms which the speech community would generally find to be deviant from the accepted standard. These items were placed on the questionnaire at random to control for recognition effects. Four different usage types were offered (cf (2)).

(2) Correct (weight factor 9)
Okay (weight factor 7)
Wouldn't use it (weight factor 1)
Incorrect (weight factor 2)

As was pointed out native speakers vary in their acceptability judgements. It is necessary, therefore, to weight the response choices. The weights assigned mirror the assessed difference between the choices offered. Thus, there is only a small weight difference between the *incorrect* and *wouldn't use it* responses and a somewhat larger difference between *correct* and *okay*. Both groups, however, constitute generally positive and generally negative reactions and are separated by a fairly large difference.

The items

Of a total of 48 items on the questionnaire constituting 15 different grammatical constructions adverbials were selected for the application of the game theoretic model. There are five such items (cf. (3)). The alternate forms of each set of divided usages are provided in square brackets.

(3)

I-10	She spoke *loud* and *clear*.	[She spoke loudly and clearly]
I-12	He drove too *quick*.	[He drove too quickly]
I-26	They now drive *slower*.	[They now drive more slowly]
I-30	He wanted to *quickly* return the book when he realised that he hadn't read it himself.	[He wanted quickly to return the book...]
I-47	Don't take it too *serious*.	[Don't take it too seriously]

All items deviate from what is normally considered the correct form. I-30 shows the adverb in a position which causes a *split* infinitive. The other items do not show the normally mandatory derivational *ly*-ending. On analysing the grammatical structure of the items with respect to the form of the adverb we find five different characteristics. These characteristics can be related to an index of complexity (cf. Kunsmann, Gordesch & Dretzke, 1998). I-10 shows two conjoined adverbs. In analogy to similar constructions elsewhere in the grammar of English, speakers would recognise the conjunction as a *standing phrase*. I-12 contains a simple adverb in that it is monosyllabic. It also shows grading with *too*. This item contrasts with I-47 where the adverb is disyllabic and thus more complex than *quick* in I-12. It is also graded with *too*. I-26 contains a monosyllabic adverb. The complexity here is provided by grading with the comparative *-er*. Finally, in I-30 we see a complex syntactic structure. The adverb is shown in a deviant position of the utterance. (4) shows a summary of the characteristics of the utterances.

(4)
I-10 A standing phrase of two conjoined monosyllabic adjectives used adverbially.
I-12 A monosyllabic adjective graded with *too* and used adverbially.
I-26 A monosyllabic adjective graded with the comparative *-er* and used adverbially.
I-30 An adverb derived from a monosyllabic adjective placed in a deviant position.
I-47 A disyllabic adjective graded with *too* and used adverbially.

Absolute and relative frequencies

Based on the choices made by native speakers the following results can be computed. Table 2 shows the absolute frequencies and Table 3 the relative frequencies.

Table 2

Usage	*I-10*	*I-12*	*I-26*	*I-30*	*I-47*
Correct	91	10	84	53	30
Okay	15	7	21	24	5
Wouldn't use it	7	14	30	42	4
Incorrect	92	175	68	82	164
Missing	2	1	4	6	4
Total	207	207	207	207	207

Table 3

Usage	*I-10*	*I-12*	*I-26*	*I-30*	*I-47*
Correct	0.444	0.049	0.414	0.264	0.148
Okay	0.073	0.034	0.103	0.119	0.025
Wouldn't use it	0.034	0.068	0.148	0.209	0.020
Incorrect	0.449	0.850	0.335	0.408	0.808

Reward

When making a specific utterance in a speech situation, i.e. when choosing a specific usage type on the questionnaire, linguistic, social and situational factors provide an explanation for the selection. Depending on the relative gain for the individual speakers, they accommodate to or diverge from the speech of others (cf Giles, 1980). Their choices involve a complex set of unconscious beliefs, attitudes and perception of the speech situation.

After relative frequencies or weights are determined, credits are given to the usage types, and the products of the relative frequencies or weights and the pertaining credits are computed. Finally, the linguistic form promising the highest success is searched for.

The selection of credit types corresponds to the different linguistic positions.

- All credits are equal to one, and thus the relative frequencies or the weights constitute the 'pay-offs'.
- The credit of *correct* is equal to one, the credits of all the other evaluations are zero.
- The credits of *correct* and of *okay* are set to positive numbers, the credits of the remaining evaluations are zero.

- In the general model, the credits are assessed by experts or other authorities, which leads to a fixed assignment of quantities to *correct* etc. as in (2) above.
- In the general model, the credits are estimated from real world data involving statistical procedures.
- A fuzzy set theoretic procedure is used to assign values to the 'linguistic variables' as the labels *correct* etc. are called in fuzzy set theory.

Table 4

	Correct	*Okay*	*Wouldn't use it*	*Incorrect*
I-10	3.995	0.512	0.034	0.898
I-12	0.437	0.238	0.068	1.699
I-26	3.724	0.724	0.148	0.670
I-30	2.373	0.836	0.209	0.816
I-47	1.330	0.172	0.020	1.616

Optimisation principles

Classical game theory (*minimax principle*: Hurwicz's principle with $\omega = 0$ or Savage's principle with $\gamma = 0$) looks for a safe (maximum of column) but not too common solution (minimum of row). In our case, a minimax solution exists and is equal to the pair (I-30, *wouldn't use it*), yielding a value of 0.209. Measurement and sampling errors, however, are high, and the difference of the entries 0.209 and 0.148 should not be overestimated. Thus we could consider (I-26, *wouldn't use it*) for a possible solution.

A more natural approach is to look for the maximal element in each row, and then to choose the maximum among these elements (*Hurwicz' principle* with $\omega = 1$). This leads to the strategies (I-10, *correct*) and a reward of 3.995, or close to it, (I-26, *correct*) with a reward of 0.724.

Mixed strategies: Instead of using only one optimal linguistic form one could also use a mixture of forms with appropriate probabilities, e.g. I-10, I-26 etc. are used with probabilities equal to the relative frequencies of *correct* found in the survey. This approach seems quite appropriate, it suffers, however, from a severe drawback: Selecting a strategy is done completely blindfold, i.e. without considering the context.

The *Bayesian strategist*, using equal probabilities, would take the average of each row, and then select the maximal value among (1.360, 0.610, 1.317, 1.058, 0.784), which would make I-10 the choice (I-26 is also very close to the optimal strategy).

For the *minimum regret* model ($\gamma = 0.1$, slight regret), the pay-off matrix is transformed by subtracting one tenth of the maximal value in each row from this row. Note that - if required - by a scaling transformation pay-offs can all be made positive. Then any convenient strategy can be applied.

Table 5

	Correct	*Okay*	*Wouldn't use it*	*Incorrect*
I-10	3.596	0.113	-0.365	0.498
I-12	0.267	0.068	-0.102	1.529
I-26	3.352	0.352	-0.225	0.298
I-30	2.136	0.599	-0.028	0.579
I-47	1.197	0.039	-0.113	1.483

For instance, the Bayesian solution using equal probabilities and ergo the mean vector (0.960, 0.441, 0.944, 0.821, 0.651) is I-10 (I-26 and I-30 are also very close to the maximal value).

Discussion of results

Osgood (1966) noted that basic linguistic principles are derived from 'universals of humanness'. These principles in turn govern linguistic regularities which show up in actual utterances. It should be possible, therefore, to determine the relative strengths of these principles on the basis of a game theoretic analysis. The results of such an analysis, i.e. the application of optimisation principles to the data in (3), allow the ordering of the principles in (4).

***Conjoined utterances* (I-10)**
The choice of I-10 seems to be dependent on the fact that *loud and clear* is considered a *standing phrase*, in particular one in which coordination is of crucial importance. Coordination can be viewed as a process basic to many areas of human activity. In various languages we find coordinated structures constrained by competing hierarchies. The order of coordinated elements in English, for instance, is determined by linguistic (syllabic structure, ellipsis phenomena) and non-linguistic (generational, gender, status) hierarchies. Thus, we find in the unmarked cases of utterances *shipping and receiving*, but not *receiving and shipping* and *Mary came in and sat down*, but not *came in and Mary sat down*. Moreover, it is *father and son* and *husband and wife*, respectively, and not *son and father* or *wife and husband*. In utterances such as I-10 a basic principle of coordination seems to be at work that triggers the choice for the divided usage item.

***Comparative constructions* (I-26)**
The choice of this item depends on the mental costs involved in finding an alternative. It is simpler to use the grammatically incorrect form of *slower* than to search for a proper comparative. *Slowerly*, as one alternative, is not available, while *more slowly* cannot be accessed readily.

***Change in linguistic level* (I-30)**
The utterance involved in this principle relates categorical aspects of linguistic analysis with syntactic and phonological ones. Generally, adverbials may not be placed between the infinitival *to* and its verb. However, in the present utterance the *split infinitive* does not cause an interruption of the intonation pattern of the sentence which accounts for the relatively high values.

***Focus on form* (I-12) *and* (I-47)**
Focus on form is related to the notion of complexity of an utterance. The two items in question allow high focus on form, and their relative differences with respect to the optimisation principles is due to the relative higher complexity of utterance (I-47). Both items, however, have relatively low values, reflecting on the status of *focus on form*.

From the results of the optimisation test we can, thus, conclude that on the basis of the choices made on the questionnaire a hierarchy of competing principles can be established. Additional studies of a similar nature will have to be made for further support. Nevertheless, making choices on a principle of *coordinated items* seems to provide greater reward than choosing on a principle of *focus on form*.

Models of Text Construction

Prescriptive theory of text construction

Two-person games
Each actor chooses a strategy which enables him to do the best he can, given that the opponent knows the strategy he is following (full information). The pay-off matrix (i.e. Table 4) we have just analysed satisfies the so-called *saddle point* condition:

$$\max_{\text{all rows}} (\text{row minimum}) = \min_{\text{all columns}} (\text{column maximum})$$

Any two-person zero-sum game satisfying the above condition is said to possess a saddle point (a *pure* strategy). Thus, if a two-person zero-sum game has

a saddle point (a *pure* strategy). Thus, if a two-person zero-sum game has a saddle point, the sender should choose any strategy (row) attaining the maximum on the left side of the saddle point condition, and the recipient should choose any strategy (column) attaining the minimum on its right side. The common value of both sides is called the value v of the game. The reward for a saddle point must be the smallest number in its row and the largest number in its column. Thus, like the centre point of a horse's saddle, a saddle point for a two-person zero-sum game is a local minimum in one direction (looking across the row) and a local maximum in another direction (looking up and down the column). A saddle point can also be thought of as an equilibrium point in that neither player can benefit from a unilateral change in strategy, i.e. a saddle point is stable in that neither player has an incentive to move away from it.

A more general concept is that of a constant-sum game where the row player's reward and the column player's reward add up to a constant c (in a zero-sum game $c = 0$). In general, the optimal strategies and the value of the game may be found by the same methods used with zero-sum games.

Not all two-person zero-sum games possess saddle points. The set of feasible solutions has to be expanded so as to cover *randomised* (*mixed*) strategies: Each actor selects a probability of playing each strategy. An n-tuple $(x_1, x_2, \ldots x_n)$ is called a randomised or mixed strategy if for each i $x_i > 0$ and $x_1 + x_2 + \ldots + x_n = 1$ hold. Any mixed strategy $(x_1, x_2, \ldots x_n)$ reduces to a pure strategy if any of the x_i equals 1. Any mixed strategy where an actor gets an expected reward at least equal to the value of the game is an optimal strategy. If the sender departs from the optimal strategy the recipient may have a strategy that reduces the sender's expected reward below the value of the game, and vice versa for the recipient. A strategy i of an actor is dominated by a strategy i' if, for each of the other actor's possible strategies, strategy i' results in at least the same reward, and if for at least one of the other actor's strategies, strategy i' is superior to strategy i.

Linear programming provides an efficient tool for finding the value and the optimal strategies for any two-person zero-sum game. A denotes the pay-off matrix, A' its transpose, x and y the pertaining probability vectors, and v and w the respective value of the game.

Sender's game:

$$\max z = v,$$
$$v \le A'x,$$
$$\sum_{i=1}^{n} x_i = 1,$$
$$\forall i: x_i \ge 0, v \ge 0$$

Recipient's game:

$$\min z = w,$$
$$w \ge Ay,$$
$$\sum_{j=1}^{m} y_j = 1,$$
$$\forall_{j=1}^{m} j: y_j \ge 0, w \ge 0.$$

The two linear programs are dual to each other, and the values of the objective functions are equal.

Finally, the value and the optimal strategies for any two-person zero-sum or constant sum-game can be found using the following procedure:

Step 1: Check for a saddle point. If there is no saddle point, go on to step 2.
Step 2: Eliminate any of the sender's dominated strategies, then any of the recipient's dominated strategies. Continue until no more dominated strategies can be found, then proceed to step 3.
Step 3: Solve the game by using a linear programming method.

N-person games

In many competitive situations, there are more than two competitors. In language, grammaticality, avoidance of complexity, frequency of use, etc. may enter into competition. Let $N = \{1, 2, \ldots, n\}$ be the set of actors. Any game with n actors is an n-person game. It is specified by its characteristic function:

For each subset S of N, the characteristic function v of the game gives the amount $v(S)$ that the members of S can be sure of receiving if they form a coalition and act together. Hence $v(S)$ can be determined by calculating the amount that the members of S can obtain without any help from actors who are not in S. The function v possesses the property of superadditivity:

$$v(A \cup B) \ge v(A) + v(B)$$

Several concepts of solution of an n-person game make sense. However, they must obey the following restrictions:
Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ be a vector such that the actor i receives a reward x_i (reward vector). Apparently, the x_i must satisfy

$$v(N) = \sum_{i=1}^{n} x_i,$$

$$\forall_{i=1}^{N} i: x_i \geq v(\{i\}).$$

Then x is called an imputation. The first line of the formula given above denotes group rationality, i.e. any reasonable reward vector must give all the actors an amount equal to the amount that can be attained by the coalition consisting of all actors. The second line denotes individual rationality, i.e. it implies that actor i must receive a reward $v(\{i\})$ at least as large as the one he can get on his own.

Besides the Shapley value of a game, the core of an n-person game is a common solution concept. The core of an n-person game (Neumann-Morgenstern solution) is the set of all imputations not dominated by any other. Given an imputation $x = (x_1, x_2, \ldots, x_n)$, we define that the imputation $y = (y_1, y_2, \ldots, y_n)$ dominates x through a coalition (in symbols $y >^S x$) if

$$\sum_{i \in S} y_i \leq v(S),$$

$$\forall_{i \in S} i: y_i > x_i.$$

If $y >^S x$, then the following must be true:

1. Each member of S prefers y to x.
2. The members of S can attain the rewards given by y.

Consequently, if y dominates x, then x should not be considered a possible solution because the actors in S can always enforce their rewards y by forming a coalition.

Algorithms for determining the core of a game often start from the following theorem:

An imputation $x = (x_1, x_2, \ldots, x_n)$ lies in the core of an n-person game if and only if for each subset S of N,

$$\sum_{i \in S} x_i \geq v(S).$$

The theorem states that an imputation x is in the core (i.e. x is dominated) if and only if for every coalition S, the total of the rewards received by the actors in S is at least as large as $v(S)$.

For additional mathematical details concerning two- and n-person games refer e.g. to Owen (1982) or Thomas (1986).

Descriptive theory of text construction

In the von Neumann and Morgenstern theory of games, 'optimal strategy' is a severely restricted concept akin to that of homo oeconomicus, hypothesising intelligent and battle-tried opponents, and, in the case of mixed strategies, unlimited repetition as well as acceptance of less favourable strategies. It is not surprising that it often fails to explain or predict human behaviour. Various attempts have been made to find optimisation procedures more adequate to human reality (some interesting aspects of decision and game theoretic modelling can be found in Krabs (1997), and a comprehensive principle was given by Gordesch (1969). In most cases each opponent lists his own strategies and the respective pay-offs for each counter-strategy of his opponent. Finally the single items are combined into a global judgement, and a general strategic plan is developed. Some of these optimisation principles can be obtained by transforming the pay-off matrix taking into account the best or the worst case by replacing the pay-offs by a linear combination of the maximal and the minimal value or by subtracting a certain fraction of the maximal pay-off.

Let e_i be the reward ('utility', 'preference') for strategy i with respect to the counterstrategies j,

$$e_i = F_i(g_{i1}, g_{i2}, \ldots g_{in}), \qquad \text{or} \qquad e = F(G).$$

Then the global strategy is given by the maximisation of a preference function Z of the preference vector e and the parameter vector s,

$$Z = \Phi(e_1, e_2, \ldots e_n; s_1, s_2, \ldots s_r) \quad \text{or} \quad Z = \Phi(e, s).$$

For α and $\beta > 0$, Z and $Z' = \alpha Z + \beta$ are maximised for the same e and s (invariance under positive linear transformations, i.e. independence from scaling). Utility is defined as a preference relation, i.e. a weak ordering (a transitive and identitive relation) that is invariant under positive linear transformations:

$P(u,u') \wedge P(u',u'') \Rightarrow P(u,u'')$,
$P(u,u)$ is valid,
u and v are equivalent if and only if
$\alpha, \beta > 0 \Rightarrow u = \alpha v + \beta$.

(For a more general definition of utility cf Gordesch,1969). Linear functions Z allow the derivation of the common optimization principles of Hurwicz and of Savage, the minimax principle, and Bayesian strategies.

The principle of Hurwicz

Let g_i be the minimal, G_i the maximal elements of the i-th row of $[g_{ik}]$:

$$g_i = \min (g_{i1}, g_{i2}, \dots, g_{ir}),$$
$$G_i = \max (g_{i1}, g_{i2}, \dots, g_{ir}), \quad i = 1, 2, \dots, n,$$

and

$$F_i = (1 - \omega)g_i + \omega G_i ,$$

where

$0 \le \alpha \le \omega \le \alpha' \le 1$, ω some measure of optimism, α α' limits for ω.

Then the optimal strategy is found by

$$Z = \max (e_1, e_2, \dots, e_n) .$$

Using the expected value of e instead of selecting its largest component is also possible (cf the following).

The principle of Savage (minimum regret)

The e_i are the maximal elements of the i-th row of the pay-off matrix:

$$e_i = \max (g_{i1}, g_{i2}, \dots, g_{ir}), \quad i = 1, 2, \dots, n.$$

The parameters s_i are interpreted as the sender's probabilities of choosing the i-th strategy. The preference function is given by

$$Z = [P_{Ai}]' [R_{ik}] [P_{Ak}].$$

$[R_{ik}]$ is the so-called regret matrix

$$[R_{ik}] = [g_{ik}] - \gamma \begin{bmatrix} e_1 & \dots\dots & e_i, \\ \dots & \dots\dots & \dots, \\ e_n & \dots\dots & e_n \end{bmatrix}, \quad 0 \le \gamma \le 1.$$

The minimax principle

The minimax principle, mostly used in the normative theory of games, derives from Hurwicz' principle for $\omega = 0$, or from Savage's principle for $\gamma = 0$.

Bayesian strategies

P_{Bk} are the known probabilities of the recipient's strategies. Then

$$[e_i] = [g_{ik}] [P_{Bk}] ,$$

and

$$Z = \max e_i .$$

For equal probabilities the equations simplify (Laplace strategies):

$$[e_i'] = [g_{ik}] \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ 1 \end{bmatrix}, \quad Z = \max(e_1, e_2, \dots, e_n).$$

(For some statistical aspects cf Gordesch, 1972).

References

Giles, H. (1980). Accommodation Theory: Some New Directions. In S. de Silva (Ed.), *Aspects of Linguistic Behaviour.* York, England: York University Press.

Gordesch, J. (1969). An Optimization Principle in the Descriptive Theory of Games. *Metroeconomica,* 21, 166-179.

Gordesch, J. (1972). On the Equality of Pay-off Matrices. *Metrica,* 19, 140-149.

Krabs, W. (1997). *Mathematische Modellierung. Eine Einführung in die Problematik.* Stuttgart: B.G. Teubner.

Kunsmann, P., Gordesch, J., & Dretzke, B. (1998). Native Speakers Reactions to Modern English Usage. *Journal of Quantitative Linguistics,* 5,3, 214-223.

Osgood, Ch. (1966). Universals and Psycholinguistics. In J. Greenberg (Ed.), *Universals of Language* (pp. 299-322), Cambridge, MA: MIT Press.

Owen, G. (1982). *Game Theory.* Orlando, Fla.: Academic Press.

Thomas, L.C. (1986). *Games, Theory and Applications.* Chichester, England: Ellis Harwood,

Zur Satz- und Teilsatzlänge zweigliedriger formelhafter Sprichwörter

Peter Grzybek

0. Einleitung

Ungeachtet der Vielzahl von Untersuchungen, die es mittlerweile zum Sprichwort gibt – man vergleiche allein die umfassenden internationalen Bibliographien von Mieder (1982, 1990, 1993) – fehlt es nach wie vor an umfassenden systematischen Untersuchungen zur sprachlichen Struktur dieses Genres. Die Ursache dieses Forschungsdefizits dürfte in erster Linie darin zu suchen sein, daß die Untersuchung von Sprichwörtern für lange Zeit eine Domäne der Folkloristik bzw. Volkskunde gewesen ist, während sie von der Sprach- und Literaturwissenschaft, von einigen richtungsweisenden Ausnahmen abgesehen, stark vernachlässigt wurde. So führen Röhrich und Mieder noch 1977 in ihrer synoptischen Einführung zum Thema *Sprichwort* bezeichnenderweise zwar die ‚*Sprachgeschichte*' als einen der zentralen Wissenschaftszweige an, die sich der Erforschung des Sprichworts verschrieben haben, nicht aber die Linguistik allgemein. Die damals noch im selben Jahr von Peukes (1977) vorgelegte Arbeit zur Semantik und Syntax deutscher Sprichwörter entsprach insofern durchaus einem gewissen Desiderat, doch zeichnete auch sie sich durch eine Eigenschaft aus, die man mit einer gewissen Portion an Böswilligkeit durchaus als ‚symptomatische Linguistik des Sprichworts' bezeichnen könnte: Es werden an einzelnen ausgewählten Beispielen syntaktische Charakteristika von Sprichwörtern beispielhaft demonstriert (!), ohne daß auch nur die Frage aufgeworfen würde, inwiefern die aufgezeigten sprachlichen Strukturen für das Sprichwort typische Eigenschaften sind oder nicht, welche Frequenz sie aufweisen, welchen Anteil sie innerhalb eines größeren Sprichwortkorpus haben, usw. Aus dieser Sicht heraus leisten also solche Arbeiten wie die genannte von Peukes (1977) oder auch neuere Arbeiten wie die *Linguistische Analyse eines Sprichworttyps* von Lenz (1993) zwar überaus akribische linguistische Beschreibungen – das grundlegend bestehende Defizit an einer systematischen Linguistik des Sprichworts können sie allerdings nicht beheben. Denn solange ausschließlich symptomatische, nicht aber systematische Untersuchungen vorliegen, wird man sich mit solchen mehr oder weniger pauschalen Verweisen auf die

„Knappheit" und „schlagfertige Kürze und geschliffene Prägnanz" (Röhrich & Mieder, 1977:56) des Sprichworts als dessen wichtigstes Stilmerkmal begnügen müssen – Aussagen, die im Grunde genommen nicht über die Einsichten hinausgehen, die spätestens seit Friedrich Seilers (1922) *Deutsche Sprichwörterkunde* als Allgemeingut angesehen werden können.

Seilers Sprichwörterkunde ist für den deutschsprachigen Bereich eine der wenigen Ausnahmen, die sich vergleichsweise früh um eine sprachliche Analyse des Sprichworts bemüht haben. Schon Seiler (1922:180) hob nämlich die Kürze als „das oberste Stilgesetz des Sprichworts" hervor; er verwies in diesem Zusammenhang allerdings nicht nur allgemein auf die Tendenz zur ‚Breviloquenz', zur Kurzrede also, sondern äußerte darüber hinaus die Vermutung, daß hierdurch zugleich Parallelstrukturen auf verschiedenen sprachlichen Ebenen – angefangen von der Verwendung von Antonymen auf der Ebene der Lexik bis hin zur Ausbildung von Parallelstrukturen auf der Ebene der Syntax – gefördert würden (ibd., 182). Die Tendenz zur Ausbildung von Parallelismen diskutierte Seiler (1922:186ff.) ebenso wie 50 Jahre nach ihm Röhrich und Mieder (1977:56ff.) in unmittelbarem Zusammenhang mit der Funktion formelhafter Wendungen und typischer Satzverbindungen, worunter bei ihnen u.a. der formelhafte Gebrauch von Relativsätzen verschiedener Art (*Wer ..., der ...; Wer ..., den ...; Wer ..., ...*), die Verbindung von Relativ- und Demonstrativsätzen durch die Verbindung *Je ..., je ...*, seltener auch *Je ..., desto ...*, der formelhafte Gebrauch der Negation (*Ohne ... kein ...*), Komparative wie *Besser/Lieber ... als/denn ...* u.a.m. verstanden werden. Alle diese Formeln verleihen dem Sprichwort einen vermeintlich „typischen Charakter", auch wenn zugestandenermaßen einige dieser Formeln „verhältnismäßig selten, andere ungemein häufig" sein sollen (Seiler, 1922:186).

Ungeachtet der Tatsache, daß in den parömiologischen Studien immer wieder auf eine Tendenz zur Kürze verwiesen wird – ohne daß dabei freilich dargelegt würde, worin diese Kürze eigentlich besteht –, werden offenbar auch Parallelismus und Formelhaftigkeit zwar immer wieder in identischem Kontext gesehen und diskutiert; ob aber eine wechselseitige Abhängigkeit zwischen diesen beiden Faktoren besteht, wird nicht gesagt (und eigentlich auch erst gar nicht explizit gefragt). Diese Frage kann auch solange nicht angemessen behandelt werden, wie sich die Tendenz der symptomatischen Linguistik fortsetzt, in der das sprichwörtliche Material zum Zwecke der Demonstration verwendet wird, nicht aber als eigentliches Material quantifizierender Untersuchungen. Dringend notwendig erscheinen vor diesem Hintergrund systematisch-quantitative Untersuchungen, die sich sowohl mit einzelnen der genannten Faktoren wie auch mit möglichen Wechselbeziehungen zwischen ihnen beschäftigen.

Die vorliegende Arbeit kann natürlich nicht all diese im Laufe der Geschichte der Sprichwortforschung vernachlässigten Fragen mit einem Male behandeln, geschweige denn beantworten. Ungeachtet dessen soll ein Versuch

unternommen werden, zumindest exemplarisch eine auf die o.a. Faktoren bezogene Analyse vorzustellen, die die Fruchtbarkeit eines entsprechenden Herangehens aufzeigen und womöglich den Weg für weiterführende Untersuchungen ebnen könnte.

1. Fragestellung und Erstellung des Untersuchungskorpus

Ausgangspunkt einer quantifizierenden Untersuchung kann nur ein umfassendes Sprichwortkorpus sein. Wir haben zu diesem Zweck die Sammlung *Deutsche Sprichwörter* von Karl Simrock aus dem Jahre 1846 ausgewählt; in ihr sind insgesamt 12980 Sprichwörter enthalten; da eine Reihe von ihnen aus mehr als einem Satz bestehen, beläuft sich die Gesamtsumme der Sätze auf 13017 Einheiten.[1] Die Auswahl eines solchen Sprichwörterkorpus – welches sich mit dem finnischen Parömiologen Matti Kuusi nicht zu unrecht als „Massengrab von Sprichwortleichen" verstehen läßt – birgt natürlich insofern Gefahren in sich, als es mehr als wahrscheinlich ist, daß hier zahlreiche unbekannte, von dem entsprechenden Herausgeber sprachlich redigierte Sprichwörter Aufnahme gefunden haben, die mit dem tatsächlich in Verwendung befindlichen Sprichwortkorpus womöglich nur entfernte Ähnlichkeit aufweisen. Solange es jedoch kein solches empirisch abgesichertes Korpus von Sprichwörtern gibt, werden wir nicht umhin kommen, unsere Analysen auf der Basis vorhandener Sprichwortsammlungen durchzuführen; dies läßt sich entweder mit der Hoffnung, daß wir es mit solchen sprachlichen Strukturen zu tun haben, die den tatsächlich im Umlauf befindlichen einigermaßen entsprechen, verbinden, oder aber mit der Frage, inwiefern als typisch erachtete Sprichwörter in besonderem Maße Eingang in die Sammlung gefunden haben.

Die genannte Sammlung von Simrock haben wir u.a. auch deshalb ausgewählt, weil sie bereits in einer anderen Arbeit Gegenstand einer quantifizierenden Untersuchung war, in der es um die in Worten gemessene durchschnittliche Satzlänge der Sprichwörter ging (Grzybek, 1995). In der vorliegenden Studie soll es in erster Linie um andere Fragen als diejenige der Satzlänge gehen; ungeachtet dessen seien die Ergebnisse hier in aller gebotenen Kürze referiert, da wir unten auf sie Bezug nehmen werden.

Die durchschnittliche (in Wörtern gerechnete) Satzlänge aller 13017 Sätze der Sammlung von Simrock beträgt $\bar{x} = 7.83$ bei einer Standardabweichung

[1] Die gleichwertige Behandlung von Sprichwörtern, die aus mehr als einem Satz bestehen, mag natürlich in der einen oder anderen Hinsicht problematisch erscheinen. Erstens aber handelt es sich bei den Sprichwörtern, die aus mehr als einem Satz bestehen, um eine verschwindend große Menge von weniger als 0.3% des gesamten Materials; zweitens wurde - und das ist wichtiger - bei den Berechnungen die Gesamtmenge der Sätze, nicht die der Sprichwörter zugrundegelegt, so daß sich durchaus auch unter Einbezug dieser Sprichwörter Einsichten in die durchschnittliche Satzlänge von Sprichwörtern gewinnen lassen.

von $s = 3.34$. Bei der ausführlichen Diskussion divergierender Ergebnisse aus anderen Untersuchungen zur Satzlänge von Sprichwörtern haben wir in der genannten Arbeit die Hypothese aufgestellt, daß die durchschnittliche Satzlänge mit der durchschnittlichen Bekanntheit der jeweiligen Sprichwörter korreliert (ohne daß freilich eine Vermutung über die Richtung des Zusammenhangs aufgestellt werden konnte).

In der vorliegenden Arbeit soll ein anderer Versuch der Quantifizierung gemacht werden, der auf die Untersuchung des oben angesprochenen Zusammenhangs von Formelhaftigkeit, Parallelismus und Satzlänge abzielt. Zu diesem Zweck wurden aus der Gesamtmenge der 12980 Simrockschen Sprichwörter in einem ersten Schritt all diejenigen Einheiten herausgefiltert, die eine vergleichbare syntaktische Struktur aufweisen und neben dem Satzendezeichen lediglich ein Komma als Satzzeichen beinhalten. Das Ergebnis dieses ersten vorbereitenden Schrittes war die Reduzierung des Gesamtkorpus auf eine Menge von 5504 Sprichwörtern (42.40% des Ausgangsmaterials). Diese 5504 Sprichwörter wurden in einem zweiten Schritt weiter gefiltert, bei dem ausschließlich diejenigen Sprichwörter in das Untersuchungskorpus übernommen wurden, die durch „klassische" Formelanfänge charakterisiert sind: *Je ...; Was ...; Wem ...; Wen ...; Wenn ...; Wer ...; Wie ...; Wo ...; Wohin ...* Auf diese Weise ergab sich ein verbleibendes Korpus von 2114 Sprichwörtern (16.29% des ursprünglichen Ausgangsmaterials), die sich einerseits durch eine syntaktische Zweiteilung, andererseits durch eine formelhafte Einleitung auszeichnen. Nicht berücksichtigt wurde bei der Zusammensetzung des Untersuchungskorpus also zunächst, ob auch der jeweils zweite Teil eine explizite formelhafte Einleitung (*Je ..., je / desto / um so ...*; u.a.) aufweist oder nicht.

In einer ersten Kategorisierung des so erhaltenen Untersuchungskorpus wurde sodann untersucht, aus welchen syntaktischen Strukturen dieses sich zusammensetzt, wenn man nach den jeweiligen Einleitungsformeln differenziert. Die folgende Tabelle 1 veranschaulicht die Ergebnisse dieser Kategorisierung, und zwar sowohl für jedes einzelne Syntagma als auch kumulativ für die jeweiligen Einleitungsformeln.

Tabelle 1
Verzeichnis der im Korpus vorkommenden Strukturformeln

Kolon (I)	*Kolon* (II)	*f*	%	% (*kum*)
Je	Desto	2	0,09	4,26
	Je	87	4,12	
	- - -	1	0,05	
Was	Da	4	0,19	14,62
	Das	90	4,26	
	Des	1	0,05	
	Dessen	1	0,05	
	Die	2	0,09	
	So	2	0,09	
	Wenn	16	0,76	
	- - -	193	9,13	
Wem	Dem	6	0,28	1,23
	Den	3	0,14	
	Der	15	0,71	
	- - -	2	0,09	
Wen	Dem	3	0,14	1,32
	Den	8	0,38	
	der	13	0,61	
	so	1	0,05	
	- - -	3	0,14	
Wenn	da	1	0,05	20,77
	dann	3	0,14	
	der	1	0,05	
	so	188	8,89	
	wer	1	0,05	
	- - -	245	11,59	
Wer	dem	96	4,54	44,56
	den	43	2,03	
	der	172	8,14	
	des	2	0,09	
	so	3	0,14	
	was	2	0,09	
	wem	1	0,05	
	wenn	1	0,05	
	wer	1	0,05	
	- - -	621	29,38	
Wes	das	2	0,09	0,19
	des	2	0,09	
Wie	also	4	0,19	4,30
	der	1	0,05	
	die	1	0,05	
	so	73	3,45	
	wie	2	0,09	
	- - -	10	0,47	
Wo	da	119	5,63	8,66
	dahin	1	0,05	
	so	2	0,09	
	wer	1	0,05	
	- - -	60	2,84	
Wohin	da	1	0,05	0,09
	dahin	1	0,05	

Wie der Tabelle 1 zu entnehmen ist, macht die Anzahl derjenigen Sprichwörter, die zwar mit einer Formel eingeleitet, nicht aber im zweiten Teil explizit formelhaft fortgesetzt werden, etwas mehr als die Hälfte aus: Von den 2114 Sprichwörtern des Untersuchungskorpus werden 979 (46.31%) im zweiten Teil explizit formelhaft weitergeführt, 1135 Sprichwörter (53.69%) hingegen werden ohne Formel weitergeführt. Im folgenden werden wir zwischen diesen beiden Teilmengen unterscheiden: Diejenigen Sprichwörter, die nur im ersten Kolon eine einleitende Formel aufweisen, werden wir als *F*1-Sprichwörter bezeichnen, und diejenigen, die nicht nur im ersten Kolon eine einleitende, sondern auch im zweiten Kolon eine weiterführende Formel aufweisen, werden wir ‚explizit formelhafte Sprichwörter' nennen und mit dem Kürzel *F*2-Sprichwörter bezeichnen. Es ist aus der Tabelle 1 leicht erkenntlich, daß sich die *F*2-Sprichwörter auf eine überschaubare Menge besonders produktiver Konstruktionen konzentrieren: Wenn wir unter ‚produktiv' bedingt diejenigen Formeltypen verstehen, die innerhalb der explizit formelhaften (*F*2)-Sprichwörter auf einen Anteil von mindestens 5% kommen, so handelt es sich um die folgenden Typen: *Je ..., je ...; Was ..., das ...; Wenn ..., so ...; Wer ..., dem / der ...; Wie ..., so ...; Wo ..., da ...* – Die Sprichwörter mit diesen produktiven Konstruktionen – die wir im folgenden als *P*-Sprichwörter bezeichnen (und die nichts anderes als eine Teilmenge der *F*2-Sprichwörter sind) – machen eine Menge von 825 Texten aus, was einem Anteil von 84.27% der explizit formelhaften (*F*2)-Sprichwörter bzw. 39.03% der überhaupt mit einer Formel eingeleiteten (*F*1- und *F*2)-Sprichwörter entspricht; im Hinblick auf die 12980 Sprichwörter des Ausgangsmaterials handelt es sich hier-

bei um einen Anteil von 6.36%. Fig. 1 veranschaulicht die Verteilung der *F*1- und *F*2- bzw. *P*-Sprichwörter.

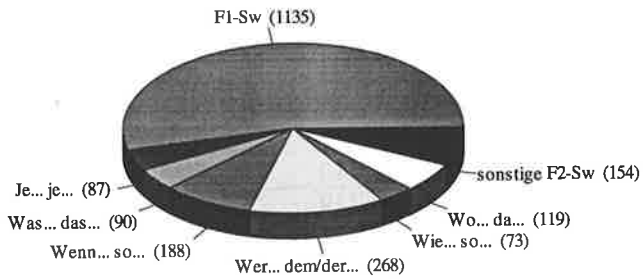


Fig. 1

2. Analysen

2.1. Satzlänge

Im folgenden wollen wir versuchen, eine Reihe quantifizierender Aussagen über die linguistische Struktur der Sprichwörter des Untersuchungskorpus zu erarbeiten. Wenden wir uns als erstes der Analyse der (in Wörtern gemessenen) durchschnittlichen Satzlänge zu.

Die Satzlänge der 2114 Sprichwörter des Untersuchungskorpus liegt mit einer durchschnittlichen Satzlänge von $\bar{x} = 8.89$ bei einer Standardabweichung von $s = 2.51$ signifikant ($p < 0.001$) über der durchschnittlichen Satzlänge des gesamten Simrock-Korpus (s.o.), ebenso wie auch die durchschnittliche Satzlänge der 1135 für sich genommenen *F*1-Sprichwörter ($\bar{x} = 8.59$, $s = 2.17$) bzw. der 979 explizit formelhaften *F*2-Sprichwörter ($\bar{x} = 9.24$, $s = 2.81$). Die Tatsache der größeren Durchschnittslänge ist insofern nicht überraschend, als ja alle 6306 syntaktisch eingliedrigen Sprichwörter (immerhin 48.55% der Simrockschen Sammlung), die sich im Vergleich zum gesamten Korpus als deutlich kürzer erweisen ($\bar{x} = 6.29$, $s = 2.81$) keinen Eingang in das Untersuchungskorpus gefunden haben. Tabelle 2 faßt die Ergebnisse in anschaulicher Form zusammen.

Tabelle 2
Durchschnittliche Satzlänge der formelhaften Sprichwörter im Korpus

	F-Sw	*F1-Sw*	*F2-Sw*
$\bar{x}$	8.89	8.59	9.24
s	2.51	2.17	2.81
n	2114	1135	979

Aufgrund dieser Befunde kommen wir nicht um die Feststellung hin, daß sich vermutlich kein Zusammenhang zwischen der durchschnittlichen Satzlänge der Sprichwörter und dem Faktor der Formelhaftigkeit herstellen läßt: Da die formelhaften Elemente selbst Bestandteil der syntaktischen und lexikalischen Struktur sind, erweisen sie sich logischerweise von ihrer Tendenz her länger als die eingliedrigen Sprichwörter.

2.2. Teilsatzlänge

Ergiebiger als die Berechnung der durchschnittlichen Satzlänge ist vermutlich jedoch im Hinblick auf die zweigliedrigen Sprichwörter die Frage, ob sich quantifizierende Aussagen über das Längenverhältnis zwischen dem ersten und zweiten Kolon der Sprichwörter des Untersuchungskorpus machen lassen: Läßt sich eine Tendenz nachweisen im Hinblick darauf, ob eines der beiden Kola länger ist als das andere?

Untersuchen wir diese Frage zunächst im Hinblick auf die 2114 Sprichwörter des gesamten Untersuchungskorpus. Für unsere Zwecke läßt sich recht gut der McNemar-Test der Signifikanz von Veränderungen verwenden, insofern es uns weder um kontinuierliche Quantitäten noch um die absoluten Unterschiede zwischen erstem und zweitem Kolon geht. Dieses Verfahren hat Altmann (1965) im Hinblick auf eine ähnliche Fragestellung bei der Untersuchung malayischer Pantuns angewendet: Es bietet sich nicht zuletzt deswegen an, weil unserer explorativen Untersuchung die Analyse auf die Wortebene und nicht etwa auf die (gegebenenfalls sinnvollere) Silbenebene ausrichten, und insofern bestenfalls eine Hypothese über eventuell vorliegende Unterschiede erarbeiten wollen. Wir testen insofern lediglich die Richtung des Unterschieds, d.h. wir halten lediglich fest, ob das erste Kolon (K-I) mehr oder weniger Wörter als das zweite Kolon (K-II) aufweist und vernachlässigen das tatsächliche Ausmaß vorhandener Längenunterschiede: Der Test basiert unter Vernachlässigung der in bezug auf die Länge identischen Einheiten lediglich auf den Differenzen und untersucht, ob die Gesamtmenge der Differenzen eine bestimmte Richtung aufweist oder nicht. Eine Antwort auf die Frage von Längenunterschieden bei formelhaften Sprichwörtern wäre u.a. auch deshalb von Interesse, weil sich hier eine Aussage darüber treffen

ließ, ob diese Texte eher eine Tendenz zur Klimax oder aber zur Anti-Klimax aufweisen – eine Frage, die in späteren Untersuchungen dann unbedingt auf der Silbenebene weiterzuverfolgen wäre.

Wenden wir uns den Ergebnissen zu: Die 627 Fälle, in denen das zweite Kolon länger ist als das erste, wurden entsprechend des McNemar-Tests mit dem Buchstaben ‚A' bezeichnet, die 791 Fälle, in denen das zweite Kolon kürzer ist als das erste, mit ‚B', die übrigen 696 Fälle mit ‚0'. Dies ergibt einen $?^2$-Wert von 18.74[2], welcher bei $FG = 1$ einer Wahrscheinlichkeit von $p < 0.001$ entspricht. Damit können wir davon ausgehen, daß in den zweigliedrigen formelhaft eingeleiteten Sprichwörtern insgesamt eine hochsignifikant ausgeprägte Tendenz zur Anti-Klimax besteht: demnach ist das *erste Kolon länger als das zweite Kolon*.

Interessant ist nun des weiteren die Frage, ob sich die explizit formelhaften Sprichwörter (mit formelhaften Einleitungen im ersten und im zweiten Kolon) in dieser Hinsicht anders verhalten oder nicht. Aus diesem Grunde wollen wir in einem nächsten Schritt diejenigen 1135 Sprichwörter, die im zweiten Teil keine explizite Einleitungsformel aufweisen, und die explizit formelhaften getrennt analysieren.

Beginnen wir mit den 1135 *F*1-Sprichwörtern (ohne Einleitungsformel im zweiten Kolon). Von diesen gibt es bei 343 Sprichwörtern keinen Längenunterschied zwischen erstem und zweitem Kolon, 509 Sprichwörter weisen ein längeres erstes, 283 ein längeres zweites Kolon auf. Der entsprechende $?^2$-Wert liegt bei 63.92, was einer Wahrscheinlichkeit von $p < 0.001$ entspricht. Es stellt sich somit heraus, daß bei den *F*1-Sprichwörtern die Tendenz des längeren ersten Teils sehr viel deutlicher ausgeprägt ist als wir dies im vorigen Schritt im Hinblick auf das Untersuchungskorpus in toto feststellen konnten. Dies macht natürlich die Frage, wie sich die Längenverhältnisse bei den explizit formelhaften *F*2-Sprichwörtern in dieser Hinsicht ausnehmen, um so spannender.

Von den 979 *F*2-Sprichwörtern weisen 353 keinen Längenunterschied zwischen erstem und zweitem Kolon auf, bei 344 Sprichwörtern ist das erste Kolon (K-I) länger, bei 282 Sprichwörtern das zweite (K-II). Dies ergibt nach dem McNemar-Test einen $?^2$-Wert von 5.94, der zwar deutlich niedriger ist als bei den *F*1-Sprichwörtern, der aber ungeachtet dessen nach wie vor ein Signifikanzniveau erreicht ($p < 0.015$). Damit können wir als zweites wesentliches Ergebnis festhalten, daß bei den zweigliedrigen formelhaften Sprichwörtern zwar insgesamt eine vergleichsweise einheitliche Tendenz zur Anti-Klimax besteht, daß diese Tendenz allerdings bei den *F*2-Sprichwörtern deutlich weniger ausgedrückt ist.

[2] Die χ^2-Werte des McNemar-Tests werden hier und im folgenden nach Kontinuitätskorrektur angegeben.

Versuchen wir im folgenden, den Grundlagen dieser Tendenz näher auf die Spur zu kommen.

Zu diesem Zweck nehmen wir nochmals Bezug auf die oben angeführten besonders produktiven Formeltypen bzw. die von ihnen geprägten 825 *P*-Sprichwörter; von ihnen läßt sich vermuten, daß sie aufgrund ihrer Produktivität einer bestimmten Prototypikalität entsprechen. Ohne Frage gibt es, was die Frage der Prototypikalität betrifft, auf den verschiedenen Sprachebenen unterschiedliche Formen von Prototypen, von denen man annehmen kann, daß sie bestimmten Bedingungen optimal entsprechen: Im Hinblick auf die uns im gegebenen Zusammenhang vorrangig interessierende Frage des Zusammenhangs von Formelhaftigkeit und Parallelstruktur scheint es deshalb durchaus sinnvoll, die Gruppe der *P*-Sprichwörter im Hinblick auf Längenunterschiede zwischen erstem und zweiten Kolon eigens zu betrachten.

In der Tat weichen die 825 *P*-Sprichwörter eindeutig von den übrigen Sprichwörtern des Untersuchungskorpus ab: Bei 303 *P*-Sprichwörtern gibt es keinen Längenunterschied zwischen K-I und K-II, bei 282 ist das zweite Kolon länger als das erste, bei 240 ist das erste Kolon länger als daß zweite. Der X^2-Wert ist mit 3.22 deutlich niedriger als bei der Gesamtmenge der *F*2-Sprichwörter und erreicht kein Signifikanzniveau ($p = 0.073$). Damit können wir festhalten, daß sich die produktiven (prototypischen) unter den explizit formelhaften *P*-Sprichwörter durch eine Tendenz auszeichnen, die sich als klare Tendenz zur Parallelisierung auf der lexikalisch-syntaktischen Ebene verstehen läßt.

Die aufgezeigten Tendenzen bestätigen sich im übrigen weitgehend auch bei der konkreten Berechnung der Teilsatzlängen (vgl. Tabelle 3) auf der Grundlage der durchschnittlichen Wortanzahl pro Kolon. Bei den 1135 *F*1-Sprichwörtern (ohne explizite Formeleinleitung im zweiten Kolon) verteilt sich die oben berechnete durchschnittliche Satzlänge von insgesamt $\bar{x} = 8.59$ ($s = 2.17$) wie folgt auf die beiden Kola: K-I weist eine durchschnittliche Länge von $\bar{x} = 4.46$ ($s = 1.30$), K-II von $\bar{x} = 4.13$ ($s = 1.39$) auf; der Unterschied zwischen K-I und K-II der *t*-Test für abhängige (gepaarte) Stichproben weist diesen Unterschied als signifikant aus ($p < 0.001$).

Bei den 979 *F*2-Sprichwörtern weist das erste Kolon im Vergleich dazu eine durchschnittliche Länge von $\bar{x} = 4.58$ ($s = 1.69$), das zweite eine durchschnittliche Länge von $\bar{x} = 4.66$ ($s = 1.69$) auf; der *t*-Test weist diesen Unterschied allerdings als nicht signifikant auf ($p = 0.206$); dasselbe gilt auch für die 825 *P*-Sprichwörter, bei denen die Werte für K-I bei $\bar{x} = 4.58$ ($s = 1.66$), für K-II bei $\bar{x} = 4.61$ ($s = 1.70$) liegen; auch dieser Unterschied erweist sich als nicht signifikant ($p = 0.630$). Die folgende Tabelle 3 resümiert die Ergebnisse.

Tabelle 3
Durchschnittliche Teilsatzlänge

		K-I	*K*-II
*F*1-*Sw*	$\bar{x}$	4.46	4.13
(*n* = 1135)	*s*	1.30	1.39
*F*2-*Sw*	$\bar{x}$	4.58	4.66
(*n* = 979)	*s*	1.69	1.69
P-*Sw*	$\bar{x}$	4.58	4.61
(*n* = 825)	*s*	1.66	1.70

Im Vergleich zeigt sich deutlich, daß sich nicht nur die relativen Längenverhältnisse ändern, sondern daß es in der Tat das zweite Kolon bei den explizit formelhaften *F*2-Sprichwörtern ist, welches sich als länger erweist. Die damit verbundene Feststellung, daß die Länge des zweiten Kolons nicht nur relativ im Verhältnis zur Länge des ersten Kolons, sondern absolut zunimmt, scheint auf den ersten Blick trivial zu sein, insofern im zweiten Kolon bei diesen Sprichwörtern eben eine explizite Einleitungsformel „hinzukommt"; doch ergibt sich damit insgesamt eine eindeutig gesicherte Tendenz zur Parallelisierung der lexikalisch-syntaktischen Struktur, wohingegen die *F*1-Sprichwörter (ohne Einleitungsformel im zweiten Kolon) eine deutliche Tendenz zur Anti-Klimax aufweisen; diese Tendenz ließe sich ihrerseits vermutlich am besten als Prägnanz- oder Pointierungstendenz verstehen.

3. Zusammenfassung

Resümieren wir abschließend nochmals die wichtigsten Ergebnisse unserer Untersuchung. Bei den zweigliedrigen formelhaften Sprichwörtern, die das Untersuchungsmaterial der vorliegenden Studie darstellen, läßt sich insgesamt eine Tendenz nachweisen, derzufolge das erste der beiden Kola sich im Vergleich zum zweiten Kolon als länger erweist. Diese Tendenz, die sich als Tendenz zur Anti-Klimax verstehen läßt, ist allerdings nicht von allgemeiner Gültigkeit. Während sie nämlich Sprichwörtern mit einer formelhaften Einleitung nur im ersten Kolon in besonderem Maße eignet, besteht bei den Sprichwörtern mit formelhafter Einleitung im ersten und zweiten Kolon eine andere Tendenz (die allerdings nur bei den Sprichwörtern mit besonders produktiven „prototypischen" Strukturen signifikant zum Ausdruck kommt): Hier stellt sich heraus, daß erstes und zweites Kolon dazu tendieren, gleich lang zu sein, so daß Formelhaftigkeit in der Tat in einer syntaktisch-lexikalischen Parallelstruktur resultiert.

Aus diesen Ergebnissen leitet sich die Notwendigkeit für weitere Untersuchungen ab: Zum einen gilt es, die entsprechende Analyse auf der Silbenebene durchzuführen, um festzustellen, ob sich die beobachtete Tendenz zur Parallelisierung auch hier bestätigt; zum anderen müssen die Analysen unter Berücksichtigung des Bekanntheitsgrades der Sprichwörter durchgeführt werden, damit festgestellt werden kann, ob sich bei den bekannteren Sprichwörtern dieselben Tendenzen herausstellen lassen oder ob es in Abhängigkeit vom Bekanntheitsgrad zu abweichenden Ergebnissen kommt.

Literatur

Altmann, G. (1966). The climax in Malay pantun. *Asian and African Studies*, 1, 13-20.

Grzybek, P. (1995). Zur Frage der Satzlänge von Sprichwörtern (unter besonderer Berücksichtigung deutscher Sprichwörter. In Baur, S. Rupprecht & Ch. Chlosta (Hrsg.), *Von der Einwortmetapher zur Satzmetapher. Akten des Westfälischen Arbeitskreises „Phraseologie / Parömiologie* (1994/95)". Bochum: Brockmeyer. [= Studien zur Phraseologie und Parömiologie; 6]

Lenz, B. (1993). Hundert Sprichwörter, hundert Wahrheiten. Linguistische Analyse eines Sprichworttyps. *Wuppertaler Arbeitspapiere zur Sprachwissenschaft*, 8.

Mieder, W. (1982). *International Proverb Scholarship. An Annotated Bibliography.* New York: Garland.

Mieder, W. (1990). *International Proverb Scholarship. An Annotated Bibliography. Supplement I (1800-1981).* New York: Garland.

Mieder, W. (1993). *International Proverb Scholarship. An Annotated Bibliography. Supplement II (1982-1991).* New York: Garland.

Röhrich, L., & Mieder, W. (1977). *Sprichwort.* Stuttgart: Metzler.

Seiler, F. (1922). *Deutsche Sprichwörterkunde.* München: Beck.

Simrock, K. (1846). *Die deutschen Sprichwörter.* Stuttgart: Reclam, 1988.

Polylexie lexikalischer Einheiten in Texten

Christiane Hoffmann

Einführung

Ziel dieser Arbeit ist die Untersuchung der Lesartenverteilung lexikalischer Einheiten im Text. Diese ist nun erstmals möglich, da ein Teil des Brown Corpus semantisch annotiert vorliegt. Neben der Lesartenverteilung werden weitere Fragestellungen zur Polylexie im Text behandelt, die die Untersuchung des Zusammenhangs von Wortlänge und Wortpolylexie und das „semantische Spektrum" zum Gegenstand haben.

Im folgenden wird der Begriff *Polylexie* in Anlehnung an die Definition in Köhler (1986) verwendet[1]:

> *"Die Anzahl der verschiedenen Bedeutungen, die eine lexikalische Einheit zu einem gegebenen Zeitpunkt trägt [...], soll die POLYLEXIE dieser Einheit genannt werden. Dieser Begriff soll nicht zwischen semantischen und grammatischen Bedeutungen differenzieren, so daß auch den Funktionswörtern eine Polylexie größer Null zukommt."* (Köhler, 1986:59; Auszeichnungen im Original)

Der Begriff der *Polysemie*, von Bréal eingeführt (vgl. Bréal, 1924/1979:143f.), wird traditionellerweise von dem Begriff der *Homonymie* differenziert, wozu ein etymologisches Abgrenzungskriterium herangezogen wird. Diese Differenzierung wird von Köhler nicht verwendet. Weder ist das verwendete Kriterium ein verläßliches,[2] noch spielt es für die von Köhler eingenommene Perspektive eine Rolle.[3] Diese besteht zunächst darin, die quantitativen Verhältnisse, die sich aus der mangelnden Eineindeutigkeit von Form und Funktion ergeben, auf ihre vermutlich durch Sprachgesetze festgelegten Zusammenhänge hin zu untersuchen.

Im folgenden wird ein fragmentarischer Überblick über quantitative Arbeiten zur Polysemie/-lexie gegeben. Danach werden einige Untersuchungen beschrieben, die mit einem semantisch annotierten Korpus als Datengrundlage durchgeführt wurden.

Arbeiten zur Polysemie/Polylexie

Zipfs Hypothesen

Mit Zipf (1949) kann Polysemie als das Resultat eines Diversifikationsprozesses angesehen werden, durch den sich das Bedeutungspotential eines Lexems vergrößert, also an Facettenreichtum gewinnt. Zipf identifiziert das *Prinzip des geringsten Aufwands* (principle of least effort) als Motor der Diversifikation: SprecherInnen möchten aus Gründen z.B. des Gedächtnisaufwands möglichst mehrere/alle Bedeutungen mit einem Wort ausdrücken. Eine Erhöhung der Polysemie „verschlechtert" das Verhältnis der Eineindeutigkeit zwischen Form und Funktion. Auf der Rezeptionsseite wird dieses Bestreben – ebenfalls aus Gründen des geringsten Aufwands – nicht begrüßt, da es zu einem erhöhten Disambiguierungsaufwand führt.

Zipf (1945) untersucht den Zusammenhang zwischen Bedeutungszahl und Frequenz an den 20.000 häufigsten Wörtern einer Textfrequenzzählung, die Material im Umfang von zehn Millionen Texttoken berücksichtigt hat. Es werden Frequenzklassen von je tausend Wörtern gebildet, deren Bedeutungsanzahl gemittelt wird. Beim Auftragen des Ranges der Frequenzklassen als unabhängige und der Bedeutungszahl der Frequenzklassen als abhängige Variable auf eine doppelt logarithmische Skala ergibt sich ein Regressionskoeffizient von ca. -0,5. Je seltener die Wörter, desto geringer ihre Bedeutungszahl im Durchschnitt. Zipf interpretiert den Faktor ½ als Kompromiß zwischen Sprecherin, die *M* Bedeutungen mit einem Wort ausdrücken möchte, und Hörer, der sich *M* Bedeutungen mit *M* Wörtern ausgedrückt wünscht.

Zipf stellt nicht nur einen Zusammenhang zwischen der Texteigenschaft (Parole) „Frequenz" eines Lexems und der Lexikoneigenschaft (Langue) „Poly-

[1] Andere Arbeiten sprechen von der *Bedeutungskomplexität* (vgl. Altmann, Beöthy & Best, 1982), oder vom *semantischen Umfang* (vgl. z.B. Levickij (erscheint); Tuldava, 1998).

[2] Interessant in diesem Zusammenhang ist jedoch die Abgrenzungshypothese der Typologie, daß Polyseme jene Homonyme seien, deren Bedeutungen sich sprachenübergreifend ähneln (vgl. Croft, 1990:166).

[3] Wobei man jedoch anmerken muß, daß das etymologische Kriterium synchron für den Sprachverstehensprozeß „unbekannter" Wörter eine Rolle spielen könnte, indem es für diesen die Möglichkeit der metaphorischen Transferleistung eröffnet. Diese „Ähnlichkeit" in den Bedeutungen ist es wohl auch, die mit dem etymologischen Kriterium operationalisiert werden soll.

semie“ her, sondern untersucht auch die Frage nach dem Verhältnis zwischen dem Grad der Polysemie *PL* und der Anzahl *n* der Lexeme mit diesem Polysemiegrad im Lexikon. Dafür stellte er folgendes Verhältnis auf, in dem *c* eine Konstante darstellt:

$$n_{PL} = \frac{c}{PL^2}$$

Empirisch konnte dieser Zusammenhang, der in einer mathematisch veränderten Form – fälschlicherweise – den Namen Krylov-„Gesetz“ trägt, bislang trotz einiger Modifikationen nicht bestätigt werden oder gar den Status einer gut überprüften Hypothese/eines Gesetzes erlangen.[4]

Polylexie/Polysemie, Länge und Frequenz

Ein Motivationsversuch für den Zusammenhang zwischen Polylexie und Länge besteht in seiner Interpretation durch das Menzerathsche Gesetz, welches die Größe eines Konstrukts in Zusammenhang zu der Größe der in ihm enthaltenen Teile setzt und bisher zumeist für formale linguistische (Ausdrucks-) Einheiten formuliert wurde.

> *“Die Möglichkeiten sind aber damit nicht erschöpft, denn die Größe der Konstrukte (x) braucht nicht unbedingt in der Zahl der als y zu ihnen in bezug gesetzten Konstituenten gemessen zu werden. [...] Eine der Beziehungen, die nach unseren Annahmen dem Menzerathschen Gesetz folgen sollte, ist die Verringerung der Menge der lexikalischen Bedeutungen des Wortes bei wachsender Wortlänge [...].“* (Altmann, Beöthy & Best, 1982:537)

Der funktionale Mechanismus hinter dieser Beziehung ist darin zu sehen, daß eine Verlängerung des Wortes, die meistens durch Morphe vorgenommen wird, zu einer Bedeutungsspezifikation führt, welches wiederum einer Reduktion der möglichen Anzahl der Bedeutungen entspricht (*Hochhaus* ist spezifischer und kann somit in weniger Kontexten verwendet werden bzw. hat weniger Lesarten als *Haus*). Es ist nicht unbedingt plausibel, das Menzerathsche Gesetz für den vorliegenden Zusammenhang heranzuziehen, denn Konstrukt und Komponenten stellen in diesem Zusammenhang dieselben Einheiten dar.[5]

Die erste Überprüfung erfolgt in Altmann, Beöthy & Best (1982) an Daten

[4] Vgl. Krylov (1982), Hammerl (1991:138ff.) für einen Überblick und Tuldava (1998: 120ff.).

[5] Zur weiteren Diskussion vgl. Köhler (1986:10f.), der auf diese Herleitung verzichtet, vgl. unten.

des Deutschen, Slowakischen und Ungarischen, wobei die AutorInnen bei der Operationalisierung der Bedeutungskomplexität als ‚Anzahl der Lesarten im Lexikon‘ zwischen Homonymie und Polysemie unterscheiden, wie ihre Erklärungen zur Datenerhebung andeuten. Die Ausprägung $y = ax^b$ des Menzerathschen Gesetzes erzielt signifikante Ergebnisse für die Operationalisierung der Länge sowohl in Anzahl der Buchstaben als auch in Anzahl der Silben. Der Parameter *a* muß als mittlere Zahl der Bedeutungen von Wörtern der Länge 1 interpretiert werden, da $y = a$, wenn $x = 1$ (vgl. Altmann, Beöthy & Best, 1982:542).

Das angepaßte Modell verläuft asymptotisch (mit der Asymptote $y = 0$), was Wörter mit einer Polysemie von 0 nahelegt. Dieser Gedanke befremdet zunächst; Eigennamen können aber durchaus als Wörter mit null Lesarten interpretiert werden, wie die Autoren meinen. Weitere Untersuchungen in dieser Richtung führt u.a. Rothe (1983) für die romanischen Sprachen Französisch, Portugiesisch und Spanisch mit signifikanten Ergebnissen durch. Hier wird der Parameter *a* ebenfalls als „theoretische mittlere Zahl der Bedeutungen für die Wörter mit der Länge 1“ (Rothe, 1983:102) interpretiert. Tabelle 1 vermittelt einen Überblick über einige Ergebnisse.

Tabelle 1
Parameter für die Anpassungen bei Buchstaben- und Silbenoperationalisierung der Länge in verschiedenen Sprachen, nach Rothe (1983), Altmann, Beöthy, Best (= ABB) (1982), Fickermann, Markner-Jäger, Rothe (= FMR) (1984) und Köhler (1986)

	Buchstaben		*Silben*	
Sprache	a	b	a	b
Slowakisch (ABB)	14,7	-1,0875	3,6	-0,7136
Deutsch (Köhler)	12,5	-0,8280	Nicht erhoben	
Deutsch (ABB)	4,9	-0,4149	2,7	-0,2885
Ungarisch (ABB)	9,1	-0,6852	4,2	-0,5636
Französisch (Rothe)	21,5	-1,0486	5,2	-0,8309
Portugiesisch (Rothe)	28,7	-1,1198	11,3	-1,1665
Spanisch (Rothe)	57,4	-1,4450	14,5	-1,3335
Schwedisch (Rothe)	6,62	-0,7421	2,46	-0,4987
Indonesisch(FMR)	3,61	-0,3386	2,8	-0,3758
Englisch (FMR)	33,2	-1,3556	5,2	-0,9802

Vergleicht man den Parameter *b*, der nach Köhler (s.u.) den Synthetizitätsgrad einer Sprache ausdrückt, aus Buchstaben- und Silbenoperationalisierung und die Rangfolge, die er den Sprachen zuordnet, so wird der statistische

Spearman-Korrelationstest signifikant:[6] Beide Operationalisierungen ordnen die Sprachen in derselben Art und Weise hinsichtlich ihres Synthetizitätsgrads.

Es befremdet, daß die beiden *b*-Werte für das Deutsche, die in unterschiedlichen Untersuchungen erhoben wurden, anscheinend erheblich voneinander abweichen. Die Datengrundlage bei Altmann, Beöthy, Best (1982) besteht aus jedem dritten Wort jeder Seite aus Wahrig (1978)[7], was einen Stichprobenumfang von 913 Wörtern ergibt, Köhler (1986) ermittelt die Polylexie aus einer anderen Wahrig-Ausgabe[8] für eine Stichprobe von 1325 Lemmata des LIMAS-Korpus. Die Unterschiede im Deutschen sind wahrscheinlich mit den unterschiedlichen Stichprobengrößen zu erklären, die große Streuung der Werte insgesamt vielleicht nicht nur durch die unterschiedlichen Sprachen selbst, sondern auch durch unterschiedliche Wörterbuchpraktiken. Vergleicht man bspw. die Einträge im ALD[9] mit denen aus Webster[10], so ergibt sich das in Tabelle 2 dargestellte Bild, welches vermuten läßt, daß die Polysemieinformation gleicher Lemmata in verschiedenen Wörterbüchern signifikant divergiert.

Tabelle 2
Zahl der Bedeutungsangaben in ALD und Webster bei verschiedenen Einträgen der Wortlänge 1 (in Buchstaben)

Eintrag	*Webster*	*ALD*
a^1	6	1
a^2	4	12
a^3	1	-
a^4	1	-
a^5	1	-
B	6	1

Fickermann, Markner-Jäger, Rothe (1984) stellen eine direkte Abhängigkeit zwischen den Parametern *a* und *b* fest, für die interessanterweise wiederum das Modell $y = ax^b$ angepaßt werden kann. Des weiteren vermuten die Autoren, daß es sich bei *b* um eine sprachenspezifische Größe handelt, und interpretieren den

[6] Korrelation 0,933 auf 0,01 %-Niveau

[7] Wahrig, G. (Hrsg.): dtv-Wörterbuch der deutschen Sprache. München, 1978. (Zitiert nach Altmann, Beöthy & Best, 1982)

[8] Wahrig, G.: Deutsches Wörterbuch. Gütersloh, 1981. (zitiert nach Köhler, 1986)

[9] Hornby, A.S. (Hrsg.): Oxford advanced learner's dictionary of current English. Berlin u.a.: Cornelsen u.a., 3.Aufl., 11. Dr., 1980

[10] Webster's seventh new collegiate dictionary. Based on Webster's third new international dictionary. Springfield, Mass.: Merriam, 1972

Parameter im Sinne von Köhlers (1984) Erklärungsansatz zum Menzerathschen Gesetz allgemein als „Maß für den Umfang der Strukturinformation, die durchschnittlich für ein einkomponentiges Konstrukt erforderlich ist".

Altmann und Schwibbe (1989:77ff.) gelingt es, den funktionalen Zusammenhang

$$Polysemie = a * Häufigkeit^b$$

an slowakische und russische Daten von Frequenzrängen und mittleren Bedeutungszahlen anzupassen, jedoch nur mit einem schwach signifikanten Ergebnis für das Russische. Die Autoren nennen den Zusammenhang das „Zipf-Guitersche Gesetz", an Guiter (1974) erinnernd, der bereits den Zusammenhang zwischen Polysemie, Wortlänge und Frequenz an verschiedenen Sprachen untersuchte.

Das synergetische Modell der Lexik

Köhler (1986) integriert die lexikalische Eigenschaft Polylexie als Systemvariable in ein synergetisches Modell der Lexik. Dort steht die Polylexie in direktem Zusammenhang zur Wortlänge, der Zusammenhang wird jedoch anders als in Altmann, Beöthy & Best (1982), vgl. oben, nicht über das Menzerathsche Gesetz begründet, sondern über den ebenfalls bereits von diesen Autoren angedeuteten Mechanismus der Bedeutungsspezifikation durch morphosyntaktische Mittel, die die Länge des jeweiligen Ausdrucks erhöhen und seine Polylexie reduzieren.

> *„Die Abhängigkeit der Polylexie von der Länge ist um so stärker, je mehr eine Sprache von morphologischen gegenüber syntaktischen Mitteln zur Bedeutungsspezifikation Gebrauch macht. Diese typologische Eigenschaft einer Sprache soll Synthetizität heißen."* (Köhler, 1986:60)

Köhler interpretiert den Parameter *a* als Kompromiß/Fließgleichgewicht, das aus dem Einfluß der Bedürfnisse *Minimierung des Kodierungsaufwands* und *Minimierung des Dekodierungsaufwands*, welche als die Zipfschen Kräfte der Unifikation und Diversifikation verstanden werden können, resultiert. Der absolute Wert des Parameters *b* wird als Synthetizitätsgrad einer Sprache gedeutet, gibt also das Maß der Verwendung morphologischer Mittel zur Bedeutungsspezifikation an. Dieser Deutung der Parameter steht die oben vorgestellte von *a* als durchschnittlicher Polysemie der Wörter der Länge 1 und *b* als Maß für den Umfang der Strukturinformation gegenüber. Köhlers Formulierung von *a* kann als funktionale Präzisierung der bisherigen Deutung verstanden werden. Auch die von Fickermann, Markner-Jäger und Rothe (1984) festgestellte Abhängigkeit der Parameter *a* und *b* ist bei Köhlers Deutung der Parameter motivierbar: Je höher

die durchschnittliche Polylexie der Wörter, desto stärker wird von u.a. lexikalischen Mitteln der Bedeutungsspezifikation Gebrauch gemacht.[11]

Interessant für die nachfolgenden Untersuchungen ist Giesekings (1993, erscheint) Beitrag, in dem Köhlers (1986) Modell für das Englische überprüft wird. Bei der Überprüfung der indirekten Abhängigkeit zwischen Länge und Polylexie zeigt sich bei der Operationalisierung der Länge als Anzahl der Phoneme bzw. Grapheme eine nicht monotone empirische Verteilung, die zunächst bis zur Länge drei bzw. vier ansteigt, um danach stetig zu fallen. Gieseking paßt ebenfalls das Menzerathsche Gesetz an, jedoch auch in der Ausprägung $y = ax^b e^{cx}$, und erzielt bessere Anpassungsergebnisse als mit der Ausprägung $y = ax^b$. Sie stellt die Vermutung an, daß diese in bisherigen Untersuchungen nicht aufgetretene Eigenschaft der Daten daraus resultiert, daß Lexeme mit geringer Länge typischerweise „Funktionswörter" wie Präpositionen, Konjunktionen, Pronomen sind, deren Bedeutungsnuancen im Lexikon nicht adäquat wiedergegeben werden. Bei einer genaueren Inspektion der Stichprobe kommt sie jedoch zu dem Schluß, daß es sich – zumindest bei den Wörtern der Längen drei und vier – um Inhaltswörter handelt. Auch eine nach Wortklassen homogenisierte Untersuchung der Stichprobe erbrachte für einzelne Klassen wie Substantive und Verben den Effekt der Nicht-Monotonie.

Alle Untersuchungen, die mit der Operationalisierung der Länge in Anzahl der Silben arbeiten, resultieren in empirisch monotonen Verteilungen.

Des weiteren wirkt die Polylexie auf die Systemgröße Polytextie, die Anzahl der verschiedenen Kontexte, in denen ein Wort vorkommt: je mehr Bedeutungen ein Wort besitzt, desto kontextunabhängiger kann es verwendet werden. Ebenso beeinflußt die Polylexie die Lexikongröße: je höher die Polysemie der Lexikoneinträge im Durchschnitt, desto kleiner das Lexikoninventar bei konstantem Ausdrucksbedürfnis.

Als indirekte Abhängigkeit modelliert Köhler (vgl. Köhler, 1986:113ff.) den Zusammenhang zwischen Polylexie und Frequenz, der auch bereits von Zipf untersucht wurde, vgl. oben. Da im synergetischen Modell die Länge direkt von der Frequenz abhängt und die Polylexie direkt von der Länge, kann durch Einsetzen die indirekte Beziehung der Abhängigkeit der Polylexie von der Frequenz gewonnen werden. Die empirische Überprüfung erzielt eine hohe Anpassungsgüte.

Weitere Arbeiten zur Polylexie

Aufbauend auf den Überlegungen Zipfs zur Diversifizierung der semantischen Bedeutung sprachlicher Einheiten formulieren Altmann (1985) und Altmann, Best und Kind (1987) das „Gesetz der semantischen Diversifikation". Methodische und empirische Untersuchungen dazu finden sich in Rothe (1991) dokumentiert.

[11] Köhler, pers. Mitteilung

Weitere Überlegungen zur Herleitung einer Verteilungsfunktion für die semantische Diversifikation findet sich in Leopold (1998).

Für einen Überblick zu quantitativen Untersuchungen zur Polysemie vgl. auch Levickij (erscheint).

Die Daten

In der vorliegenden Arbeit werden verschiedene Untersuchungen zur Polylexie im Text durchgeführt. Mit dem Teil des Brown Corpus, der mittels Word Net® semantisch annotiert vorliegt, ist es möglich, quantitative Untersuchungen zur Verteilung der Polylexie an einer größeren Menge von Texttoken und zum Zusammenhang der Polylexie mit anderen sprachlichen Eigenschaften der Lexik auf der Textebene durchzuführen.

Word Net®

Word Net®[12] ist ein maschinenlesbares Lexikon des Englischen, das an der Princeton University erstellt wird und ein Modell des mentalen Lexikons darstellen soll. Die Lemmata der offenen Wortklassen werden zu sogenannten „Synsets" zusammengefaßt, die Lemmata „gleicher" bzw. annähernd gleicher Bedeutung gruppieren. Zusätzlich sind semantische und lexikalische Relationen zwischen Synsets bzw. Lemmata definiert. Die Anzahl der Bedeutungen eines Lemmas läßt sich mit WordNet als Anzahl der Synsets, in denen es fungiert, operationalisieren. Das Lexikon hat in der Version 1.6 einen Umfang von 121.962 Lemmata, 99.642 Synsets und 173.941 Lesarten insgesamt.

Das Korpus

Im Rahmen des Word Net-Projekts wurden 186 Dateien des Brown Corpus (ca. 360.000 laufende Wortformen) zunächst syntaktisch und in einem weiteren Schritt semantisch annotiert, (vgl. Landes, Leacock & Tengi, 1998).[13] Die semantisch annotierten Dateien sind wiederum aufgeteilt auf Brown1 und Brown2. Das Brown Corpus ist ein Korpus der sog. „ersten Generation", das aus insgesamt einer Million laufender Wortformen besteht. Das Korpus setzt sich aus 500 Dateien zusammen, die jeweils ca. 2000 Wörter enthalten und

[12] vgl. Fellbaum (1998). Das Lexikon ist frei verfügbar, um Registrierung wird gebeten. Auf der Homepage http://www.cogsci.princeton.edu/~wn/ finden sich Informationen zum Download, zur Konzeption und Literaturhinweise zu Untersuchungen von und Anwendungen mit WordNet.

[13] Es liegen weitere 166 Dateien des Brown Corpus vor, in denen nur die Verben semantisch annotiert sind. Des weiteren wurde ein kompletter Roman (Stephen Crane: The Red Badge of Courage) semantisch annotiert, der jedoch nicht frei verfügbar ist.

Textstücke verschiedener Textsorten darstellen. In der Auswahl für das semantische Tagging sind ebenfalls verschiedene Textsorten berücksichtigt.

Die semantischen Annotationen sind nur an jenen Wortformen (ca. 170.000) vorgenommen, deren Lemma im Lexikon WordNet vertreten ist, und befinden sich somit nur an den Wortformen der sogenannten „offenen Wortklassen": Verben, Substantive, Adjektive, Adverbien. Die Nominalphrasen, die Eigennamen enthalten, wurden mit einer „generischen" semantischen Annotation „Eigen-name" versehen, die nach *person*, *location*, *group* und *other* differenziert ist. Diese Wortformen wurden für die vorliegenden Untersuchungen ausgeschlossen, obwohl sich auch *für* ihren Einschluß argumentieren läßt.

Untersuchungen zur Polylexie im Text

Operationalisierung der untersuchten Eigenschaften

Notwendige Voraussetzung für die Überprüfung quantitativer Hypothesen ist die Operationalisierung der verwendeten Variablen. Letztere repräsentieren die Eigenschaften von sprachlichen Einheiten, die im Modell der Synergetischen Linguistik wiederum als Systemgrößen fungieren.

Bei den im folgenden untersuchten Variablen Länge und Polylexie handelt es sich um Eigenschaften von Lexemen, die auf unterschiedliche Weise operationalisiert werden können. Die *Länge* eines Lexems kann in der Anzahl der Grapheme, Phoneme, Phone, Morpheme oder Silben gemessen werden. Altmann und Schwibbe untersuchten bspw. den Zusammenhang zwischen Polylexie und Länge, gemessen sowohl in Anzahl der Silben als auch in Anzahl der Buchstaben, und fanden bei beiden Arten der Operationalisierung signifikante Ergebnisse (vgl. Altmann & Schwibbe, 1989:67ff.). Dies ist ideal und bedeutet auch, daß zwischen den einzelnen Metrisierungen eine numerische Transformation vorgenommen werden kann. Weiterhin können viele sprachliche Eigenschaften sowohl im Lexikon als auch im Text gemessen werden. Im folgenden wurde die Länge in Buchstaben im Lexikon gezählt, also die Lemmalänge, nicht die Wortformenlänge berücksichtigt.

Polylexie wurde bislang zumeist als Anzahl der Lesarten im Lexikon operationalisiert (vgl. Köhler, 1986:92). Auch in WordNet kann man die Polylexie als die Anzahl der „Synsets" verstehen, in denen ein Lemma Mitglied ist.

Anhand der vorliegenden semantisch annotierten Daten ist es nun erstmals möglich, auch einen Wert für eine Text- bzw. Stichprobenpolylexie zu ermitteln und diese – statt des Polylexiewerts aus dem Lexikon – zu verwenden. Der Wert für die Textpolylexie ist nicht unabhängig von dem im Lexikon: die Lexikonpolylexie bildet den maximal möglichen Wert für die Textpolylexie. Im Text gibt die Lesartenannotation in einer eindeutigen Identifikation an, welchem Synset die Lesart zuzuordnen ist.

Verteilung der Lesarten: das Spektrum

Eine Frage, die sich in bezug auf die Polylexie lexikalischer Einheiten in der Parole stellen läßt, betrifft die der Lesartenfrequenzen. Während die Lesartenangaben zu einem Lexem im Wörterbuch zumeist ohne weitere quantitative Angaben erfolgen,[14] können korpuslinguistische Untersuchungen Aufschluß über Lesartenfrequenzen geben. Abgesehen von rein deskriptivem und theoretischem Interesse hat die Antwort auf diese Frage Konsequenzen für angewandte Teilbereiche der Linguistik. So wäre es sinnvoll, den Lernwortschatz im Fremdsprachenunterricht auf die Frequenzen der einzelnen Wörter/Lesarten hin zu überdenken. Ebenso wie beim probabilistischen Tagging und Parsing die Frequenzen der einzelnen Konstruktionen und Kombination beim Verfolgen und Ermitteln der wahrscheinlichsten Lösungen eingehen, können Daten über die Verwendungsfrequenzen von Lesarten Unterstützung bei der automatischen Übersetzung bieten oder in anderen Disambiguierungsfällen verwendet werden.

Tabelle 3
Spektrum: Anpassungsgüte für verschiedene Verteilungen

Verteilung	*Freiheitsgrade*	χ^2	$P(\chi^2)$	*Kontingenz-koeff.*
Neg. bin.-Poisson (k,a,p) $k = 0{,}1$; $a = 422{,}24$; $p = 0{,}98$	47	46,41	0,0043	0,0023
Poisson-Pascal (a,k,q) $a = 0{,}01$; $k = 3{,}33$; $q = 0{,}001$	144	165,19	0,1092	0,0050
Waring (b,n) $b = 1{,}09$; $n = 1{,}27$	302	339,10	0,0	0,0103
Zipf-Mandelbrot (a,b,n) $a = 2{,}08$; $b = 0{,}73$; $n = 5364$	309	338,69	0,0	0,0103

Die Untersuchung der Lesartenfrequenzen insgesamt als *Spektrum* beantwortet die Frage danach, wie viele Lesarten der Frequenz x im Korpus vorkommen. Berücksichtigt wurden die Lesarten von 32867 Wortformen in den semantisch annotierten Korpusteilen Brown1 und Brown2. Es ergibt sich bspw., daß es 15209 Lesarten gibt, die nur einmal vorkommen, 5518 Lesarten, die zweimal vorkommen usw. Tabelle 4 zeigt in Spalte 2 den Anfang und das Ende der empirischen Frequenzklassenverteilung.

[14] In WordNet® spielen jedoch die Lesartenfrequenzen in der Präsentation der Synsets eine Rolle (die häufigste zuerst). Um dieses Merkmal implementieren zu können, wurde u.a. das semantisch annotierte Korpus erstellt.

Tabelle 4
Spektrum der Lesarten

Lesarten-frequenz	*empirisch n=32867*	*Negative binomial-Poisson*	*Poisson-Pascal*	*Waring*	*Zipf-Mandelbrot*
1	15209	32803,22	32697,61	15209,00	14910,17
2	5518	0	0	5736,53	5790,40
3	2849	0	0	2983,87	3030,89
4	1798	0	0	1819,38	1851,75
5	1261	0	0	1221,05	1243,86
...	...	...	...	...	...
5364	1	0	20,15	0	0

Leider besteht momentan keine Hypothese darüber, welcher Verteilung das Spektrum der Lesarten folgt, außer der schwachen Analogiehypothese, daß das Spektrum der Lesarten möglicherweise derselben Verteilung wie das Spektrum der Wortformen, also der Zipf-Mandelbrot-Verteilung oder der Waring-Verteilung, folgt. Mittels der Software Altmann-Fitter® wurde eine automatische Anpassung vorgenommen, die die in Tabelle 3 vorgestellten Ergebnisse brachte. Diese möchte ich zunächst nur als deskriptive Formulierung der Daten verstanden wissen.

Bei der vorliegenden Stichprobengröße kann die Güte der Anpassung nur schlecht mittels $P(\chi^2)$ beurteilt werden, statt dessen wird der Kontingenzkoeffizient herangezogen.[15] Hier zeigt sich zunächst eine numerisch akzeptable Anpassung mit den Verteilungen Neg. binomial-Poisson und Poisson-Pascal. Die theoretischen Werte (vgl. Spalten 3 und 4 in Tabelle 4), die nur für die erste (bzw. letzte) Klasse nennenswerte Werte vorhersagen, legen nahe, diese Verteilungen aus inhaltlichen Gründen zu verwerfen. Hier rächt sich nun die wenig Theorie-geleitete Vorgehensweise: zwar erzielen wir mit den beiden Verteilungen akzeptable Inferenz-statistische Ergebnisse; diese ergeben jedoch ohne theoretische Annahme im vorliegenden Fall keinen Sinn: Aus linguistischer Sicht sollte keine Verteilung von Interesse sein, die vorhersagt, daß in einem Text fast alle Lesarten nur einmal vorkommen, oder gar eine, die vorhersagt, daß die zweithäufigste Lesartenfrequenz die Klasse mit der häufigsten Lesart ist!

Die Anpassungsversuche mit der Waring-Verteilung und der Zipf-Mandelbrot-Verteilung sind aus linguistischer Sicht interessanter, da sie die Analogie-

[15] Dieser wird im Altmann-Fitter® mittels der Formel $C = \chi^2/n$, mit n = Stichprobengröße, bestimmt. Bei der Signifikanzbeurteilung gilt, daß die H_0 (hier die Wunschhypothese) nicht abgelehnt werden muß, wenn gilt: $C < 0{,}015$.

hypothese stützen. Abb. 1 zeigt das empirische Spektrum und die theoretischen Werte der Waring-Verteilung in logarithmischer Darstellung.

Spektrum - Waring-Anpassung

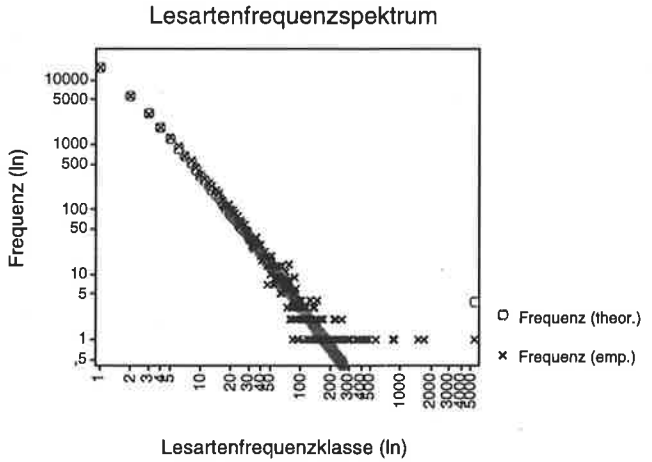


Abb. 1. Spektrum: Waring-Verteilung (ln)

Spektrum - Zipf-Mandelbrot-Anpassung

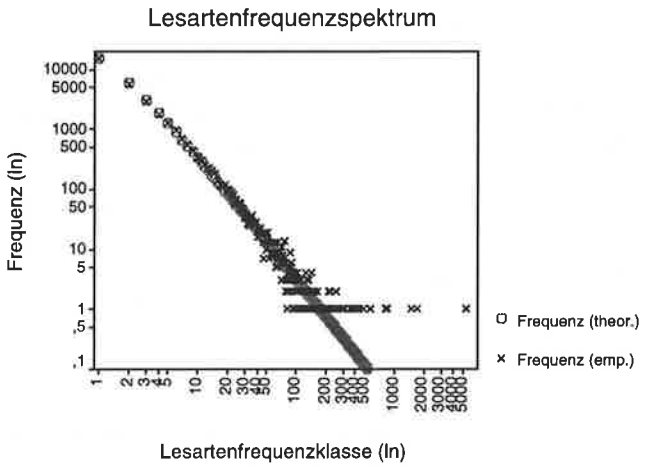


Abb. 2. Spektrum: Zipf-Mandelbrot-Verteilung (ln)

Der Zusammenhang Polylexie – Länge

Dimensionen für die Untersuchung Polylexie – Länge
In den ersten Untersuchungen zur Abhängigkeit der Polylexie von der Länge wurden die Daten zumeist so erhoben, daß für alle Längenklassen im Lexikon die durchschnittliche Polylexie im Lexikon ermittelt wurde.[16] Bei der Untersuchung der Abhängigkeit der Polylexie von der Länge ist es prinzipiell möglich, entlang (mind.) fünf unterschiedlicher Dimensionen verschiedene Messungen vorzunehmen:

- **Sprachliche Datengrundlage**

Die Untersuchung kann am Lexikon oder am Text vorgenommen werden. Wird sie am Text vorgenommen, so kann
1. die Messung der Eigenschaften unterschiedlich gestaltet werden (s.u.) und
2. die Untersuchung auf *Types* oder *Token* durchgeführt werden (s.u.).

- **Messung der Eigenschaften**

Die Messung der Eigenschaften Länge und Polylexie kann im Text/Korpus oder im Lexikon vorgenommen werden. Wird die Länge im Text gemessen, so wird die Länge der Wortform gemessen. Im Lexikon wird die Länge des Lemmas gemessen. Wird die Polylexie im Lexikon (im folgenden *Lexikonpolylexie*) gemessen, so wird dort die Anzahl der Lesarten für ein Lemma ermittelt. Wird sie im Text/Korpus (im folgenden *Textpolylexie*) gemessen, so wird für ein Lemma in einem bestimmten Korpus/Text ermittelt, in wieviel verschiedenen Lesarten es vorkommt. Die Textpolylexie kann nur an einem semantisch annotierten Korpus ermittelt werden. Ihr Wert ist immer kleiner oder gleich der Lexikonpolylexie in bezug auf das Lexikon, mit dessen Hilfe Text bzw. Korpus semantisch annotiert wurden.

Die Unterscheidung in Lexikon- und Textpolylexie kann in Untersuchungen die Frage beantworten helfen, ob sich der Ausnutzungsgrad der Textpolylexie proportional zur Lexikonpolylexie verhält bzw. ob er so gestaltet ist, daß die funktionale Abhängigkeit zwischen Länge und Polylexie weiterhin gilt.

- **Type oder Token**

Wird die Untersuchung im Text durchgeführt, können für jede laufende Wortform eines Texts Länge und Polylexie ermittelt werden, vgl. Gieseking (1993,

[16] Vgl. jedoch Hammerl (1990) für weitere Operationalisierungen.

erscheint). Dadurch, daß die Längenklassenbildung und die Mittelung der Polylexie auf den gesamten *Token* eines Textes beruht, ergibt sich ein anderes Bild als bei bisherigen Untersuchungen, die auf Lexikoneinträgen basierten, oder als bei Untersuchungen, die die *Types* eines Texts/Korpus berücksichtigen.

- **Unterscheidung nach Wortarten**

Weiterhin ist es möglich, Homonyme, die verschiedenen Wortkategorien angehören, als einen oder als zwei bzw. mehrere Lexikoneinträge zu betrachten. Dementsprechend ergeben sich unterschiedliche Polylexiewerte.

- **Direkte Gewichtung bei der Anpassung**

Bei der Anpassung eines Modells sollten die einzelnen Werte für die abhängige empirische Variable Polylexie mit der absoluten Frequenz der Werte, aus denen sie gemittelt wurden, gewichtet werden. Dies führt zu einem „realistischeren" Bild der Anpassung.

Untersuchungsdesign
In den im folgenden beschriebenen Untersuchungen wurden entlang der oben geschilderten Dimensionen drei verschiedene Untersuchungsdesigns gewählt. Für alle drei Untersuchungen gilt, daß

- Types betrachtet werden,
- bei der Anpassung direkt gewichtet wird
- und daß Wortarten unterschieden werden.

Untersuchung I ist die „*klassische*" Untersuchung des Zusammenhangs im Lexikon.
Untersuchung II berücksichtigt nur die Lemmata, die im Korpus Brown1 vertreten sind. Dabei wird der Polylexiewert aus der *Lexikonpolylexie* ermittelt, der Längenwert am Lemma.
Untersuchung III entspricht Untersuchung II, nur das hier die *Textpolylexie* verwendet wurde.

Untersuchungsergebnisse
Die Untersuchungsergebnisse wurden, dem mathematischen Modell $Polylexie = a*Länge^b$ mittels nichtlinearer Regression angepaßt.[17] Die Güte der Anpassung wird mit dem Determinationskoeffizienten bestimmt. Die Untersuchungsergeb-

[17] Für eine linguistische Motivation des Zusammenhangs vgl. oben.

nisse der Erhebungen mit unterschiedlichen Einstellungen in den Dimensionen finden sich in Tabelle 5.

Tabelle 5
Polylexie – Länge: Untersuchungen im Lexikon und im Text

Untersuchungsdesign	*I*	*II*	*III*
Datengrundlage	Lexikon	Text	Text
Operationalis. d. Polylexie	Lexikonpolylexie	Lexikonpolylexie	Textpolylexie
Anpassung			
Parameter a	4,35	2,41	2,95
Parameter b	-0,53	-0,247	-0,38
Determinationskoeff.	0,86478	0,86054	0,76223

Abb. 4 zeigt die klassische Untersuchung des Zusammenhangs im Lexikon. Der Determinationskoeffizient ist relativ gut. Abb. 5 stellt die Ergebnisse der Untersuchung am Text dar. Auch diese Anpassung ist relativ gut. Vergleicht man I, II und III, so sind vor allem Unterschiede in den Parametern festzustellen, die aus der sehr unterschiedlichen Datengrundlage resultieren und nur schlecht miteinander verglichen werden können. Vergleiche sollten eher in der Art angestellt werden, wie es oben zu den Daten verschiedener Sprachen durchgeführt wurde, die in Tabelle 1 dargestellt sind: hier wurden sprachenübergreifende Ergebnisse von zwei Untersuchungsdesigns im Hinblick auf die konsistente Interpretierbarkeit des Parameters *b* verglichen. Untersuchung III ergibt eine weniger gute Anpassung an die Zusammenhangshypothese.

Wichtig ist festzuhalten, daß der Zusammenhang zwischen Länge und Polylexie auch bei Operationalisierungen mit Textpolylexie recht gut haltbar ist. Dies läßt ein einigermaßen konstantes Verhältnis zwischen Lexikon- und Textpolylexie vermuten, das es weiter zu untersuchen gilt. Auch wäre es interessant, der Frage nachzugehen, inwieweit Untersuchungsdesignunterschiede hinsichtlich Types bzw. Token die Anpassungsergebnisse verändern.

Als charakteristisch, wie bereits oben diskutiert, erweist sich in allen drei Untersuchungen die Unimodalität der empirischen Daten. Eine Anpassung des mathematischen Modells $y = ax^b e^{cx}$ brächte sicherlich gute Ergebnisse, erschwerte aber die linguistische Interpretation.

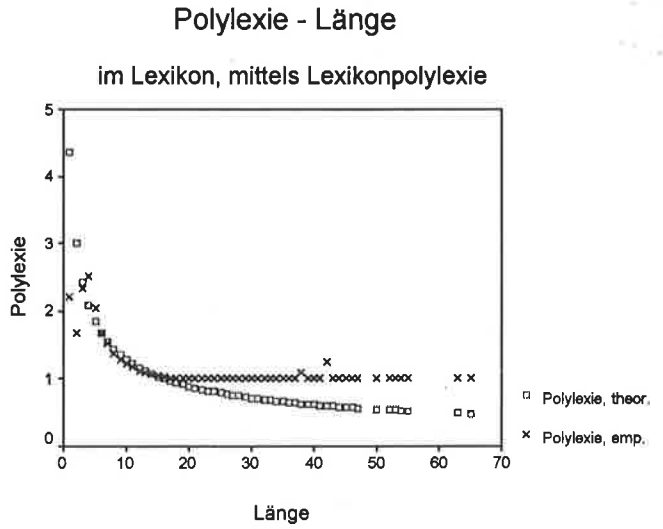


Abb. 3. Polylexie – Länge, untersucht (I) im Lexikon, mittels Lexikonpolylexie

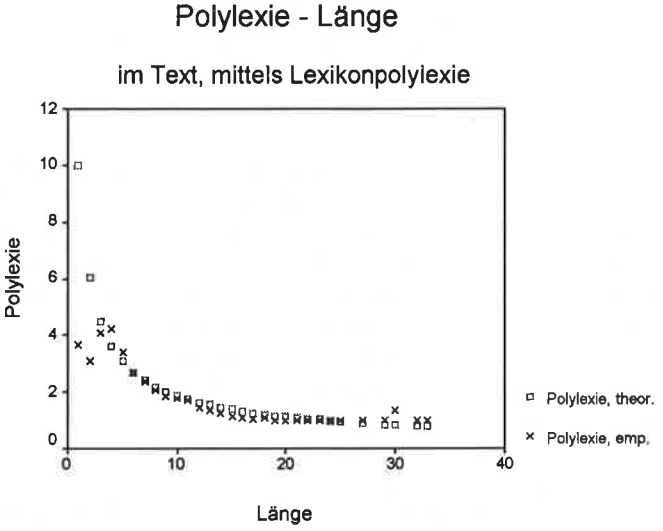


Abb. 4. Polylexie – Länge: untersucht (II) im Text, mittels Lexikonpolysemie

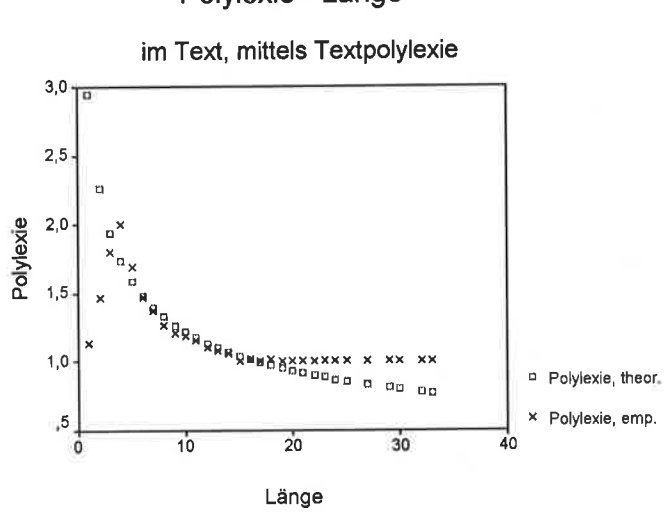


Abb. 5. Polylexie – Länge: untersucht (III) im Text, mittels Textpolylexie

Das semantische Spektrum

In ihrem Artikel von 1982 zum Zusammenhang von Polylexie und Wortlänge erwähnen Altmann, Beöthy und Best in einem Nebensatz eine recht plausible „ad-hoc-Hypothese", die nur anhand von Textmaterial, nicht jedoch im Lexikon untersucht werden kann:

> *„Die vorgestellte Beziehung könnte auch für die semantische Charakterisierung eines Textes verwendet werden. Im Text wird zwar die Bedeutung des Wortes durch den Kontext eingeschränkt, aber in unterschiedlich starkem Maß. In wissenschaftlichen Texten ist die Präzisierung stärker als in poetischen Texten, in denen dem Leser eine bestimmte Interpretationsfreiheit bleibt. Wissenschaftliche Texte enthalten z.B. viele eindeutige Termini (und gleichzeitig die längsten Wörter), poetische Texte dagegen benutzen gern vieldeutige Wörter."* (Altmann, Beöthy & Best, 1982:542f.)

Ausgehend von dieser Überlegung wurden aus dem semantisch annotierten Brown Corpus zwei Textsorten-spezifische Stichproben aus wissenschaftlichen und Prosatexten erstellt. Die Hypothese gemäß dem obigen Zitat lautet, daß in Prosatexten mehr Wortformen mit hoher Polylexie im Lexikon vertreten sind als in wissenschaftlichen Texten und daß umgekehrt mehr Wortformen mit niedriger

Polylexie in wissenschaftlichen Texten vertreten sind als in Prosatexten. Um diese Hypothese zu überprüfen, wurden zwei empirische Verteilungen erhoben, die die Häufigkeiten von Wortformen mit der Lexikonpolylexie 1, 2 etc. in den beiden Stichproben erfassen. Ein Überblick über die Daten beider Verteilungen[18] findet sich in Tabelle 6 und in Abb. 6.[19] Abb. 6 enthält in den Klassen nicht die absoluten, sondern die relativen Häufigkeiten.

Tabelle 6
Semantisches Spektrum: wissenschaftliche (wiss) und Prosatexte (pros)

Poly-lexie	Genre		Polylexie	Genre		Polylexie	Genre	
	pros	wiss		pros	wiss		pros	wiss
1,00	5348	77	19,00	500	477	37,00	137	22
2,00	3629	5528	20,00	261	167	38,00	112	21
3,00	2873	4012	21,00	244	153	39,00	65	29
4,00	2829	3822	22,00	381	522	40,00	5	8
5,00	2300	3156	23,00	530	238	42,00	122	74
6,00	2146	3019	24,00	91	80	43,00	108	48
7,00	1483	2430	25,00	154	102	44,00	21	23
8,00	1435	1424	26,00	40	16	45,00	53	21
9,00	1476	1729	27,00	160	78	46,00	106	119
10,00	971	981	28,00	171	56	47,00	24	18
11,00	1198	1147	29,00	61	15	48,00	41	26
12,00	1157	893	30,00	82	28	50,00	148	160
13,00	2987	3034	31,00	70	55	53,00	40	11
14,00	568	391	32,00	110	49	72,00	7	6
15,00	481	367	33,00	32	19	78,00	9	12
16,00	452	288	34,00	160	33			
17,00	719	757	35,00	22	72			
18,00	381	464	36,00	70	57	Gesamt	36570	44022

[18] Vollkommen offen hingegen bleibt bislang die Frage nach der theoretischen Verteilung des semantischen Spektrums, die sich keiner dem Altmann-Fitter® bekannten Verteilung mit signifikantem Ergebnis anpassen ließ.

[19] Die Verteilung der Polysemie über die Wortformen einer Stichprobe wird hier in Anlehnung an Altmann u.a. (s.o.) als semantisches Spektrum bezeichnet. Der Begriff ist jedoch nicht mit dem des *Spektrums* als *Verteilung der Lesartenfrequenzen* zu verwechseln.

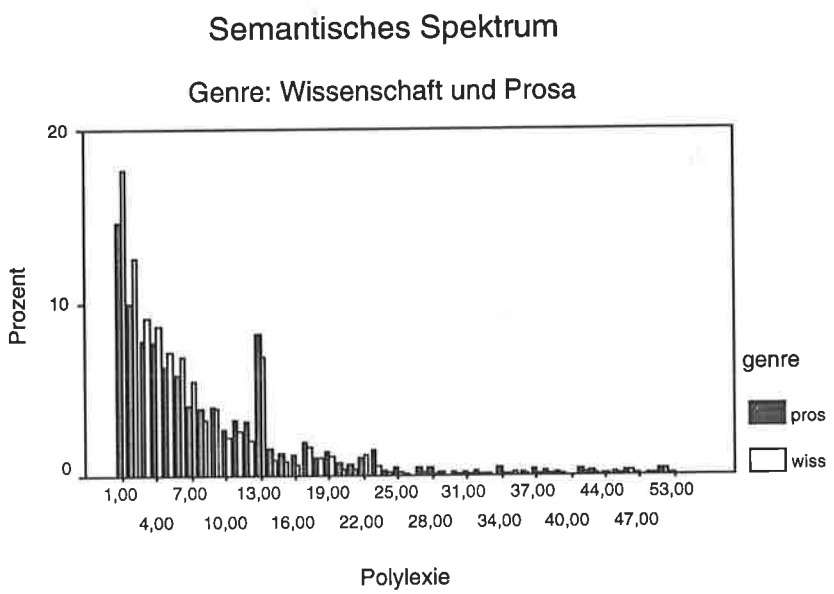


Abb. 6. Semantisches Spektrum der Lexikonpolylexie in wissenschaftlichen (wiss) und in Prosatexten (pros) (relative Häufigkeiten) [20]

Eine erste Inspektion der Daten in tabellarischer und Diagrammdarstellung spricht für die Hypothese: in den wissenschaftlichen Texten (Mittelwert: 7,1) liegt die Zahl der Wortformen mit geringer Lexikonpolylexie höher als in Prosatexten (Mittelwert: 9,0).

Statistisch absichern läßt sich dieses Ergebnis mit einem nichtparametrischem Mittelwerttest, z.B. dem Mann-Whitney-Test (auch: U-Test), da wir nicht davon ausgehen können, daß die Grundgesamtheiten, aus der die beiden Stichproben stammen, normalverteilt sind. Der U-Test überprüft die Hypothese H_1, daß die Wahrscheinlichkeit, mit der ein Element, das zufällig aus der Stichprobe *pros* gezogen wird, größer ist als ein Element, das zufällig aus der Stichprobe *wiss* gezogen wird, größer ist als 50 %. Ein Element x_{pros} der Stichprobe *pros* ist stochastisch größer als ein Element x_{wiss} der Stichprobe *wiss* (vgl. Siegel (1987:113): $H_1 = (x_{pros} > x_{wiss}) > 0{,}5$.

Die Alternativhypothese H_0 lautet demnach: $H_0 = (x_{pros} > x_{wiss}) \leq 0{,}5$.

[20] Der „Ausreißer" bei Polysemieklasse 13 kommt durch *be* zustande, welches recht häufig ist, aber auch eine relativ hohe Polylexie besitzt.

Mithilfe der Statistik-Software SPSS® wird das in Tabelle 7 und Tabelle 8 dargestellte Ergebnis erzielt.

Tabelle 7
Rangdaten der semantischen Spektren

	genre	*N*	*Mittlerer Rang*	*Rangsumme*
PL	pros	36570	43040,16	1573978496,00
	wiss	44022	38017,29	1673596928,00
	Gesamt	80592		

Tabelle 8
Ergebnis des Mann-Whitney-Tests für semantische Spektren (Gruppenvariable: genre)

Polylexie	
Mann-Whitney-U	704606720,000
Wilcoxon-W	1673596928,000
Z	-30,641
Asymptotische Signifikanz (2-seitig)	0,000

Das Testergebnis ist signifikant, was bedeutet, daß die H_1 momentan nicht verworfen werden muß. Da es sich im vorliegenden Fall um einen sehr großen Stichprobenumfang handelt, liegt die Vermutung nahe, daß dieser ein signifikantes Ergebnis verhindert. In einem kleinen Permutationstest, der aufgrund des großen Stichprobenumfangs nur mit 38 Zufallsstichproben durchgeführt wurde, wurden die Meßwerte zufällig den Genres zugeordnet und einem Test unterzogen. Von 38 Zufallsstichproben waren 36 nicht signifikant, die H_0 konnte in diesen Fällen nicht verworfen werden.[21]

Ausblick

Die Untersuchung der semantisch annotierten Teile des Brown Corpus hat zum ersten Mal eine Vorstellung davon vermittelt, welcher Verteilung Lesarten in Texten folgen könnten und die Analogiehypothese zur Wortverteilung, die Zipf-Mandelbrot oder Waring als Verteilungsform annimmt, gestützt. Die Explikation des zugrunde liegenden Mechanismus steht noch aus. In zukünftigen Untersuchungen an möglicherweise anderen Sprachen sollte diese Fragestel-

[21] Herzlicher Dank für methodische Hilfe geht an Bernhard Baltes-Götz.

lung weiter verfolgt werden. Besonders interessant ist auch die Hypothese, daß verschiedene Textsorten unterschiedliche semantische Spektren aufweisen, die hier einer ersten Untersuchung standhielt. Ebenso stellt sich für zukünftige Projekte die Frage nach dem Zusammenhang zwischen Lexikon- und Textpolylexie, der in einer Verteilung der Unterschiede zwischen beiden Operationalisierungen differenzierter betrachtet werden müßte. Auch die Möglichkeiten des Untersuchungdesigns „im Lexikon, mit Texttypes oder mit Texttoken" sollten weiter exploriert und interpretiert werden.

Während bislang Diversifikationsuntersuchungen an vereinzelten, zumeist grammatischen Wörtern unter differenzierter Betrachtung ihrer Polysemie, die über die Auflösung des Lexikons hinausging, durchgeführt wurden, bieten die vorliegenden Daten die Möglichkeit, Diversifikationsuntersuchungen im „großen Stil" und auch an Autosemantika durchzuführen.

Literatur

Altmann, G. (1985) Semantische Diversifikation. *Folia linguistica*, 19,177-200.

Altmann, G., Best, K.-H., & Kind, B. (1987). Eine Verallgemeinerung des Gesetzes der semantischen Diversifikation. In I. Fickermann (Hrsg.), *Glottometrika 8* (S. 130-139), Bochum: Brockmeyer.

Altmann, G., Beöthy, E., & Best, K.-H. (1982). Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 35, 537-543.

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

Bréal, M. (1924). *Essai de sémantique. Science des significations.* Genève: Slatkine. Repr., 1976. Reprint der Ausgabe Paris 1924.

Croft, W. (1990). *Typology and Universals.* Cambridge: University Press. (Cambridge textbooks in linguistics)

Fellbaum, Chr. (Hrsg.)(1998). *Wordnet: an electronical Database.* Cambridge, Mass. u.a.: MIT Press. (Language, speech, and communication)

Fickermann, I., Markner-Jäger, B., & Rothe, U. (1984). Wortlänge und Bedeutungskomplexität. In J. Boy & R. Köhler (Hrsg.), *Glottometrika 6*, (S. 115-126), Bochum: Brockmeyer. (Quantitative linguistics ; 25)

Gieseking, K. (1993). *Synergetische Aspekte von Struktur und Dynamik der englischen Lexik.* Unveröff. Magisterarbeit, Universität Trier.

Gieseking, K. Untersuchungen zur Synergetik der englischen Lexik. In R. Köhler (Hrsg.), *Beiträge zur quantitativen und systemtheoretischen Linguistik* (erscheint).

Guiter, H. (1977). Les relations /fréquence-longueur-sens/ des mots (langues romanes et Anglais) In *Actes du 14ème congrès international de linguistique et philologie romanes.* Napoli 15-20 Aprile 1974. 1. Aufl. Bd. 4, 1977-373-381

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

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer. (Quantitative linguistics; 13)

Köhler, R. (1984). Zur Interpretation des Menzerathschen Gesetzes. In J. Boy & R. Köhler (Hrsg.), *Glottometrika 6* (S. 177-183), Bochum: Brockmeyer. (Quantitative linguistics; 25)

Krylov, Ju.K. (1982). Eine Untersuchung statistischer Gesetzmäßigkeiten auf der paradigmatischen Ebene der Lexik natürlicher Sprachen. In H. Guiter & M.V. Arapov (Hrsg.), *Studies on Zipf's Law* (S. 234-262), Bochum: Brockmeyer. (Quantitative linguistics; 20)

Landes, Sh., Leacock, C., & Tengi, R.I. (1998), Building semantic Concordances. In Fellbaum, 1998:199-216

Levickij, V.V. Polysemie. In G. Altmann, R. Köhler & R.G. Piotrowski (Hrsg.), *Quantitative Linguistik. Ein internationales Handbuch - Quantitative Linguistics. An International Handbook,* Berlin u.a.: de Gruyter (erscheint).

Leopold, E. (1998). *Stochastische Modellierung lexikalischer Evolutionsprozesse.* Hamburg: Kovac. (Schriftenreihe Philologia; 30). Zugl.: Trier, Univ., Diss., 1998

Rothe, U. (1983). Wortlänge und Bedeutungsmenge: eine Untersuchung zum Menzerathschen Gesetz an drei romanischen Sprachen. In R. Köhler & J. Boy (Hrsg.), *Glottometrika 5* (S. 101-112), Bochum: Brockmeyer. (Quantitative linguistics; 20)

Rothe, U. (Hrsg.) (1991). *Diversification Processes in Language: Grammar.* Hagen: Rottmann.

Siegel, S. (1987). *Nicht-parametrische statistische Methoden.* Eschborn. (Methoden in der Psychologie; 4)

Tuldava, J., Altmann, G., u.a. (Übers.) (1998). *Probleme und Methoden der quantitativ-systemischen Lexikologie.* Trier: Wissenschaftlicher Verlag Trier. (Quantitative linguistics; 59)

Zipf, G.K. (1972). *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology.* Reprint. New York: Hafner. [Faks. d. Ausg. von 1949]

Zipf, G.K. (1945). The Meaning-Frequency Relationship of Words. *Journal of General Psychology*, 33, 251-256.

Wort- und Satzlänge als parallele stilistische Parameter

Kontexte der französischen Demonstrativpronomina *celui-ci* und *celui-là*

Marc Hug

Die französischen Demonstrativpronomina *celui-ci* und *celui-là*, mit ihren morphologischen Varianten *celle-ci*, *celle-là ceux-ci*, *ceux-là*, *celles-ci*, *celles-là* werden von den Grammatiken, ebenso von den neueren, linguistisch inspirierten wie von den traditionellsten, meistens so beschrieben, als würden sie vor allem paarweise gebraucht und als sei jeweils eine in *-ci* endende Form mit einer in *-là* endenden Form verbunden; in solch einem Gebrauch geben die beiden Endungen dem jeweiligen Formpaar ungefähr den Gegensatzwert der deutschen Pronomina *dieser* vs. *jener*. Solch eine Darstellung bedeutet nicht unbedingt, daß die Grammatiker diese kontrastive Verwendung als die verbreitetste betrachten, sondern vielmehr, daß die kontrastive Darstellung den jeweiligen Stellenwert beider Demonstrativa am besten zu verstehen gibt. Beim Leser kann aber dennoch der Eindruck entstehen, es handle sich hier um die „wichtigste" Verwendung dieser Einheiten.

Nun findet man aber in den gleichen Grammatiken auch die Meinung, die in *ci* endenden Formen werden immer seltener, und die heutige Sprache gebe den là-Formen ganz deutlich den Vorzug. Diese letztere Meinung wird von einem umfangreichen Korpus aus der heutigen frz. Schriftsprache eindeutig widerlegt, wie ich vor kurzem zeigen konnte (Hug, 1998). Außerdem ist sie aber mit der anderen, ebenfalls traditionellen Auffassung im Widerspruch, die meint, beide Sorten Pronomina werden vor allem im kontrastiven Gebrauch verwendet. Wenn nämlich ein *celui-là* vor allem dort zu erwarten wäre, wo es im Gegensatz zu einem *celui-ci* steht, könnte eine der Formen der anderen nicht vorgezogen werden, da sie beide zusammen vorkommen müssten.

Zwei andere, oft auftretende Ideen sind folgende: einmal, daß *celui-ci* eine nahe Wirklichkeit (ob im Kontext oder in der Sprechsituation), *celui-là* eine entferntere bezeichnet; und zum anderen, daß *celui-ci* einen kataphorischen

Wert haben kann, während *celui-là* ausschließlich anaphorisch interpretiert werden kann.

All diese Ideen beruhen auf teilweise gerechtfertigten Intuitionen, sind aber als Allgemeinregeln nicht vertretbar. Es soll hier gezeigt werden, daß eine in den Grammatiken meist eher angedeutete als ausgesprochene Idee einen wesentlichen Teil des semantischen und stilistischen Werts dieser Demonstrativa audrückt, nämlich die Idee, die besagt, *celui-ci* gehöre zu einem gehobeneren, anspruchsvolleren oder jedenfalls formelleren Sprachniveau als *celui-là*. Dies muß allerdings zur Folge haben, daß entweder beide Demonstrativa in einem semantisch ähnlichen Kontext mit gleichem Sinn zur Verfügung stehen können, oder daß in Sprechsituationen, die den Gebrauch von *celui-ci* erlauben, desto weniger Gelegenheiten auftreten, *celui-là* zu gebrauchen.

1. Das Korpus

In der großen Datenbasis *Frantext*[1] wurden zuerst alle Texte ausgesondert, die im Jahre 1960 und seither erschienen sind. In diesen Texten kommen im Ganzen *celui-ci* und seine Varianten 3791mal vor, während *celui-là* nur 955mal auftritt. Aus diesem Korpus wurde vor wenigen Jahren eine Stichprobe von 724 unter den damals 3620 verfügbaren Verwendungen von *celui-ci* erhoben, während sämtliche 800 von *celui-là,* die damals vorhanden waren, behalten wurden. Jedes dieser 1524 Pronomina wurde in einem Kontext von insgesamt drei „Sätzen" eingetragen, nämlich dem Satz, in dem das Pronomen vorkommt, dem Satz davor und dem Satz nachher. Nun wurde aber die Einheit „Satz" von dem Befragungsprogramm von *Frantext* etwas anders definiert als in meinen eigenen Programmen, was dazu führt, daß die Anzahl der Sätze nicht genau dreimal die Zahl der Pronomina betrifft, sondern bedeutend mehr[2].

Im Folgenden wird gelegentlich von „beiden Korpora" die Rede sein, womit in einer etwas nachlässigen, aber bequemen Terminologie die beiden Stichproben, d.h. die 724 Verwendungen von *celui-ci* bzw. die 800 Verwendungen von *celui-là* in ihrem jeweiligen Kontext gemeint sind.

Es kann intuitiv von vorneherein als gewiß erscheinen, daß Wort- oder Satzlänge in den beiden Korpora nicht gleich sind, denn die Kontexte wurden in beiden Fällen mit genau den gleichen Kriterien abgegrenzt, und trotzdem (s. Tabellen 1 und 2) ist das Korpus mit 724 Kontexten von *celui-ci* nicht unerheblich länger als das mit den 800 Kontexten von *celui-là*. Auch fällt bei der Be-

[1] *Frantext*, Institut National de la langue française, 44, avenue de la Libération, F 54000 Nancy. Das Abonnement wird an der Straßburger Universität II (Université des Sciences humaines) vom *Centre de Calcul appliqué aux Sciences humaines* getragen.

[2] Z.B. wird in *Frantext* der Doppelpunkt nicht als Satzgrenze betrachtet, während er hier als solche gilt.

trachtung der Tabellen 2 auf, daß im *celui-ci*-Korpus weniger Sätze, aber mehr Wörter sind als im *celui-là*-Korpus. Wer mit statistischen Messungen vertraut ist, kann daher im voraus raten, daß der Vergleich dieser beiden Eintragungsreihen zu signifikanten Resultaten führen wird.

Tabelle 1
Wortlänge

A: 724 Verwendungen von *Celui-ci*

x_i	n_i	n_ix_i	$n_ix_i^2$
1	8094	8094	8094
2	18642	37284	74568
3	12960	38880	116640
4	10556	42224	168896
5	6932	34660	173300
6	6525	39150	234900
7	6045	42315	296205
8	5281	42248	337984
9	4327	38943	350487
10	3371	33710	337100
11	1946	21406	235466
12	1375	16500	198000
13	798	10374	134862
14	456	6384	89376
15	230	3450	51750
16	81	1296	20736
17	23	391	6647
18	22	396	7128
19	7	133	2527
∑	87671	417838	2844666

Mittlere Wortlänge: 4.766
Varianz: 9.733
Std-Abweichung: 3.120

B: 800 Verwendungen von *Celui-là*

x_i	n_i	n_ix_i	$n_ix_i^2$
1	7617	7617	7617
2	17479	34958	69916
3	10945	32835	98505
4	11264	45056	180224
5	7301	36505	182525
6	6044	36264	217584
7	4910	34370	240590
8	3869	30952	247616
9	2592	23328	209952
10	1740	17400	174000
11	912	10032	110352
12	514	6168	74016
13	271	3523	45799
14	133	1862	26068
15	65	975	14625
16	26	416	6656
17	8	136	2312
18	2	36	648
∑	75692	322433	1909005

Mittlere Wortlänge: 4.260
Varianz: 7.075
Std-Abweichung: 2.660

In Tabelle 2 werden die n_ix_i und $n_ix_i^2$ nicht mehr abgedruckt. Sie können aber anhand der beiden anderen Spalten nachgerechnet werden, und dadurch können auch Mittelwert und Varianz geprüft werden.

Tabelle 2
Satzlänge

A: Celui-ci

x_i	n_i	x_i	n_i	x_i	n_i
1	249	40	33	81	1
2	480	41	31	83	1
3	565	42	27	84	1
4	178	43	30	86	1
5	266	44	27	88	1
6	191	45	25	93	1
7	156	46	16	95	1
8	183	47	20	97	1
9	91	48	19	98	2
10	98	49	26	99	1
11	123	50	26	100	1
12	93	51	14	109	1
13	88	52	13	111	1
14	100	53	16	123	1
15	82	54	17	143	1
16	87	55	5	242	1
17	92	56	3		
18	91	57	6		
19	98	58	14		
20	95	59	7		
21	100	60	8		
22	96	61	4		
23	96	62	4		
24	95	63	6		
25	72	64	8		
26	86	65	6		
27	72	66	6		
28	91	67	5		
29	76	68	7		
30	78	69	1		
31	64	70	2		
32	43	71	3		
33	57	72	4		
34	55	73	3		
35	61	75	1		
36	39	76	2		
37	40	77	1		
38	45	78	1		
39	46	80	3		

∑:5198

Mittlere Satzlänge (Zahl der Wörter): 16.866
Varianz: 260.958
Std-Abweichung: 16.154

B: *Celui-là*

x_i	n_i	x_i	n_i	x_i	n_i
1	262	40	17	83	1
2	255	41	11	84	3
3	897	42	21	88	1
4	344	43	12	90	1
5	435	44	13	91	1
6	347	45	13	93	2
7	287	46	14	102	1
8	258	47	17	107	1
9	216	48	11	109	1
10	171	49	7	112	1
11	148	50	9	114	1
12	162	51	4	139	1
13	160	52	14	141	1
14	134	53	4	145	1
15	115	54	7	146	1
16	108	55	12	161	1
17	117	56	7	172	1
18	92	57	8	181	1
19	101	58	5	208	1
20	93	59	12	232	2
21	85	60	5	281	1
22	78	61	6		
23	72	62	6		
24	71	63	2		
25	52	64	2		
26	62	65	2		
27	49	66	2		
29	42	68	2		
30	43	69	3		
31	38	70	1		
32	33	71	1		
33	35	72	1		
34	23	74	3		
35	26	75	3		
36	24	76	3		
37	39	77	4		
38	25	79	2		
39	22	82	1		
$\sum$: 5843					
Mittlere Satzlänge (Zahl der Wörter): 12.954 Varianz: 227.505 Std-Abweichung: 15.083					

x_i = Satzlänge (Wortzahl)
n_i = Zahl der Sätze mit x_i Wörtern

2. Die mittlere Wortlänge

Es wurde davon ausgegangen, daß in einer „einfachen", anspruchslosen Ausdrucksweise die mittlere Wortlänge kürzer ausfallen sollte als dort, wo die Sprache gehobener sein muß. Wenn wir also annehmen, daß die *ci*-Formen zu einer gehobeneren Sprache gehören als die *là*-Formen, so muß die Wortlänge im Durchschnitt im Kontext der *ci*-Formen höher sein, als im Kontext von *là*-Formen. Tatsächlich finden wir die folgenden Resultate:

$\bar{x}_{ci} = 4.766 \quad Var\, x_{ci} = 9.733 \quad n_{ci} = 87671$
$\bar{x}_{ci} = 4.260 \quad Var\, x_{là} = 7.075 \quad n_{là} = 75692$

(wo jeweils $\bar{x}$ einen Mittelwert bedeutet, *Var x* die dazu gehörige Varianz und *n* die Gesamtzahl der Wörter in der jeweiligen Stichprobe). Kann dieser Unterschied als signifikant angesehen werden? Da die Anzahl der Wörter, die diese Mittelwerte ergeben, Zehntausende betragen, ist die Variation des Mittelwertes zweifellos einer Gaußschen Verteilung gemäß. Die Varianz wurde gleichzeitig mit dem Mittelwert errechnet[3]. Um die beiden Mittelwerte miteinander zu vergleichen, wird davon ausgegangen, daß sie Zufallswerte einer und derselben Variablen sind; wenn diese Hypothese zutrifft, muß der Unterschied $\bar{x}_{ci} - \bar{x}_{là}$ symmetrisch um den Nullpunkt verteilt sein, und zwar nach einer Gaußschen Verteilung $N(0; \sigma 2)$,

$$\sigma^2 = \frac{Var\, x_{ci}}{n_{ci}} + \frac{Var\, x_{là}}{n_{là}} = \frac{9.733}{87671} + \frac{7.075}{75692} = 0.000204488;$$

die Varianz der jeweiligen Wortlänge wird durch die Zahl der vorkommenden Wörter geteilt; so wird die Varianz der mittleren Wortlänge in jedem den beiden Textsets errechnet.

Der normierte Unterschied zwischen den beiden Mittelwerten wird mit *u* bezeichnet.

$$u = \frac{\bar{x}_{ci} - \bar{x}_{là}}{\sigma} = \frac{4.766 - 4.260}{\sqrt{0.000204488}} = 35.38$$

[3] Es könnte auch ein Vergleich der beiden Varianzen vorgenommen werden, wobei die Varianz der Wörter um *celui-ci* bedeutend größer als die der Wörter um *celui-là* ausfallen würde. Dies kommt daher, daß der Großteil der sehr kurzen Wörter aus unentbehrlichen grammatischen Wörtern besteht. Dieser Unterschied der Varianzen ist also eine notwendige Folge des Unterschieds der Mittelwerte.

Solch eine Abweichung von Mittelwert hat nach dem Gaußschen Gesetz so gut wie gar keine Aussicht, durch reinen Zufall einzutreten. Es kann praktisch mit Sicherheit behauptet werden, daß die durchschnittliche Wortlänge im Kontext der Verwendung von *celui-ci* bedeutend größer ist als im Kontext von *celui-là*.

3. Die mittlere Satzlänge

Hier wurde als Maßstab nicht die Zahl der Buchstaben, sondern die Zahl der Wörter genommen, die jeden Satz bilden. Würde nämlich die Buchstabenzahl gemessen, so wäre gleich selbstverständlich, daß bei gleicher Wortzahl die Sätze um *celui-ci* länger wären als die Sätze um *celui-là*. Wenn wir hingegen von der Wortzahl ausgehen, ist diese Messung von der vorigen unabhängig.

Allerdings ist hier die Variation viel größer als bei der Wortlänge; ein Satz wurde so definiert, daß jede „starke" Interpunktion als Satzgrenze angesehen wurde[4]. Die Sätze haben in den verschiedenen Texten Längen, die sich von 1 bis zu 281 Wörtern erstrecken. Hier ist die Satzlänge noch weiter davon entfernt, einer Normalverteilung zu folgen, als dies bei der Wortlänge der Fall war. Aber auch hier bleibt gültig, daß die *durchschnittliche* Länge eines jeden der beiden Korpora einer Normalverteilung gemäß variiert, weil die Anzahl der betroffenen Sätze sehr groß ist.

Wir bezeichnen hier mit y_{ci} und $y_{là}$ die Variablen „Satzlänge im Kontext von resp. *celui-ci* und *celui-là*". Diesmal interessiert uns der Unterschied $\bar{y}_{ci} - \bar{y}_{là}$.

Die dazu gehörige Varianz wird wie im vorigen Abschnitt ausgerechnet, nur mit verschiedenen Parameterwerten.

$$\sigma^2_{y(ci)} = \frac{260.958}{5198} = 0.0502, \qquad \text{und } \sigma^2_{y(là)} = \frac{227.505}{5843} = 0.0389.$$

Der genormte Unterschied beträgt daher

$$u = \frac{16.866 - 12.954}{\sqrt{0.0502 + 0.0389}} = \frac{3.912}{0.29856} = 13.10.$$

Wenngleich diese Abweichung vom theoretischen Mittelwert 0 nicht so enorm wie bei der Wortlänge ausfällt, so übersteigt sie doch bei weitem alle Werte, die auf den Tabellen des Gaußschen Gesetzes vermerkt sind. Es liegt also

[4] Dies ist kein sehr zuverlässiges Merkmal, kann aber deswegen hier verwendet werden, weil die Stellen, wo dadurch falsche, sehr kurze Pseudo-„Sätze" identifiziert werden, sich vor allem in dem Korpus befinden, das ansonsten die längsten Sätze hat. Eine Kontrolle in den betreffenden Texten hat gezeigt, daß diese Stellen selten sind, und daß sie größtenteils aus Ziffern bestehen, die, von einem Punkt gefolgt, einen neuen Abschnitt in einem Kapitel einleiten.

auf der Hand, daß die Sätze, in denen *celui-ci* auftritt, in der Regel länger sind als diejenigen, in denen *celui-là* verwendet wird.

4. Interpretation dieser Resultate

Jede Texteinheit[5] in *Frantext* wird durch einen von einer dreistelligen Zahl gefolgten alphabetischen Charakter identifiziert. Dabei trägt der alphabetische Charakter dazu bei, die Texteinheiten zwischen verschiedenen Kategorien zu sortieren; so sind die durch **R***nnn* identifizierten Texteinheiten meist Romane, jedenfalls grundsätzlich fast immer narrative Prosa, während Texteinheiten, deren Identifizierung z.B. mit **P** beginnt, nicht-literarische Texte sind, wie etwa *Encyclopédie de l'éducation en France, L'Univers économique et social*, Bourbakis *Éléments d'histoire des mathématiques* usw. Nun sagt uns Tabelle 3, wie sich die verschiedenen Verwendungen unserer Demonstrativa nach diesen Identifizierungen verteilen. Es ist sogleich zu sehen, daß gut drei Viertel aller *celui-là* in R-Texteinheiten auftreten, aber in den gleichen Texten nur etwa 11% aller *celui-ci*. Diese Feststellung paßt sehr gut zu dem, was durch statistische Tests schon festgestellt worden ist[6].

Tabelle 3

	-ci	*-là*
L	44	29
P	595	151
Q	6	2
R	79	618

N.B. S. Text (§ 4). Die neben R stehenden Pronomina kommen in Romanen vor, die anderen in sonstigen Texten.

Ein Roman enthält oft Dialoge, und kann auch ansonsten eventuell in einer ziemlich ungezwungenen Sprache geschrieben sein. Natürlich muß das nicht so sein; auch muß ja ein Satz in einem Roman nicht kurz sein: wir sehen in Tabelle 2, daß *celui-là* 15mal in wenigstens 100 Wörter langen Sätzen vorkommt, *celui-ci* aber nur 6mal. Aber die Sprache eines Romans kann einfach und familiär sein.

[5] Was in *Frantext* als Texteinheit (*unité textuelle*) bezeichnet wird, ist entweder ein ganzes Werk – Roman, Lehrbuch, Gedichtesammlung oder dgl. – oder ein Teil eines sehr ausgedehnten Werkes. Im Durchschnitt hat eine Texteinheit eine Länge von etwa 200 bis 300 Seiten.

[6] Es wäre natürlich möglich, auch über die Daten in Tabelle 3 durch einen χ^2-Test zu bestätigen, daß sich die Verteilung der beiden Demonstrativa nicht gleichmäßig zwischen die verschiedenen Identifikatoren verteilt, aber diese Ungleichmäßigkeit ist so augenfällig, daß sich solch ein Test erübrigt.

Eine wissenschaftliche Schrift und ein didaktisches Werk können nicht die gleiche Sprache sprechen: hier ist der Stil einheitlicher, was sich dadurch sichtbar macht, daß ebenso die längsten wie auch die kürzesten Sätze weniger zahlreich sind als in den Romanen. In den *celui-là*-Kontexten machen die sehr kurzen Sätze (1 bis 10 Wörter) zwei Drittel aller Sätze aus, während es um *celui-ci* nur weniger als die Hälfte sind; gleichzeitig, wie eben erwähnt, sind die meisten sehr langen Sätze auch in den *celui-là*-Kontexten zu finden.

Celui-là ist in der Umgangssprache so geläufig, daß viele Grammatiken meinen, *celui-ci* sei wie die anderen mit *-ci* endenden Demonstrativa relativ selten, wo es doch in den neueren Texten in Frantext etwa viermal öfter vorkommt als *celui-là*. Das kommt daher, daß in wissenschaftlich-didaktischen Schriften, wie sie in *Frantext* für die Zeitspanne 1960-1975 recht zahlreich sind[7], *celui-ci* noch viel stärker vorherrscht als im gesamten Korpus. Es ist daher gar nicht erstaunlich, daß in den Schriften, die am meisten *celui-ci* verwenden, die Wortlänge größer ist als in den andern Texten, denn einerseits geht die moderne Wissenschaft bekanntlich gern mit einer etwas schwerfälligen Terminologie um, während diese Termini in der Umgangssprache, sofern sie in dieselbe aufgenommen worden sind, oft abgekürzt werden, was ein Anzeichen dafür ist, daß sie für den normalen Sprachgebrauch zu lang erscheinen. Andererseits ist naturgemäß eine wissenschaftliche Ausführung auch in logischer Hinsicht meist komplizierter als die Themen eines familiären Gesprächs; daher erklärt sich, daß auch die Satzlänge in den gleichen Schriften größer ist; nicht, daß die Umgangssprache keine oder wenige untergeordnete Sätze gebrauche; aber diese Sätze sind doch meist so strukturiert, daß die gesamte syntaktische Struktur leicht übersichtlich bleibt.

Man darf also annehmen, daß die Frequenz der beiden Pronomina *celui-ci* und *celui-là* zu einem erheblichen Teil von einem stilistischen Faktor abhängig ist. Wahrscheinlich ist die rechte Frage nicht „Wie geschieht die Wahl zwischen *celui-ci* und *celui-là*?", sondern eher diese: „Wenn man von den syntaktischen und semantisch-pragmatischen Stellenwerten der beiden Pronomina ausgeht, in welchen Texten hat jedes der beiden die größten Aussichten, vorkommen zu können?" Man muß nämlich hier zu unserem Ausgangspunkt zurückkehren: der Normalfall ist nicht der, daß die beiden miteinander gebraucht werden; wir fügen nun hinzu: er ist auch nicht, daß die beiden miteinander in Konkurrenz stehen (obgleich dies vorkommen kann). Jedes hat seinen Bereich, beide Bereiche überschneiden sich nur teilweise, und die Kontexte, in denen jedes vorkommen kann, sind großenteils verschieden.

[7] Leider ist das Verhältnis zwischen der Anzahl literarischer und nicht-literarischer Texte in *Frantext* gar nicht stabil geblieben. Im ganzen genommen überwiegen die literarischen Texte stark, aber für diese Periode sind eine große Anzahl von Lehrbüchern in die Datenbasis eingegangen. Es wäre zu wünschen, daß dies auch für andere Perioden der Fall wäre.

Von diesen Kontexten soll hier nur andeutungsweise die Rede sein. Unter den Verwendungen von *celui-ci*, die hier untersucht wurden, sind 686 (fast 95%) einfache 724 Anaphoren, d.h. das Pronomen repräsentiert eine im Kontext vorhandene Nominalphrase und wiederholt deren ganzen Inhalt. So etwa in diesem einfachen Beispiel :

... une véritable mobilisation de l'industrie privée française. <Celle-ci> emploie ...

wo die Phrase *l'industrie privée française* durch *celle-ci* wieder aufgegriffen wird. Diese Art der Verwendung kommt bei *celui-là* nur in 128 von 800 Verwendungen vor (also 16%). Es muß dabei unterstrichen werden, daß die Anwendung solch einer Anaphora nur in relativ komplexen Texten notwendig ist, da in einer einfacheren Rede, wo nur *ein* Beziehungswort in Frage kommt, meist ein Personalpronomen ausreicht. Daneben findet man aber bei *celui-là* 95 teilweise Anaphoren, in denen das Pronomen gleichzeitig von zwei verschiedenen Kontextstellen abhängt: zuerst von einer notionellen Verankerung, und dann von einer referentiellen. Betrachten wir z.B. folgendes Zitat:

Dans certains domaines je suis très susceptible, maladivement susceptible, dit-elle. Il y a beaucoup de défauts dont je puis sans doute me corriger, mais de <celui-là> je ne crois pas. (M. Droit)

Celui-là stützt sich notionell auf den Begriff *défauts* („Fehler"), der eben vorher genannt wurde; referentiell aber deutet *celui-là* auf das vorher erwähnte *très susceptible* („empfindlich"), das eben einen „Fehler" darstellt. Das Beziehungswort *défauts* ist im Plural, das Pronomen aber bezeichnet einen besonderen „Fehler", der aus dem sonstigen Kontext zu ermitteln ist, und ist deshalb im Singular. So eine Verwendung des Demonstrativpronomens kommt mit *celui-là* aber, wie gesagt, 95mal, mit *celui-ci* jedoch nur 12mal vor.

Außerdem gibt es auch in unserem Korpus 40 Verwendungen der Sequenz *celui-là même* (qui) = „eben derselbe, der", während *celui-ci même* (qui) überhaupt nicht vorkommt.

Diese kurzen Feststellungen genügen, um zu zeigen, daß beide Pronomina zu einem großen Teil getrennte Verwendungsmöglichkeiten haben, und daß die in den Grammatiken übliche kontrastive Darstellung ein ziemlich falsches Bild von ihren normalen Verwendungen gibt. Die Tatsache, daß die am häufigsten vorkommende Verwendung von *celui-ci* von Natur aus an komplexere Kontexte gebunden ist, kann zum Teil die Unterschiede erklären, die wir zwischen Wort- und Satzlängen um beide Pronomina feststellen konnten.

Implementation Aspects and Applications of a Spelling Correction Algorithm

Viggo Kann, Rickard Domeij, Joachim Hollman, Mikael Tillenius

Introduction

How to automatically detect and correct spelling errors is an old problem. Nowadays, most word processors include some sort of spelling error detection. The traditional way of detecting spelling errors is to use a word list, which usually also contains some grammatical information, and to look up every word in the text in the word list (Kukich, 1992).

The main problem with this solution is that if the word list is not large enough, the algorithm will report several correct words as misspelled, because they are not included in the word list. For many natural languages, the size of word list needed is too large to fit in the working memory of a simple computer. In Swedish this is a large problem, because infinitely many new words can be constructed as compound words.

There is a way to reduce the size of the stored word list by using *Bloom filters* (Bloom, 1970). Then the word list is stored as an array of bits (zeroes and ones), and only two operations are allowed: checking if a specific word is in the word list and adding a new word to the word list. Both operations are extremely fast and the size of the stored data is greatly reduced.

We have developed a method for finding and correcting misspellings in Swedish texts using Bloom filters (Domeij, Hollman and Kann, 1994). In this paper we describe the method and our implementation of it, called Stava (Kann, 1998). We concentrate in particular on how the spelling error correction suggestions are ranked using quantitative linguistic methods. We also mention several applications of our methods, besides spelling error detection and correction. Examples of applications that we have studied are extension and creation of a part-of-speech lexicon, tagging of unknown words, hyphenation of one word compounds, lemmatisation and correction of search questions in information retrieval.

Preliminaries

Swedish word formation

Swedish is a morphologically rich language compared to English. An ordinary verb in Swedish has more than ten different inflectional forms. Most words can also be compounded to form a completely new word. For example, the verb *rulla* (roll) can combine with *skridsko* (skate) to form the word *rullskridsko* (roller skate). Since words can combine without limit, it is not possible to list them. This is a considerable problem for Swedish spell checkers. The many false alarms that make Swedish spell checkers impractical are caused by compound words.

As the example of Swedish compounding above shows, it is not always possible just to put two words together to form a compound. Stem alteration is often the case, which can mean that the last letter of the initial word stem is deleted or changed, depending (roughly) on what part of speech and inflectional group it belongs to. Between different compound parts an extra *-s-* is often added. However, individual words tend to behave irregularly, thus making compounding hard to describe by general rules.

Bloom filters

A Bloom filter (Bloom, 1970) is a special kind of hash table, where each entry is either 0 or 1, and where we make repeated hashings into a single table (using different hash functions each time). A word is added to the table by applying each hash function to the word and entering ones in the corresponding positions (i.e., the integer indices that the hash functions return). To check if a word belongs to the word list, you apply the same hash functions and check if all the entries are equal to 1. If not all entries are equal to 1, then the word was not in the word list.

It can happen that a word gets accepted even if it is not in the word list. The reason is that all the positions of the word may be set to 1 due to collisions with other words. Fortunately, the probability of such collisions can easily be adjusted to a specific application. All we have to do is to change the size of the table and the number of hash functions.

Suppose that the word list consists of n words, that the size of the hash table is m, and that we use k independent and evenly distributed hash functions. Then the probability that a word not in the word list will be accepted by the Bloom filter is $f(k)=[1-(1-1/m)^{kn}]^k$, as shown in Domeij, Hollman and Kann (1994). The minimum of this function is $f(k)=2^{-k}$, which is reached when $k = \ln 2 \cdot m/n$.

Example 1:

If the word list has n = 100,000 words and we choose m = 2,000,000 as the size of the hash table, we should choose $k = \ln 2 \cdot 2{,}000{,}000/100{,}000 \approx 14$, i.e., we should use 14 hash functions in the Bloom filter. The probability that a random word is accepted is $f(14) \approx 6 \cdot 10^{-5} = 0.006\%$.

Compounding and inflection

Basic structure of the word recognition

In Stava, compounding and inflection are handled by an algorithm that uses a list of suffix rules together with three different word lists: the *individual word list*, containing words that cannot be part of a compound at all, the *last part list*, containing words that can end a compound or be an independent word, the *first part list*, containing altered word stems that can form the first or middle part of a compound.

Inflection is handled in a straightforward but unconventional way. We are trying a heuristical method to reduce the number of word forms listed, and ensure that all forms of a word are represented. The *last part list* presented above does not actually contain all inflectional word forms. It contains only the basic word forms needed to infer the existence of the rest from suffix rules.

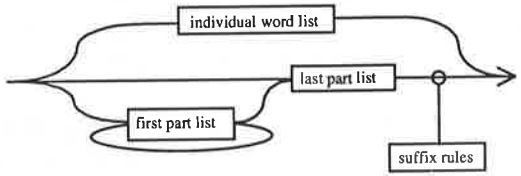


Fig. 1: Look-up scheme for handling of compounding and inflection.

When a word is checked, the algorithm consults the lists in the order illustrated in Figure 1. In the trivial case, the input word is found directly in the *individual word list* or the *last part list*. If the input word is a compound, only its last part is confirmed in the *last part list*. Then the *first part list* is looked up to find its first part. If the compound has more than two parts, a recursive look-up is performed. The algorithm optionally inserts an extra *-s-* between compound parts, to account for the fact that an extra *-s-* is generally inserted between the second and third compound parts.

The ending rule component is only consulted if the input word can be found neither in the *individual word list*, nor in the *last part list*. If the last part of the input word matches a rule ending, it is considered a legal ending under the condition that the related basic inflectional forms are in the *last part list*. In this way, only three noun forms, normally out of eight, must be stored in the *last part list*. The other noun forms are inferred by suffix rules.

Example 2

The word *docka* (doll) belongs to the first inflectional noun class in Swedish, and has the following inflectional forms: *docka* (doll), *dockan* (the doll), *dockor* (dolls), *dockorna* (the dolls), *dockas* (doll's), *dockans* (the doll's), *dockors* (dolls'), *dockornas* (the dolls').

For this ending class only *docka*, *dockan* and *dockor* are put in the last part list. We construct the following suffix rules from which the other five forms can be inferred:

-orna	←	*-a, -an, -or*
-as	←	*-a, -an, -or*
-ans	←	*-a, -an, -or*
-ors	←	*-a, -an, -or*
-ornas	←	*-a, -an, -or*

Consider the input word *porslinsdockorna* (*porslin* = porcelain). The input word can neither be found in the *individual word list*, nor in the *last part list*. Therefore the suffix rules are consulted. The first rule above should be read: If *dock-a*, *dock-an* and *dock-or* are in the *last part list*, then the word *dock-orna* is a legal word.

Finally the *first part list* is consulted. There the first part of the compound (*porslins-*) is found, thus confirming the legality of the input word.

Improved expressiveness of suffix rules

Our handling of inflections is a possible source of error. For example, the non-existent word *dekorna* can be constructed using the rule above since the words *deka* (degenerate), *dekan* (dean) and *dekor* (decor) all exist in Swedish. It is important to design the rules in such a way that the number of incorrect words that can be constructed is minimised.

There are different ways to obtain better rules. We can include a new suffix on the right hand side of the rule, and at the same time expand the word list with the corresponding inflectional word forms. Another way is to substitute a new suffix for a suffix on the right hand side. A third method is to include a *negated*

suffix, which works in the following way. If the negated suffix is included, and a word exists in the word list with the suffix, then the rule cannot be applied to that word. In our syntax for suffix rules we precede the negated suffix by ~.

Example 3

The rule *-samma ← -sam, ~samen* says for example that the plural form *varsamma* of *varsam* (careful) is a correct word since *varsam* but not *varsamen* is in the last part list. The negated suffix is added so that Stava should not accept *balsamma* on the grounds that *balsam* (balsam) exists. The definite form *balsamen* is namely also in the last part list.

In order to make the suffix rules more expressive we can state that a rule should only be applicable when the suffix is preceded by (or not preceded by) a certain letter. We use standard Unix regular expression syntax for this.

Example 4

Consider the following two rules:

[uå]-rna ← -e, *-n, -r, ~dde, ~ra*
[^sxz]-s ← -e, *-en, -er*

The first rule accepts the suffix *-rna* only if it is preceded by either *u* or *å*. Thus *basturna* (the saunas) is accepted but not *vararna*. The second rule accepts the genitive suffix *-s* when it is not preceded by *s*, *x* or *z*. This means that *films* (film's) is accepted but not *sfinxs*.

In order to compare different variants of suffix rules we generate all possible words that can be constructed from a specific rule. Using the *-orna* rule in example 2, 1,592 words can be generated, and only two of them are incorrect (*dekorna* and *traktorna*). Thus, the error is 2/1 592 ≈ 0.13%.

Exception list

It is of course unsatisfactory that the algorithm accepts a few non-existing words (like *dekorna*). If we can identify these words we can avoid this problem by putting them in an exception list. This word list should contain all words that are wrongly accepted by the algorithm and should be searched before any of the other word lists. If the exception list is stored as a Bloom filter using the same hash functions as the individual word list and the last part list, the only extra work when checking a word will be to look at a few (at most *k* but more often just one or two) positions in an array of bits.

Another advantage of introducing an exception list is that we may include words in the last part list that only exist as last elements and not as whole words. In Swedish there are a couple of such words: for example *mässig* is a very common last element in compounds like *affärsmässig* (businesslike), but cannot be an individual word. If such words are added both to the last part list and the exception list, we obtain the desired result.

In some cases there are common misspellings that coincide with very uncommon inflectional forms of other words. An Swedish example of this is the misspelling *parantes* of *parentes* (parenthesis). Unfortunately *parantes* is the masculine genitive inflection of *parant* (stylish). This means that the spelling error detector will accept *parantes* even if it almost surely is a misspelling. The solution to this problem is to add the word to the exception list.

Suffix rules without errors

The reason that we have chosen the type of suffix rules described above is that the word lists that we have access to do not contain any paradigm information for the words. Thus we have to rely on the fact that certain inflectional forms of each word in each paradigm are included in the word list. This leads to the problems with overgeneration that were described above.

If we had access to a word list where each word's paradigm is marked we could use another and completely safe type of suffix rules. In the right hand side of each rule we just have one suffix (for the primary form) and a code for the word's paradigm. In the word list (last part list) we store only the primary form of each word and attach the code of the paradigm to the end of the word.

Example 5

Suppose that the paradigm of the first inflectional noun class in Swedish has the code 17. Then we include *docka*17 in the last part list and write the suffix rule for the definite plural form as *-orna ← -a*17

Spelling error correction

Many studies, see for example Damerau (1964) and Peterson (1986), show that four common mistakes cause 80 to 90 percent of all typing errors: transposition of two adjacent letters, one extra letter, one missing letter, and one wrong letter. A method that has proven useful for generating spelling correction suggestions is to generate all words that correspond to these four types of mistakes, and see which are correct words. Words that are generated in this way are said to lie at a distance of one from the original word.

A problem with the probabilistic method is that when we generate many suggestions for a misspelled word there is a slight possibility that an incorrect

word may slip in. It is however possible to reduce such errors to a minimum by introducing a *graphotactical table* as suggested by Mullin and Margoliash (1990). This table holds all allowed n-grams, i.e. combinations of n letters, for some pre-specified limit n. We have chosen $n = 4$ and we store the graphotactical table using one bit for every possible 4-gram, 1 if there is a Swedish word that contains the 4-gram and 0 otherwise. A word is accepted as correct only if all its 4-grams appear in the table. In Swedish only a small subset of the n-grams can appear at the beginning and at the end of a word. Therefore we consider the beginning and end of the word as special letters in the n-grams. A graphotactical table for Swedish constructed in this way will be filled to about 7 percent.

The reasonableness of the generated words is checked both against the Bloom filter and the graphotactical table. The words that pass both tests will be suggested as corrections.

In earlier studies of automatic spelling correction, see for instance Takahashi et al. (1990), it has been considered impractical to use word lists larger than about 10,000 words. Using our methods, it is possible to have extremely large word lists without sacrificing speed.

Spelling correction with ranking

The need for ranking

When an interactive spell checker finds a spelling error, it usually asks the user if and how she wants to correct the error. A few spelling corrections are then presented. The algorithm suggested above will find some possible corrections at a distance of one from the original word. If there are no words at a distance of one, it can compute the words that have a distance of two from the original word instead. In any case there might be a number of suggestions, and ideally they should be ranked so that the most probable correction is given as the first alternative, the second most probable correction as the second alternative and so on. If the algorithm makes a correct guess, it is easy for the user to make the change.

In some cases (for example in OCR and in spelling correction for information retrieval) there might be need for fully automatic spelling correction, i.e. the program corrects the errors without asking the user first. In this case it is of course very important that the algorithm with high probability makes the right choice among the possible corrections.

A third possibility is a semi-interactive spelling correction that reports corrections when it is clear which word the user intended to write, and asks the user when there is no single correction that is significantly more probable than the others. Then the algorithm must be able not just to rank the suggestions but to give them a *probability*.

We have studied the ranking problem under the same objectives as before, i.e. the algorithm should be fast and the full size word list is encoded as a Bloom filter. We found that the best result was obtained when we used both a refined editing distance and word frequency information for ranking the corrections (Tillenius, 1996).

Each correction suggestion is given a *penalty*, which is a number that tells us how (un)probable this word is as the correction of the misspelled word. The penalty is a combination of an editing distance penalty and a word frequency penalty.

Refined editing distance

The editing distance penalty is dependent both on the edit operation and the letters surrounding the place of the operation. There is also a penalty for changing the first letter in a word since it is uncommon for the first letter to be wrong. It is easy to generate the penalties by collecting statistics of real spelling and typing errors.

These rules can correct all of the normal keyboard typing errors and make it possible to code their probabilities (e.g. an *a* is more often mistyped as an *s* than as a *p* on a normal keyboard). The rules are also powerful enough to correct some phonetic errors. Since the correlation between spelling and pronunciation is high in Swedish, these rules work quite well for most common Swedish phonetic errors. An example is the misspelling *gort* of *gjort* (made) that is quite common since the consonant [*j*] is spelled *g* more often than *gj*.

The insertion and deletion rules will also take care of doubling and undoubling of consonants (e.g. *spel*↔*spell*, *tik*↔*tick*), which are very common types of errors in Swedish.

Word frequency

The word frequency penalty depends on how common the word is in the Swedish language (ideally taken over the text type that the user is currently writing). More common words give a lower penalty. We chose to divide the words into 10 frequency classes named *A*, *B*,..., *J*. The word frequencies were stored in a separate Bloom filter where each word was concatenated with the letter corresponding to its frequency class. For example the very common word *och* (and) is in frequency class A and is thus stored in the Bloom filter as *ochA*. In this way the frequency class of a word can be found by at most 10 look-ups in the Bloom filter.

Evaluation of our ranking

An evaluation of the ranking method on 729 misspelled words shows that it finds the correct correction in 60% of the cases, see table 1. This is very good, especially taking into consideration that only 78% of the corrections were included in the word list of the program.

Table 1
Performance of different spelling correction methods tested on 729 misspelled words.

Method	1	2	3
none	204 (28%)	71 (10%)	16 (2%)
word freq.	356 (49%)	42 (6%)	26 (4%)
edit dist.	388 (53%)	55 (8%)	16 (2%)
edit dist.+word freq.	440 (60%)	28 (4%)	10 (1%)

Columns 1, 2, and 3 tell whether the correct word was the first, second or third suggestion. None means that no ranking was performed, and that the suggestions were presented in the order they were generated.

We also tried to use word bigrams to rank the suggestions, but this was not successful. The reason was that most correct bigrams were not included in the bigram database (containing 200,000 bigrams) that we used. Word bigrams might work better on tests with a smaller vocabulary. We did not try to use word class tag bigrams, which perhaps would improve the ranking if word tags are available.

Our implementation: Stava

We have implemented these algorithms for spelling error detection, correction and ranking as a C program of 4,000 lines. The program is called Stava. Documentation and a test version of Stava are available on the web (Kann, 1998). In the following we will describe and discuss some implementation aspects.

Word lists and suffix rules

We have used many sources of Swedish words for Stava. The main source is the word list of the Swedish Academy (1986) with 120,000 words and information about inflections. For the word frequency list we have used a source consisting of 200,000 words collected from a newspaper corpus of 1,000,000 words composed by Språkdata at the University of Gothenburg.

The last part list consists of about 100,000 words, the first part list of about 25,000 words, the individual word list of about 1,000 words and the exception list of about 1,000 words.

There are about 1,000 suffix rules in Stava. When constructing the rules we have used the Swedish morphology as described by Hellberg (1978). The rules are sorted by the suffix on the left-hand side reversed (from right to left). This means that we can use a binary search when looking for rules that match a given word. About 500,000 words can be constructed from the suffix rules and last part list.

When suffix rules are matched against a word it often happens that the same word has to be looked up in the last part list several times. In order to minimize the number of look-ups we have a special cache that remembers the last look-ups and their results.

Optimization of the hash functions

Every hash function $h_i(w)$ in the Bloom filters had the following basic structure, where c_j is the number associated with the jth character in the word w, $|w|$ is the number of characters in w, and p_i is a prime smaller than the size of the hash table.

$$h_i(w) = \sum_{j=1}^{|w|} 2^{7(j-1)} c_j \bmod p_i$$

The main part of the execution time (more than 80%) was spent on computing the hash functions. Therefore it was very important to speed up the computation of $h_i(w)$.

First we noted that the most time-consuming operation is the mod computation, since the remainder taking hides a division. We tried to get rid of the division by precomputing $1/p_i$ once for all and using floating number multiplication instead of remainder taking. This improved the total running time by a factor of two.

The next improvement was made by performing mod once per hashing instead of once for each character. This is possible without overflow for short words, but if the program is run on the same computer as the Bloom filter is built we can in fact forget about overflow – the important thing is that the computed hash value is the same each time. This improved the running time by another factor of two.

Now we wanted to get rid of the mod operation completely. If we choose the hash table size as an exponent of 2, the mod operation can be performed by a simple and extremely fast bit mask. This would destroy the even distribution of the above hash function, so we had to change to a hash function that mixes all the

bits of the hash value so that taking just the last bits still gives an even distribution. For this we used a hash function constructed by Jenkins (1997).

Finally we observed that the same hash functions (mod different numbers) are computed twice, since a word is searched both in the individual word list and in the last part list. When we had changed the program so that we re-used earlier computed hash values we had made a total optimization by a factor of ten with respect to the unoptimized program. After this Stava could check 10,000 words per second on a Sun SPARCstation 10, a Unix machine comparable to a Pentium PC.

Applications

In this paper we have seen that our methods give a good spelling error detection and correction for Swedish. We have also used these methods successfully in several other applications.

Using the method on other languages than Swedish

The spelling detection and correction method described in this paper is not limited to Swedish. We have successfully used it with an English word list and some very simple suffix rules. We have also shown (but not implemented) that the method is suitable for Russian with its quite complicated inflections (Engebretsen, 1997; Axensten, 1997).

If the method is to be used on a language where inflections change letters at the beginning or middle of the word and not just at the end, the suffix rule language has to be extended, but this should be straightforward.

Creating a part-of-speech lexicon

Building a complete part-of-speech lexicon where each word is tagged with syntactic category and inflectional morphological features is an extremely hard and time-consuming task. The work will diminish drastically when using Stava's suffix rules extended with tagging information. All inflected forms of all regularly inflected words may be constructed automatically.

Example 6

The word *dockas* is either the genitive of the noun *docka* (doll) or the passive of the verb *docka* (dock). This is reflected by the following two rules:

-as	←	*-a, -an, -or*
-as	←	*-a, -ade*

If these rules are extended with the tags *nn.utr.sin.ind.gen* and *vb.inf.sfo*, *vb.prs.sfo*[1] respectively we can use the ordinary suffix rule search of Stava to conclude that *dockas* should be tagged with the three tags *nn.utr.sin.ind.gen*, *vb.inf.sfo*, *vb.prs.sfo*.

Furthermore, by starting from the original last part list we can generate a part-of-speech lexicon. However, there are two problems with this approach: first there are no suffix rules for the inflections that are in the last part list (for example *docka*, *dockan* and *dockor*), and secondly there are no suffix rules at all for irregularly inflected words and words that are not inflected at all.

We can deal with the first problem by simply adding suffix rules also for the inflections included in the last part list. This can be done automatically by adding suffix rules for all suffixes that appear positively on the right hand side of the rule.

Example 7

For the noun suffix rules in the example above we add the following rules:

-a	←	*-a, -an, -or*	*nn.utr.sin.ind.nom*
-an	←	*-a, -an, -or*	*nn.utr.sin.def.nom*
-or	←	*-a, -an, -or*	*nn.utr.plu.ind.nom*

The second problem cannot be solved automatically. The irregular words and words without inflections have to be tagged by hand. Fortunately these are not so many in Swedish. Less than 3% (3,000 of 100,000) of the words in our last part list are of this type.

In Swedish all words in the open word classes can be inflected, which means that the number of words that have to be tagged by hand is constant.

Also note that the tagged suffix rules described above can also be used to extend an existing part-of-speech lexicon with tags for words that are already included in the lexicon. Often just the common tags for a word are included, even if it might be necessary to know uncommon tags in order to be sure that the correct tagging of a word is in the lexicon.

Finding the parts of a compound word in hyphenation

In Swedish, one word compounds can be very long, so there is a large need for hyphenation of compound words. The Swedish hyphenation rules say that a compound should preferably be hyphenated between the elements. This means

[1] We have used the Swedish tagging system defined in SUC (Ejerhed et al., 1992).

that a Swedish hyphenation algorithm cannot consist only of local hyphenation rules. It must be able to split a compound into its elements.

We have used Stava's method for doing this. In Stava a compound is accepted if there is a way to split it into one or more elements in the *first part list* and one element in the *last part list*. If there is more than one way to split a compound every possible split is investigated and the best one is chosen. We have found that a split consisting of few elements and where the last element is long is often the correct split. Therefore we used the following objective function for choosing between different splittings.

Maximize {(number of characters of last element) - 3·(number of elements)}

Using this objective function on a list of 66,000 compounds 95.5% were split correctly, 3.0% were split incorrectly, and 1.5% were not split at all.

In order to choose splits like *kvarts-ur* (quartz watch) instead of *kvart-sur* (something like quarter sour), that is the letter *s* is moved to the first part instead of the last part in spite of the fact that this gives a shorter last part, we changed the algorithm so that it prefers the first splitting. Unfortunately the gain was only 0.03% (227 more words were now correctly hyphenated, but at the same time 202 words were incorrectly hyphenated).

Part-of-speech tagging of unknown words

A part-of-speech tagger typically has a lexicon consisting of words and possible taggings of these words (for example constructed using the methods above). When tagging a new text there might be unknown words, i.e. words that are not in the lexicon. The possible tags of these unknown words have to be guessed.

In Swedish the unknown words can be divided into three main groups: new compounds, proper nouns (names) and uncommon simple words (usually technical terms or dialectal words).

A compound can be split into its elements using the method described above. The tagging of a Swedish compound is decided by the tagging of its last element, so if the last element is in the lexicon we can just look up the tags.

Proper nouns can be separated from uncommon simple words in most cases since their initial letter is a capital. Otherwise we have to guess the tags from the word's appearance in some way. A good way is to use the suffix rules again. They contain both suffixes and tags, so we can look at the last few letters of the word, see if any suffix rules apply and return the corresponding tags. If we have some frequency statistics on the rules we will be able to guess which tag is the most probable.

Lemmatisation and spelling correction in information retrieval

A common approach for an information retrieval system is to process the search question as well as the documents by removing all non-significant words (using a

stop list) and performing a lemmatisation of the rest of the words so that different inflections of a search term in the question and in the document do not matter.

The suffix rules in Stava can be used for lemmatisation. When a word matches a suffix rule we can transform it into its base form by using the first suffix on the right-hand side of the rule. Adding suffix rules as described above solves the problem with inflectional forms that are already in the last part list.

Spelling correction can also be used in information retrieval. Up to a third of the search terms given to web search engines are misspelled. Also a large number of documents available in any given database contain misspelled terms. Since the number of untrained and novice users and low-budget text producers is increasing, the need for spelling correction in information retrieval will probably increase in the future.

The users can for example be offered interactive spelling correction of misspelled search terms. This would improve search results both as regards precision and recall. Spelling correction of the indexed documents will also improve the search results, but if this is to be of practical use the correction has to be fully automatic.

We have used Stava's spelling correction method in the web version of Skolverket's Swedish-English dictionary (Skolverket, 1997) which contains 28,500 Swedish words. Every day about 20,000 questions are asked of the web dictionary. Of these 20% are misspelled. For 33% of the misspellings a single search key is at closest distance to the misspelling, so the question can be corrected automatically.

Directions for future research

We have shown that the Stava method is powerful enough to detect spelling errors and to construct and rank spelling corrections very fast. A shortcoming of the method is that it only finds spelling errors where the misspelled word is not a correct Swedish word. In many misspellings, especially of short words, the misspelled word coincides with a correct word, for example *för* (for) is easily misspelled as *frö* (seed).

A probabilistic tagger that uses word and tag frequencies as well as tag bigrams and trigrams might be able to find many misspellings of this type. We will investigate this in a new project.

In particular we will look at the special case of compound splitting when for example *bokhylla* (bookshelf) is written as *bok hylla*. This type of spelling error has become more common in Swedish, probably due to English influences. By looking at the syntactic categories and frequencies of both the separate words and the compound we hope to be able to find most cases of compound splitting.

In the new project we will try to detect and correct grammatical errors. When correcting a grammatical error where a word has a wrong inflectional form, we

know which tag the word has and which it should have. Thus we can once again use Stava's suffix rules to construct the correction.

Having access to the tagging of the words in the document and tag frequencies also makes it possible to improve the ranking of ordinary spelling corrections.

Acknowledgements

The research has been funded in the Language Engineering program (Språkteknologiprogrammet) by HSFR and Nutek.

We would like to thank Språkdata at the University of Gothenburg and Svenska Akademien for letting us use *Svenska Akademiens ordlista* as a source for words in Stava, and Per Hedelin for letting us use the SUL word list for evaluation purposes.

References

Axensten, P. (1997). *Stava ryska adjektiv* (Spell Russian adjectives). Technical report, Department of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm. In Swedish. Web version available at http://www.nada.kth.se/theory/projects/swedish.html.

Bloom, B.H. (1970). Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM,* 13, 422–426.

Damerau, F.J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7, 171–176.

Domeij, R., Hollman, J., & Kann, V. (1994). Detection of spelling errors in Swedish not using a word list en clair. *Journal of Quantitative Linguistics,* 1, 195–201.

Ejerhed, E., Källgren, G., Wennstedt, O., & Åström, M. (1992). *The linguistic annotation system of the Stockholm-Umeå corpus project. Technical Report DGL-UUM-R-33.* Department of General Linguistics, University of Umeå, Umeå.

Engebretsen, L. (1997). *De ryska böjningsmönstrens betydelse vid maskinell rättstavning* (The influence of Russian paradigms on spelling correction). Technical report. Department of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm. In Swedish. Web version available at http://www.nada.kth.se/theory/projects/swedish.html.

Hellberg, S. (1978). *The Morphology of Present-Day Swedish.* Stockholm: Almqvist & Wiksell.

Jenkins, R. J. (1997). Hash functions. *Dr Dobb's Journal,* 22, 107-109.

Kann, V. (1998). *Stava's home page*, http://www.nada.kth.se/stava.

Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24, 377-439.

Mullin, J.K., & Margoliash, D.J. (1990). A tale of three spelling checkers. *Software-Practice and Experience,* 20, 625-630.

Peterson, J.L. (1986). A note on undetected typing errors. *Communications of the ACM,* 29, 633-637.

Skolverket. (1997). *Lexin Swedish-English dictionary.* Stockholm: Norstedts. Web version available at http://www.nada.kth.se/skolverket/swe-eng.html.

Svenska Akademien (The Swedish Academy). (1996). *Ordlista över svenska språket (SAOL).* Stockholm: Norstedts. 11th edition.

Takahashi, H., Itoh, N., Amano, T., & Yamashita, A. (1990). A spelling correction method and its application to an OCR system. *Pattern Recognition,* 23, 363-377.

Tillenius, M. (1996). *Efficient generation and ranking of spelling error corrections.* Technical Report TRITA-NA-E9621, Department of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm.

Assoziativität der Phoneme im Russischen

Sebastian Kempgen

Auch wenn das Russische im allgemeinen und auch quantitativ als gut untersuchte Sprache gelten kann (vgl. auch Kempgen, 1995a), so gibt es doch erstaunlicherweise immer noch Bereiche, die nicht nach dem aktuellen methodischen Stand bearbeitet sind. Zu diesen Phänomenen gehört u.a. eine Ermittlung und Beschreibung der Phonemdistribution mit den Assoziativitätsmaßen, die von Altmann & Lehfeldt (1980) auf der Grundlage des Kalküls von Harary & Paper (1957) entwickelt worden sind. Es läßt sich plausibel machen, daß es vor allem die Ebene der Silben ist, auf der die Phonemdistribution untersucht werden sollte – die Silbe gehört als einzige Rahmeneinheit der gleichen – phonologischen – Ebene an und erfüllt u.a. genau den Zweck, die kleinsten segmentalen Einheiten zu größeren Verbünden zu organisieren. Als Besonderheit im strukturalistischen Schichtenmodell der Sprache, dessen höchste Ebene der Text ist, bleibt jedoch zu bemerken, daß Phoneme als unmittelbare Konstituenten von Silben wie auch von Morphemen betrachtet werden müssen. Morpheme können sich – im Russischen – zwar in Silben zerlegen lassen, dies ist jedoch nicht zwingend erforderlich. Auf der Ebene des Wortes überlagern sich somit zwei Organisationsschichten: Wörter (genauer: Wortformen) lassen sich in Morpheme zerlegen, und sie lassen sich in Silben zerlegen.

Der vorliegende Beitrag präsentiert die Distribution der Phoneme des Russischen im Rahmen der Silbe. Die Daten wurden durch Untersuchung zahlreicher Wörterbücher gewonnen, stellen also eine systemische Stichprobe dar, die Syllabisierung (Zerlegung der phonologischen Wörter in Silben) erfolgte nach Lehfeldt (1971). Gleichzeitig sollen die Distributionsmaße von Altmann & Lehfeldt durch Berücksichtigung der sog. 'modellbedingten Vorkommensbeschränkungen' präzisiert und modifiziert werden. Der Artikel möchte also die Beschreibung des Russischen auf phonologischer Ebene vervollständigen und gleichzeitig einen Beitrag zur Weiterentwicklung der phonologischen Assoziativitätsmaße leisten.
Bei den 'modellinternen Vorkommensbeschränkungen' (eingeführt von Kempgen [im Druck]) handelt es sich um Lücken in der Phonemdistribution, die nicht empirischer Natur sind, sondern der modellinternen Logik entspringen. Die systematische Berücksichtigung dieser Lücken macht eine Beschreibung der Distributionsfähigkeiten realitätsnäher. Ein Beispiel zur Demonstration: Im Russischen kommt die Graphemfolge *zd* vor, z.B. in *zdes'* 'hier'. An das palatalisierte [d'] paßt sich der vorausgehende Laut an, der ebenfalls erweicht wird: [z'd']. Arbeitet man mit einer Allophon-nahen Phonemisierung und notiert deshalb /z'd'/, so kann es aus Gründen der modellinternen Konsistenz keine Phonemfolge */zd'/ geben, denn jedes auftretende *zd* würde ja als /z'd'/ notiert, nach der Verbindung */zd'/ brauchen wir empirisch also gar nicht erst zu suchen. Die Berücksichtigung dieser Fälle ist insbesondere für die Bestimmung der theoretischen wie praktischen Obergrenze der Zahl derjenigen Verbindungen, in die ein Phonem eingehen kann, relevant.

Das russische Phoneminventar besteht in unserer Beschreibung aus $N = 39$ Einheiten: Vokale (5): */ a, e, i, o, u/*

Konsonanten (34):*/b b' v v' g g' d d' ž z z' j k k' l l' m m' n n' p p' r r' s s' t t' f f' x c č š /*

Tabelle 1 zeigt die Distribution der Phoneme des Russischen im Rahmen der Silbe in kategorischer Form, d.h. die Frequenz bleibt hier unberücksichtigt. Die Existenz der Verbindung *ij* wird durch '+' markiert, die modellbedingten Vorkommensbeschränkungen – von denen es mehrere Arten gibt – werden unterschiedslos durch '0' angezeigt.

Diese Daten basieren auf einer Auswertung des Materials, das im Anhang unter den Primärquellen im einzelnen genannt wird. Zu einer weiteren Auswertung des gleichen Materials vgl. Kempgen (1995b).

Das Erstellen einer solchen Matrix ist bekanntlich nur das erste Ziel einer Distributionsanalyse, nicht ihr Endpunkt. Aufbauend auf den in der Distributionsmatrix enthaltenen Daten geht es vielmehr im nächsten Schritt darum, die Phoneme hinsichtlich ihrer kombinatorischen Eigenschaften präzise zu beschreiben und, in einem weiteren Schritt, zu klassifizieren (diesen Schritt werden wir an dieser Stelle aus Platzgründen allerdings nicht durchführen).

Die von Altmann & Lehfeldt (1980) entwickelten Distributionsmaße basieren auf einigen wenigen Grundmengen, die hier in allgemeiner Form kurz resümiert seien:

1. *Ai* ist die Menge der Phoneme, die vor *i* stehen können; sie wird auch als die Menge der *Vorgängerphoneme* von *i* bezeichnet. In der Distributionsmatrix können diese Phoneme der zu *i* gehörenden *Spalte* entnommen werden.
2. *iB*. Das Gegenstück zu *Ai* ist die Menge der *Nachfolgerphoneme*, in der die Phoneme versammelt sind, die an zweiter Stelle stehen können, wenn *i* vorausgeht. Sie sind in der Matrix der zu i gehörenden *Zeile* zu entnehmen.

3. $Ai \cap iB$. Dies ist die *Durchschnittsmenge* der Vorgänger- und Nachfolgerphoneme, also die Menge der Phoneme, die sowohl vor als auch nach einem bestimmten Phonem *i* auftreten.
4. $Ai \cup iB$. Die *Vereinigungsmenge* der Vorgänger- wie Nachfolgerphoneme umfaßt alle Phoneme, die entweder vor oder nach *i* oder auch vor und nach *i* auftreten, also mindestens in einer von beiden Positionen. Mit anderen Worten: dies ist die Gesamtmenge der Phoneme, die sich überhaupt mit *i* verbinden, also ohne Berücksichtigung der Reihenfolge.
5. $Ai \otimes iB = |Ai \cup iB| - |Ai \cap ib|$. Dies ist die sog. symmetrische Differenz. Sie umfaßt diejenigen Phoneme, die sich nur in einer der beiden möglichen Positionen mit *i* verbinden, bei denen also die Reihenfolge eine Rolle spielt.

Der Tabelle 2 kann die Mächtigkeit der genannten Mengen für die Phoneme des Russische entnommen werden. Nur aus Platzgründen geben wir hier die absoluten Zahlen an, nicht die in das Einheitsintervall $< 0 ; 1 >$ transformierten Werte. Letztere erhält man ja einfach, indem die absoluten Zahlen durch den Umfang *N* des Phoneminventars (39) dividiert werden.

Die Distributionsmaße von Altmann & Lehfeldt (1980) lassen sich in drei Gruppen einteilen:

a) Assoziativitätsmaße,
b) Symmetriemaße und
c) Reflexivitätsmaße.

Außerdem lassen sich Einzelmaße und Vergleichsmaße unterscheiden. Bei den Assoziativitätsmaßen nennt man die Fähigkeit, Vorgängerphoneme zu haben, *Attraktivität*, und die Fähigkeit, Nachfolgerphoneme zu haben, *Aggressivität*. Die Assoziativitätsmaße sind jeweils definiert als das Verhältnis der Zahl der beobachteten Phonempaare (s. Tabelle 2) zur Zahl der theoretisch maximal möglichen Paare. Die absolute Obergrenze ist durch den Umfang des Phoneminventars gegeben, für die Berechnung einer relativen (individuellen) Obergrenze stehen mehrere Möglichkeiten zur Verfügung (s.u.).

Im folgenden sollen sieben Assoziativitätsmaße berechnet werden, deren genaue Definition in der genannten Arbeit leicht nachzulesen ist:

- das Maß der Attraktivität (Vorgängermenge),
- das Maß der Aggressivität (Nachfolgermenge),
- das Maß der vollständigen Assoziativität (Vereinigungsmenge),
- das Maß der Paarbildung (Vorgänger- plus Nachfolgermenge).

Bei diesen Maßen wird die Zahl der beobachteten Kombinationen jeweils auf den Umfang des Phoneminventars bezogen. Bei den nächsten drei Maßen wird hingegen die Zahl der Phoneme verwendet, mit denen das Phonem *i* überhaupt eine Verbindung eingeht, also $|Ai \cup iB|$. Wenn man diese Größe zugrundelegt, also einen Phonem-eigenen Maßstab, ergeben sich aus den oben genannten Maßen die sogenannten Maße der *internen Assoziativität*:

- das Maß der internen Attraktivität;
- das Maß der internen Aggressivität;
- das Maß der internen Paarbildung.

Das *Maß der internen Paarbildung* von *i* wird hier von uns zusätzlich eingeführt. Wir beziehen hier die Summe von Vorgänger- und Nachfolgerphonemen auf die Zahl der Verbindungen, in die *i* eingehen würde, wenn es sich mit allen seinen Vorgänger und Nachfolgerphonemen gleichermaßen als erstes und als zweites Glied verbinden würde. Es ist danach definiert als:

$$As_i^*(i) = \frac{|Ai| + |iB| - G(i)}{2(|Ai \cup iB|) - G(i)}$$

$G(i)$ steht hierbei für die Fähigkeit, eine Geminate zu bilden, die natürlich nur einmal gewertet werden darf und deshalb aus den Vorgänger- und Nachfolgermengen herausgenommen wird.

Die Werte, die diese sieben Maße für das Russische annehmen, finden sich in Tabelle 3.

Kommentieren wir einige auffallende Fälle. Das Maß *Ati*(f') nimmt den Wert 1 an. Das heißt: alle Phoneme, mit den sich /f'/ überhaupt verbindet, stehen auch vor ihm, oder andersherum: nach /f'/ stehen keine Phoneme, die nicht auch vor ihm stünden. In der Sprachgeschichte des Slavischen stellt /f/ bekanntlich insofern einen besonderen Fall dar, als es sich erst relativ spät – zuerst durch Übernahme griechischer Lehnwörter ins Kirchenslawische – eingebürgert hat, was bis heute seine Spuren in der Tatsache hinterläßt, daß das stimmhafte Gegenstück, also /v/, von Stimmtonassimilationen nicht betroffen ist.

Für die Phoneme /ž/ und /š/ nimmt das Maß *Agi*(*i*) ebenfalls den Wert 1 an, was besagt, daß alle Phoneme, mit denen sich diese beiden Zischlaute verbinden, auf jeden Fall auch nach ihnen auftreten. Unter den Konsonanten sind dies die einzigen Fälle, in denen Phoneme den Maximalwert eines Maßes erreichen.

Betrachten wir die zu /s/ gehörige Zeile der Distributionsmatrix (Tabelle 1), so stellen wir fest, daß sie vollständig entweder durch '+' oder durch '0' ausgefüllt ist, d.h. alle theoretisch denkbaren Kombinationen mit /s/, die nicht realisiert werden, fallen unter die modellbedingten Vorkommensbeschränkungen. Oder anders: alle Kombinationen mit /s/ als erstem Glied, die es im Rahmen unseres Modells überhaupt geben kann, gibt es auch. Das Maß der Aggressivität nimmt für /s/ nach der 'kanonischen Definition' den Wert 20/39 an. Man könnte nun

sagen: dieser Wert ist unrealistisch niedrig, da /s/ ja 'in Wirklichkeit' seine maximale Kombinationsfähigkeit ausschöpft. Die Berücksichtigung der modellbedingten Vorkommensbeschränkungen in den Maßen der Assoziativität ist folglich 'realitätsnäher' als die gewöhnlichen Maße.

Zusätzlich zu den Definitionen bei Altmann & Lehfeldt (1980) legen wir deshalb zunächst folgendes fest:

1. Die Relation R_m ist zu verstehen als „kann im Rahmen des Modells nicht unmittelbar gefolgt werden von…".
2. $Mi = \{ j / j \in N, ji \in R_m \}$
   Dies ist die Menge der Phoneme, die aus modellbedingten Vorkommensbeschränkungen nicht vor *i* auftreten können.
3. $iM = \{ j / j \in N, ij \in R_m \}$
   In dieser Menge sind die Phoneme vertreten, die aufgrund von modellbedingten Vorkommensbeschränkungen nicht nach *i* auftreten können.
4. Die Mengen $Mi \cap iM$ und $Mi \cup iM$ ergeben sich analog zu oben.
   Wir definieren deshalb entsprechend modifizierte Maße, in den das 'm' auf die Berücksichtigung der modellbedingten Vorkommenbeschränkungen verweist.

Das *Maß der modellinternen Attraktivität* von *i* sei definiert als:

$$At_m(i) = \frac{|Ai|}{n - |Mi|}$$

Das Maß der modellinternen Aggressivität von i sei definiert als:

$$Ag_m(i) = \frac{|iB|}{n - |iM|}$$

Das Maß der modellinternen Assoziativität sei definiert als:

$$As_m(i) = \frac{|Ai \cup iB|}{n - |Mi \cap iM|}$$

Die Ergebnisse der Berechnungen dieser Maße für das Russische lassen sich aus Tabelle 4 ablesen. Natürlich können diese Werte nur gleich hoch oder höher als die Werte der „Normalmaße" sein. Wie man sieht, erreicht nur /s/ unter den Konsonanten den Maximalwert von 1.0 in einem dieser Maße, ist also besonders kombinationsfreudig.

Um Sprachen hinsichtlich ihrer distributionellen Eigenschaften vergleichen zu können, gibt es die sog. Totalmaße, die das System als Ganzes charakterisieren.

Das Maß der totalen Assoziativität nimmt für das Russische den Wert 0.5437 an. Auch hier ist es wieder sinnvoll, die modelbedingten Vorkommensbeschränkungen zu berücksichtigen, wonach sich ein Wert von 0.6702 ergibt. Das zugehörige Maß wird von uns definiert als

$$As_m(L) = \frac{|R|}{n^2 - |R_m|}$$

R steht hier für die Menge der realisierten Phonemkombinationen. R_m für die Zahl der modellintern ausgeschlossenen Kombinationen.

Tabelle 1
Phonemdistribution im Russischen (Rahmeneinheit: Silbe)

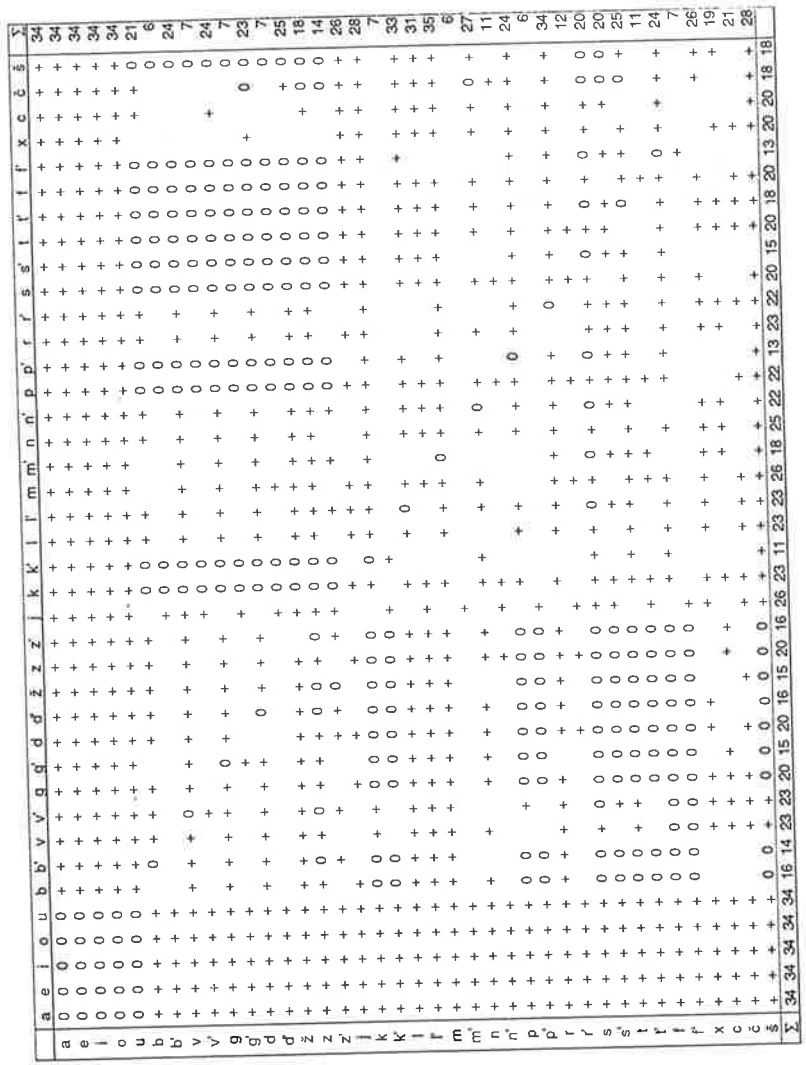


Tabelle 2
Distributionelle Eigenschaften der russischen Phoneme

i	*Ai*	*iB*	*Ai* ∪ *iB*	*Ai* ∩ *iB*	*Ai* ⊗ *iB*
a	34	34	34	34	0
e	34	34	34	34	0
i	34	34	34	34	0
o	34	34	34	34	0
u	34	34	34	34	0
b	16	21	23	14	9
b’	14	6	15	5	10
v	23	24	31	16	15
v’	23	7	24	6	18
g	20	24	27	17	10
g’	15	7	16	6	10
d	20	23	25	18	7
d’	16	7	17	6	11
ž	15	25	25	15	10
z	20	18	21	17	4
z’	16	14	22	8	14
j	26	26	38	14	24
k	23	28	29	22	7
k’	11	7	12	6	6
l	23	33	35	21	14
l’	23	31	32	22	10
m	26	35	36	25	11
m’	18	6	19	5	14
n	25	27	30	22	8
n’	22	11	26	7	19
p	22	24	26	20	6
p’	13	6	14	5	9
r	23	34	35	22	13
r’	22	12	23	11	12
s	20	20	23	17	6
s’	15	20	23	12	11
t	20	25	27	18	9
t’	18	11	19	10	9
f	20	24	25	19	6
f’	13	7	13	7	6
x	20	26	28	18	10
c	20	19	28	11	17
č	18	21	25	14	11
š	18	28	28	18	10

Tabelle 3
Assoziativitätsmaße im Russischen

	$At(i)$	$Ag(i)$	$As(i)$	$As^*(i)$	$At_I(i)$	$Ag_I(i)$	$As_I^*(i)$
a	0,8718	0,8718	0,8718	0,8831	1,0000	1,0000	1,0000
e	0,8718	0,8718	0,8718	0,8831	1,0000	1,0000	1,0000
i	0,8718	0,8718	0,8718	0,8831	1,0000	1,0000	1,0000
o	0,8718	0,8718	0,8718	0,8831	1,0000	1,0000	1,0000
u	0,8718	0,8718	0,8718	0,8831	1,0000	1,0000	1,0000
b	0,4103	0,5385	0,5897	0,4805	0,6957	0,9130	0,8043
b'	0,3590	0,1538	0,3846	0,2597	0,9333	0,4000	0,6667
v	0,5897	0,6154	0,7949	0,5974	0,7419	0,7742	0,7541
v'	0,5897	0,1795	0,6154	0,3766	0,9583	0,2917	0,6170
g	0,5128	0,6154	0,6923	0,5584	0,7407	0,8889	0,8113
g'	0,3846	0,1795	0,4103	0,2727	0,9375	0,4375	0,6774
d	0,5128	0,5897	0,6410	0,5584	0,8000	0,9200	0,8600
d'	0,4103	0,1795	0,4359	0,2987	0,9412	0,4118	0,6765
ž	0,3846	0,6410	0,6410	0,5065	0,6000	1,0000	0,7959
z	0,5128	0,4615	0,5385	0,4805	0,9524	0,8571	0,9024
z'	0,4103	0,3590	0,5641	0,3766	0,7273	0,6364	0,6744
j	0,6667	0,6667	0,9744	0,6753	0,6842	0,6842	0,6842
k	0,5897	0,7179	0,7436	0,6494	0,7931	0,9655	0,8772
k'	0,2821	0,1795	0,3077	0,2208	0,9167	0,5833	0,7391
l	0,5897	0,8462	0,8974	0,7143	0,6571	0,9429	0,7971
l'	0,5897	0,7949	0,8205	0,7013	0,7188	0,9688	0,843
m	0,6667	0,8974	0,9231	0,7792	0,7222	0,9722	0,8451
m'	0,4615	0,1538	0,4872	0,3117	0,9474	0,3158	0,6316
n	0,6410	0,6923	0,7692	0,6623	0,8333	0,9000	0,8644
n'	0,5641	0,2821	0,6667	0,4286	0,8462	0,4231	0,6346
p	0,5641	0,6154	0,6667	0,5844	0,8462	0,9231	0,8824
p'	0,3333	0,1538	0,3590	0,2468	0,9286	0,4286	0,6786
r	0,5897	0,8718	0,8974	0,7403	0,6571	0,9714	0,8143
r'	0,5641	0,3077	0,5897	0,4416	0,9565	0,5217	0,7391
s	0,5128	0,5128	0,5897	0,5065	0,8696	0,8696	0,8667
s'	0,3846	0,5128	0,5897	0,4416	0,6522	0,8696	0,7556
t	0,5128	0,6410	0,6923	0,5844	0,7407	0,9259	0,8333
t'	0,4615	0,2821	0,4872	0,3766	0,9474	0,5789	0,7632
f	0,5128	0,6154	0,6410	0,5584	0,8000	0,9600	0,8776
f'	0,3333	0,1795	0,3333	0,2468	1,0000	0,5385	0,7600
x	0,5128	0,6667	0,7179	0,5974	0,7143	0,9286	0,8214
c	0,5128	0,4872	0,7179	0,5065	0,7143	0,6786	0,6964
č	0,4615	0,5385	0,6410	0,5065	0,7200	0,8400	0,7800
š	0,4615	0,7179	0,7179	0,5844	0,6429	1,0000	0,8182

Tabelle 4
Modellinterne Assoziativität im Russischen

i	Ai	iB	Mi	iM	$Ai \cup iB$	$Mi \cap iM$	$At_M(i)$	$Ag_M(i)$	$As_M(i)$
a	34	34	5	5	34	5	1,0000	1,0000	1,0000
e	34	34	5	5	34	5	1,0000	1,0000	1,0000
i	34	34	5	5	34	5	1,0000	1,0000	1,0000
o	34	34	5	5	34	5	1,0000	1,0000	1,0000
u	34	34	5	5	34	5	1,0000	1,0000	1,0000
b	16	21	11	12	23	11	0,5714	0,7778	0,8214
b'	14	6	13	11	15	11	0,5385	0,2143	0,5357
v	23	24	2	12	31	2	0,6216	0,8889	0,8378
v'	23	7	5	11	24	3	0,6765	0,2500	0,6667
g	20	24	11	12	27	11	0,7143	0,8889	0,9643
g'	15	7	12	11	16	11	0,5556	0,2500	0,5714
d	20	23	11	13	25	11	0,7143	0,8846	0,8929
d'	16	7	13	11	17	11	0,6154	0,2500	0,6071
ž	15	25	13	11	25	11	0,5769	0,8929	0,8929
z	20	18	11	20	21	11	0,7143	0,9474	0,7500
z'	16	14	12	13	22	11	0,5926	0,5385	0,7857
j	26	26	0	0	38	0	0,6667	0,6667	0,9744
k	23	28	11	10	29	9	0,8214	0,9655	0,9667
k'	11	7	12	9	12	9	0,4074	0,2333	0,4000
l	23	33	0	1	35	0	0,5897	0,8684	0,8974
l'	23	31	3	0	32	0	0,6389	0,7949	0,8205
m	26	35	0	1	36	0	0,6667	0,9211	0,9231
m'	18	6	3	0	19	0	0,5000	0,1538	0,4872
n	25	27	0	2	30	0	0,6410	0,7297	0,7692
n'	22	11	3	0	26	0	0,6111	0,2821	0,6667
p	22	24	11	10	26	9	0,7857	0,8276	0,8667
p'	13	6	13	9	14	9	0,5000	0,2000	0,4667
r	23	34	0	1	35	0	0,5897	0,8947	0,8974
r'	22	12	1	0	23	0	0,5789	0,3077	0,5897
s	20	20	11	19	23	10	0,7143	1,0000	0,7931
s'	15	20	12	11	23	9	0,5556	0,7143	0,7667
t	20	25	11	11	27	9	0,7143	0,8929	0,9000
t'	18	11	13	9	19	9	0,6923	0,3667	0,6333
f	20	24	11	12	25	11	0,7143	0,8889	0,8929
f'	13	7	13	11	13	11	0,5000	0,2500	0,4643
x	20	26	0	0	28	0	0,5128	0,6667	0,7179
c	20	19	0	0	28	0	0,5128	0,4872	0,7179
č	18	21	7	0	25	0	0,5625	0,5385	0,6410
š	18	28	13	9	28	9	0,6923	0,9333	0,9333

Primärquellen (Korpus)

Babkin, A.M., & Levasov, E.A. (red.) (1975). *Slovar' nazvanij žitelej SSSR.* Moskva: Russkij jazyk.

Barxudarov, S.G., Protčenko, I.F., & Skvorcov, L.I. (red.) (1974). *Orfografičeskij slovar' russkogo jazyka.* Izd. 13-oe. Moskva: Russkij jazyk.

Benson, M. (1967). *Dictionary of Russian Personal Names. With a guide to stress and morphology.* 2nd. ed. Philadelphia, PA: University of Philadelphia Press.

Bielfeldt, H.H. (1972). *Russisch-deutsches Wörterbuch.* Berlin. 2. Auflage: Akademie Verlag.

Flegon, A. (1973). *Za predelami russkich slovarej.* London.

Friedrich, W., & Geis, S. (1976). *Russisch-deutsches Neuwörterbuch.* München: Max Hueber.

Galler, M., & Marquess, H.E. (1972). *Soviet Prison Camp Speech. A survivor's glossary supplemented by terms from the works of A.I. Solženicyn.* Madison, Wisc.: University of Wisconsin Press.

Obratnyj slovar' (1974). *Obratnyj slovar' russkogo jazyka.* Moskva: Sovetskaja Enciklopedija.

Vasmer, M. (ed.) (1961ff.). *Wörterbuch der russischen Gewässernamen.* Zusammengestellt unter Leitung von Max Vasmer. Bd. I-V. Berlin: Harrassowitz.

Vasmer, M. (ed.) (1964ff.). *Russisches geografisches Namensbuch.* Begründet von Max Vasmer. Bd. I–VI. Wiesbaden: Harrassowitz.

Volostnova, M.B. (red.) (1968). *Slovar' geografičeskich nazvanij SSSR.* Moskva: Nedra.

Volostnova, M.B. (red.) (1970). *Slovar' geografičeskich nazvanij zarubežnych stran.* Izd. vtoroe, ispravl. Moskva: Nedra.

sowie das in den folgenden Arbeiten enthaltene Material:
Fedorova (1969), Ivanov (1972), Jakobson (1956), Lockwood (1966), Lomtev (1972), Meredov (1974), Pilch (1967), Saunders (1970), Shapiro (1966), Toporov (1966, 1971), Torsuev (1975), Živov (1973).

Sekundärliteratur

Altmann, G., & Lehfeldt, W. (1980). *Einführung in die Quantitative Phonologie* (Quantitative Linguistics, vol. 7). Bochum: Brockmeyer.

Fedorova, N.I. (1969). O pričinach redkoj upotrebitel'nosti sočetanij "sonornyj + šumnyj" v pozicii načala slova. *Vestnik Moskovskogo universiteta, Serija filologija*, 6, 75–79.

Harary, F., & Paper, H.H. (1957). Towards a general calculus of phonemic distribution. *Language,* 33, 143-169.

Ivanov, V.V. (1972). Fonetika i fonologija sovremennogo russkogo jazyka. II: Sintagmatika i paradigmatika fonem russkogo jazyka. *Russkij jazyk v nacional'noj škole,* 4, 6-17.

Jakobson, R. (1956). Die Verteilung der stimmhaften und stimmlosen Geräuschlaute im Russischen. In *Festschrift für Max Vasmer zum 70. Geburtstag* (S. 199-202), Wiesbaden, Berlin: Harrassowitz.

Kempgen, S. (1995a). *Russische Sprachstatistik. Systematischer Überblick und Bibliographie* (Vorträge und Abhandlungen zur Slavistik, Bd. 26). München: Otto Sagner.

Kempgen, S. (1995b). Phonemcluster und Phonemdistanzen (im Russischen). In Weiss (Hg), *Slavistische Linguistik 1994* (S. 197-221), München: Otto Sagner.

Kempgen, S. (im Druck). Modellbedingte Vorkommensbeschränkungen in der Phonologie. In *Festschrift für Baldur Panzer.*

Lehfeldt, W. (1971). Ein Algorithmus zur automatischen Silbentrennung. *Phonetica,* 24, 212-237.

Lockwood, D.G. (1966). *A Typological Comparison of Microsegment and Syllable Constructions in Czech, Serbo-Croatian, and Russian.* Ph.D., University of Michigan.

Lomtev, T.P. (1972). *Fonologija sovremennogo russkogo jazyka (na osnove teorii množestv).* Moskva: Vysšaja škola.

Meredov, E. (1974). Ierarchičeskaja sistema inicial'nych dvučlennych konsonantnych sočetanij v sovremennom russkom jazyke. *Vestnik Moskovskogo universiteta, Serija filologija,* 3, 49-55.

Pilch, H. (1967). Russische Konsonantengruppen im Silbenan- und -auslaut. *To Honor Roman Jakobson,* vol. II (S. 1555-1584), The Hague, Paris: Mouton.

Saunders, R. (1970). *Phonological Constraints in Russian Syllable Margins.* Ph.D., Brown University.

Shapiro, M. (1966). On non-distinctive Voicing in Russian. *Journal of Linguistics,* 2, 189-194.

Toporov, V.N. (1966). Materialy dlja distribucii grafem russkogo jazyka. In *Strukturnaja tipologija jazykov* (S. 65-143), Moskva: Nauka.

Toporov, V.N. (1971). O distributivnych strukturach konca slova v sovremennom russkom jazyke. In *Fonetika. Fonologija. Grammatika.* (S. 52-162), Moskva: Nauka.

Torsuev, G.P. (1975). *Stroenie sloga i allofony v anglijskom jazyke (v sopostavlenii s russkim).* Moskva: Nauka.

Živov, V.M. (1973). Centr i periferija v fonologičeskoj organizacii slova. I. O sostavlenii jadernogo inventarja grupp soglasnych russkogo jazyka. *Lingvotipologičeskie issledovanija,* vyp. I, č. I, 80-163, Moskva.

The Distribution of Some Syntactic Construction Types in Text Blocks

Reinhard Köhler

Introduction

One of the basic kinds of word repetition in texts is their distribution in text blocks (cf. Altmann, 1988:174ff): A text is segmented into adjacent passages of equal size; in each block, the frequency of the given word is counted. Frumkina (1962) was the first to investigate the number of blocks with x occurrences of a given word, where x is considered a random variable. She started from the assumption that the Poisson distribution is an appropriate model of the corresponding probability; other authors (for details cf. Altmann, 1988:75) used the Normal and the Log-Normal distributions. Later, a theoretical derivation of the *Negative Hypergeometric* distribution was given, empirically tested, and baptised *Frumkina's Law* by Altmann (Altmann & Burdinski, 1982; Altmann, 1988:175ff).

Meanwhile, many investigations of data from several languages have been conducted, and all of them have confirmed the Negative Hypergeometric distribution together with its special cases (the Poisson, Binomial and Negative Binomial distributions) as an appropriate model of the frequency of occurrence of words in text blocks.

The purpose of the present paper is to make a first attempt to investigate the block distribution of linguistic units at other levels, viz. at the syntactic level, which differs – with respect to repetitions – from the word and other 'lower' levels in a fundamental way:

> *When the items of higher levels are assumed, not the repetitions of constructs, but the repetition of their structures is observed and established as a language phenomenon* (Hřebíček, 1999).

Accordingly, for frequency counts of syntactic structures in text blocks, instead of surface material, the occurrence of categories should be observed. As a

consequence, segmentation of blocks and definition of block size must also be based on the occurrence of categories, regardless of the fact that they do not define unambiguous text positions in terms of terminal elements (words). In the present study, two kinds of categories have been observed: clause types (viz. relative, infinitival, participle clauses) and function types (logical direct, indirect, and prepositional objects).

The text corpus used is the Susanne corpus (cf. Sampson, 1995), a collection of 64 English texts with a total of 128,000 running words, which is available in a syntactically analysed and annotated form (cf. Fig 1). The organisational and linguistic information for each word form is given in six columns: The first column (reference field) gives a text and line code, the second (status field) marks abbreviations, symbols, and misprints, the third gives the wordtag according to the Lancaster tag set, the fourth the word form from the raw text, the fifth the lemma, and the sixth the parse. In lines A01:0040j and A01:0050d (example in Fig. 1), for example, the :o's mark the NP "the over-all ... of the election" as logical direct object, the brackets with label Fr in lines A01:0060h and A01:0060n mean that "in which ... was conducted" is a relative clause.

A01:0010a	-	YB	<minbrk>	-	[Oh.Oh]
A01:0010b	-	AT	The	the	[O[S[Nns:s.
A01:0010c	-	NP1s	Fulton	Fulton	[Nns.
A01:0010d	-	NNL1cb	County	county	.Nns]
A01:0010e	-	JJ	Grand	grand	.
A01:0010f	-	NN1c	Jury	jury	.Nns:s]
A01:0010g	-	VVDv	said	say	[Vd.Vd]
A01:0010h	-	NPD1	Friday	Friday	[Nns:t.Nns:t]
A01:0010i	-	AT1	an	an	[Fn:o[Ns:s.
A01:0010j	-	NN1n	investigation	investigation	.
A01:0020a	-	IO	of	of	[Po.
A01:0020b	-	NP1t	Atlanta	Atlanta	[Ns[G[Nns.Nns]
A01:0020c	-	GG	+<apos>s	-	.G]
A01:0020d	-	JJ	recent	recent	.
A01:0020e	-	JJ	primary	primary	.
A01:0020f	-	NN1n	election	election	.Ns]Po]Ns:s]
A01:0020g	-	VVDv	produced	produce	[Vd.Vd]
A01:0020h	-	YIL	<ldquo>	-	.
A01:0020i	-	ATn	+no	no	[Ns:o.
A01:0020j	-	NN1u	evidence	evidence	.
A01:0020k	-	YIR	+<rdquo>	-	.
A01:0020m	-	CST	that	that	[Fn.
A01:0030a	-	DDy	any	any	[Np:s.
A01:0030b	-	NN2	irregularities	irregularity	.Np:s]
A01:0030c	-	VVDv	took	take	[Vd.Vd]
A01:0030d	-	NNL1c	place	place	[Ns:o.Ns:o]Fn]Ns:o]Fn:o]S]
A01:0030e	-	YF	+.	-	.O]

A01:0030f	- YB	<minbrk>	-	[Oh.Oh]
A01:0030g	- AT	The	the	[O[S[Ns:s.
A01:0030h	- NN1c	jury	jury	.Ns:s]
A01:0030i	- RRR	further	far	[R:c.R:c]
A01:0030j	- VVDv	said	say	[Vd.Vd]
A01:0030k	- II	in	in	[P:p.
A01:0030m	- NNT1c	term	term	[Np[Ns.
A01:0030n	- YH	+<hyphen>	-	.
A01:0030p	- NN1c	+end	end	.Ns]
A01:0040a	- NN2	presentments	presentment	.Np]P:p]
A01:0040b	- CST	that	that	[Fn:o.
A01:0040c	- AT	the	the	[Nns:s101.
A01:0040d	- NNL1c	City	city	.
A01:0040e	- JB	Executive	executive	.
A01:0040f	- NNJ1c	Committee	committee	.
A01:0040g	- YC	+,	-	.
A01:0040h	- DDQr	which	which	[Fr[Dq:s101.Dq:s101]
A01:0040i	- VHD	had	have	[Vd.Vd]
A01:0040j	- JB	over<hyphen>all	overall	[Ns:o.
A01:0050a	- NN1n	charge	charge	.
A01:0050b	- IO	of	of	[Po.
A01:0050c	- AT	the	the	[Ns.
A01:0050d	- NN1n	election	election	.Ns]Po]Ns:o]
A01:0050e	- YC	+,	-	.Fr]Nns:s101]
A01:0050f	- YIL	<ldquo>	-	.
A01:0050g	- VVZv	+deserves	deserve	[Vz.Vz]
A01:0050h	- AT	the	the	[N:o.
A01:0050i	- NN1u	praise	praise	[NN1n&.
A01:0050j	- CC	and	and	[NN2+.
A01:0050k	- NN2	thanks	thank	.NN2+]NN1n&]
A01:0050m	- IO	of	of	[Po.
A01:0050n	- AT	the	the	[Nns.
A01:0060a	- NNL1c	City	city	.
A01:0060b	- IO	of	of	[Po.
A01:0060c	- NP1t	Atlanta	Atlanta	[Nns.Nns]Po]Nns]Po]N:o]
A01:0060d	- YIR	+<rdquo>	-	.
A01:0060e	- IF	for	for	[P:r.
A01:0060f	- AT	the	the	[Ns:103.
A01:0060g	- NN1c	manner	manner	.
A01:0060h	- II	in	in	[Fr[Pq:h.
A01:0060i	- DDQr	which	which	[Dq:103.Dq:103]Pq:h]
A01:0060j	- AT	the	the	[Ns:S.
A01:0060k	- NN1n	election	election	.Ns:S]
A01:0060m	- VBDZ	was	be	[Vsp.
A01:0060n	- VVNv	conducted	conduct	.Vsp]Fr]Ns:103]P:r]Fn:o]S]
A01:0060p	- YF	+.	-	.O]

Fig. 1: First sentence of text A01 from the Susanne corpus

The corpus, or rather the grammar according to which the texts are analysed and tagged, differentiates the following clause types:

S	main clause	Ti	infinitival clause
Ss	embedded quoting clause	Tf	"for-to" clause
Fa	adverbial clause	Tb	bare nonfinite clause
Fn	nominal clause	Tq	infinitival relative clause
Fr	relative clause	W	"with" clause
Ff	fused relative	A	special "as" clause
Fc	comparative clause	Z	reduced ("whiz-deleted") relative
Tg	present participle clause	L	miscellaneous verbless clause
Tn	past participle clause		

and the following functions:

Complement Function tags

s	logical subject
o	logical direct object
i	indirect object
u	prepositional object
e	predicate complement of subject
j	predicate complement of object
a	agent of passive
S	surface (and not logical) subject
O	surface (and not logical) direct object
G	"guest" having no grammatical role within its tagma

Adjunct Function tags

p	place
q	direction
t	time
h	manner of degree
m	modality
c	contingency
r	respect
w	comitative
k	benefactive
b	absolute

Other Function tags

n	participle of phrasal verb
x	relative clause having higher clause as antecedent
z	complement of catenative

In the first case, the frequency analysis of clause types, two alternative block definitions were applied:

1. Each syntactic construction was counted as a block element,
2. only clauses were considered as block elements.

In the second case, each functionally interpreted construction, i.e. each function tag in the corpus, was counted as a block element. As the results presented in the next section show, the hypothesis that the analysed categories are block distributed according to Frumkina's law was confirmed in all cases.

Results

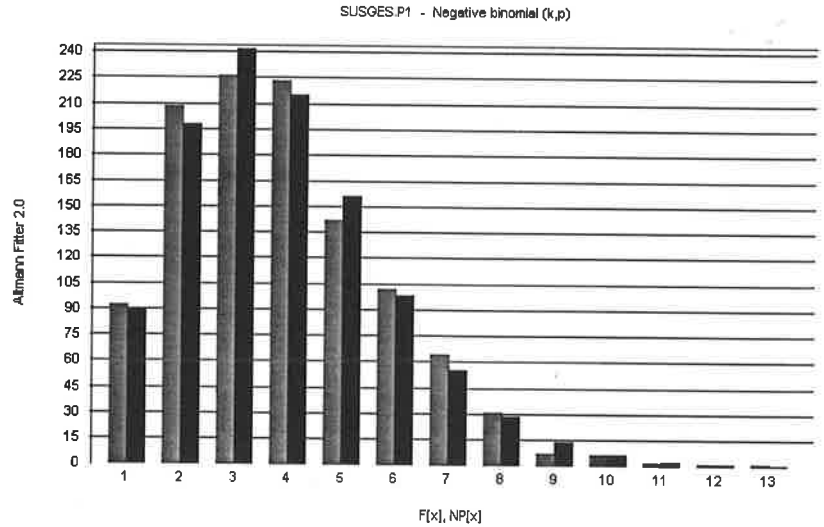
In order to form a sufficiently large sample, the complete Susanne corpus was used for each of the following tests. As types of syntactic constructions are more frequent than specific words, smaller block sizes have been chosen – depending on which block elements were taken into account, 100 or 20, whereas for words, a block size of at least several hundreds is common. All calculations were performed with the help of the Altmann Fitter (Altmann, 1994).

No. of blocks with x occurrences of (present and past) participle clauses

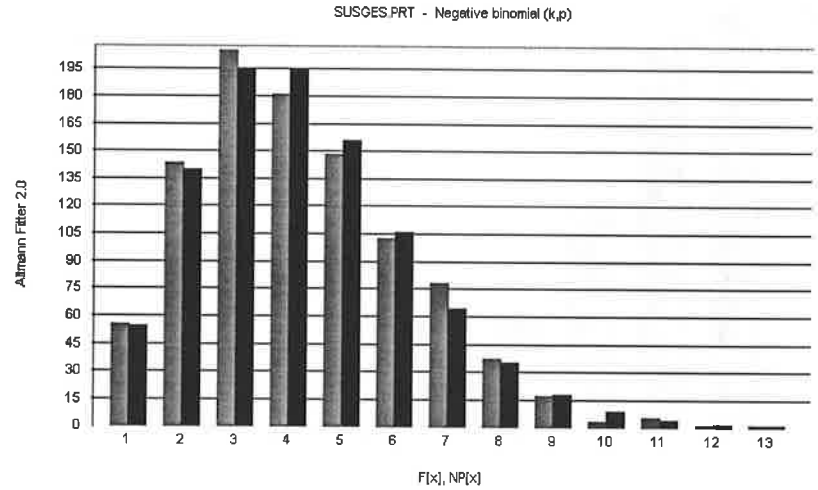
all syntactic constructions			clauses		
$X[i]$	$F[i]$	$NP[i]$	$X[i]$	$F[i]$	$NP[i]$
0	92	90.05	0	55	54.26
1	208	198.29	1	143	139.78
2	226	241.31	2	205	194.66
3	223	214.67	3	181	194.26
4	142	155.77	4	148	155.53
5	102	97.72	5	102	106.12
6	64	54.89	6	78	64.03
7	31	28.26	7	37	35.02
8	7	13.56	8	17	17.68
9	6	6.13	9	3	8.34
10	2	2.64	10	5	3.76
11	1	1.09	11	1	1.58
12	1	0.70	12	1	1.04
k = 9.4115 p = 0.7661 DF = 9 X^2 = 8.36 $P(X^2)$ = 0.50			k = 12.34131 p = 0.7912 DF = 10 X^2 = 9.33 $P(X^2)$ = 0.50		

Distribution: Negative binomial (k, p)
Sample size: 1105
Block size: 100

Distribution: Negative binomial (k, p)
Sample size: 976
Block size: 20



All syntactic constructions



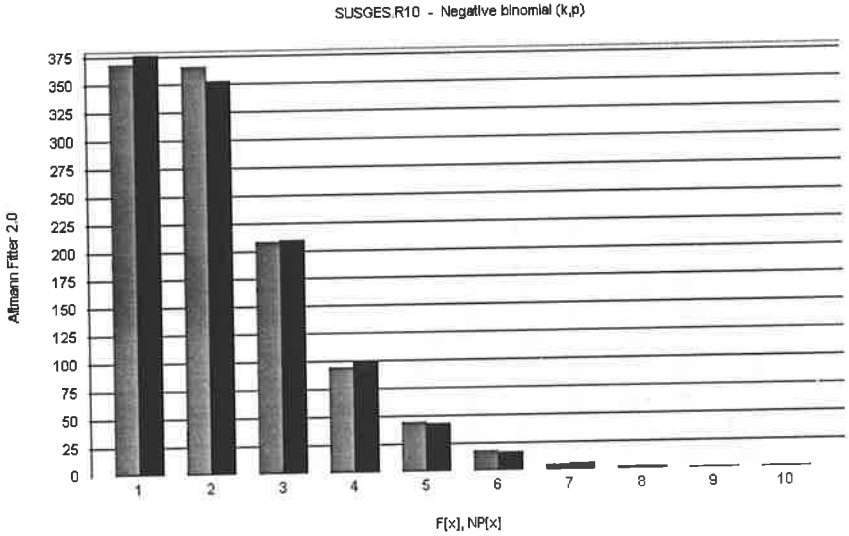
Clauses

No. of blocks with x occurrences of relative clauses

all syntactic constructions			clauses		
$X[i]$	$F[i]$	$NP[i]$	$X[i]$	$F[i]$	$NP[i]$
0	368	376.54	0	105	113.44
1	366	352.73	1	170	164.23
2	208	208.95	2	165	145.31
3	94	99.78	3	92	101.33
4	44	41.93	4	57	61.15
5	17	16.17	5	30	33.46
6	4	5.87	6	12	17.06
7	2	2.03	7	11	8.24
8	1	0.68	8	5	3.81
9	1	0.32	9	3	1.70
			10	1	1.28
k = 3.7781			k = 4.4941		
p = 0.7520			p = 0.6779		
DF = 5			DF = 8		
X^2 = 2.08			X^2 = 8.84		
$P(X^2)$ = 0.84			$P(X^2)$ = 0.36		

Distribution: Negative binomial (k, p)
Sample size: 1105
Block size: 100

Distribution: Negative binomial (k, p)
Sample size: 651
Block size: 30



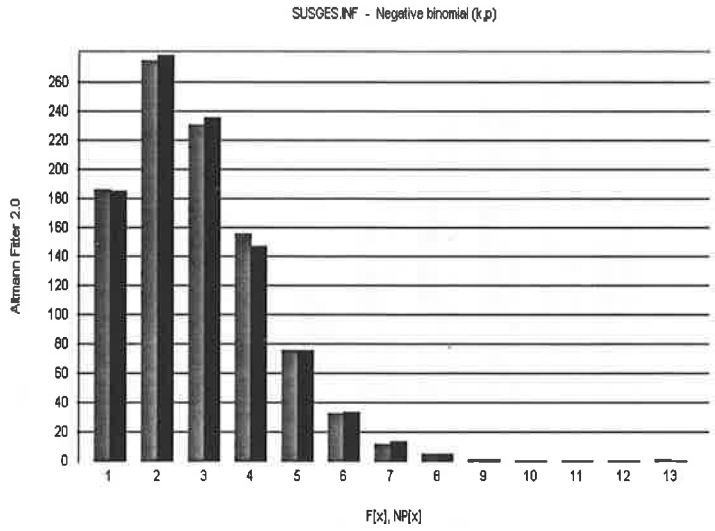
All syntactic constructions

No. of blocks with x occurrences of infinitival clauses

all syntactic constructions			clauses		
$X[i]$	$F[i]$	$NP[i]$	$X[i]$	$F[i]$	$NP[i]$
0	271	264.03	0	186	184.80
1	323	332.99	1	275	278.59
2	248	247.97	2	231	235.37
3	147	141.96	3	156	146.86
4	67	69.05	4	76	75.41
5	30	30.02	5	33	33.73
6	13	12.02	6	12	13.59
7	3	4.52	7	5	5.05
8	3	2.44	8	1	1.76
			9	0	0.58
			10	0	0.18
			11	0	0.056
			12	1	0.02
k = 5.5279			k = 8.2771		
p = 0.7719			p = 0.8179		
DF = 6			DF = 6		
X^2 = 1.44			X^2 = 0.045		
$P(X^2)$ = 0.96			$P(X^2)$ = 0.98		

Distribution: Negative binomial (k, p)
Sample size: 1105
Block size: 100

Distribution: Negative binomial (k, p)
Sample size: 976
Block size: 20



All syntactic constructions/clauses

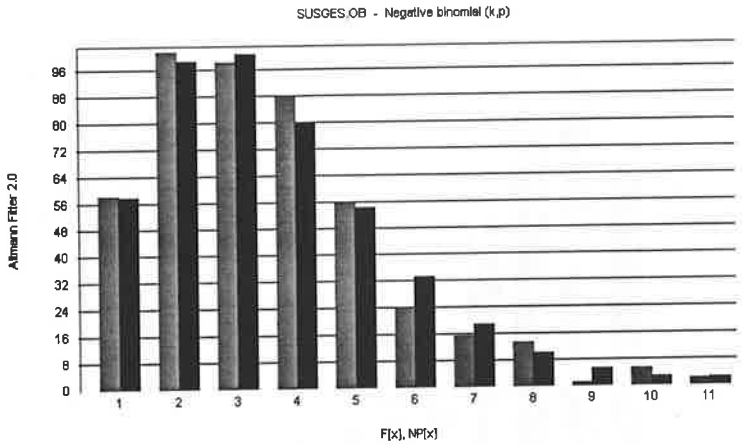
No. of blocks with x occurrences of a prepositional object

all syntactic constructions

$X[i]$	$F[i]$	$NP[i]$
0	58	57.50
1	101	98.22
2	98	100.39
3	88	79.64
4	56	54.08
5	24	33.01
6	15	18.64
7	13	9.92
8	1	5.03
9	5	2.46
10	2	2.13

k = 5.0853
p = 0.6641
DF = 8
X^2 = 11.08
$P(X^2)$ = 0.20

Distribution: Negative binomial (k, p)
Sample size: 461
Block size: 100



All syntactic constructions

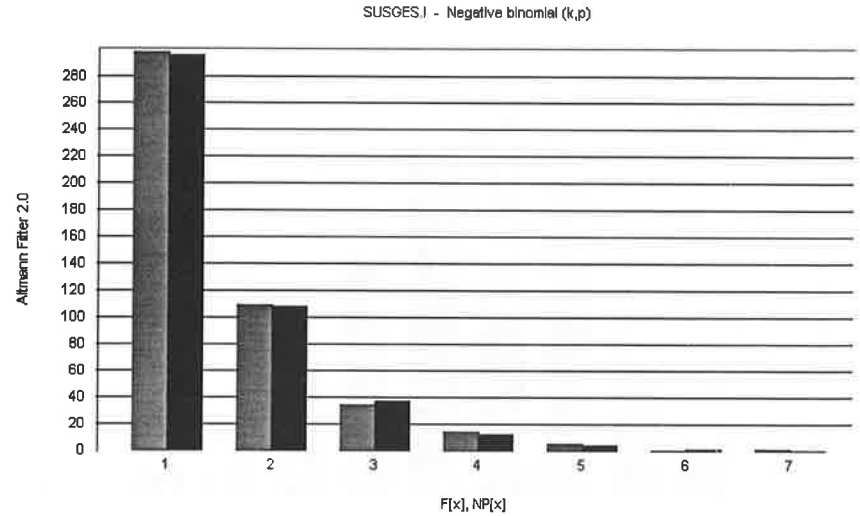
No. of blocks with x occurrences of an indirect object

all syntactic constructions

$X[i]$	$F[i]$	$NP[i]$
0	298	296.92
1	109	108.73
2	34	37.05
3	14	12.31
4	5	4.04
5	0	1.32
6	1	0.63

k = 1.1613
p = 0.6846
DF = 3
X^2 = 1.17
$P(X^2)$ = 0.76

Distribution: Negative binomial (k, p)
Sample size: 461
Block size: 100



All syntactic constructions

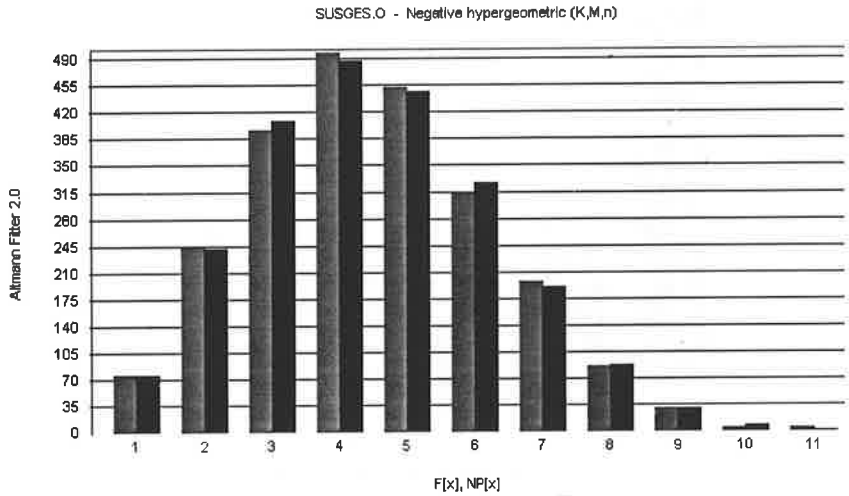
No. of blocks with x occurrences of a logical direct object

all syntactic constructions

$X[i]$	$F[i]$	$NP[i]$
0	76	76.23
1	245	240.32
2	397	408.83
3	497	487.46
4	451	446.56
5	315	326.00
6	198	191.05
7	86	88.43
8	30	30.88
9	5	7.34
10	4	0.90

K = 19.8697
M = 6.9199
n = 10
DF = 6
X^2 = 1.45
$P(X^2)$ = 0.96

Distribution: Negative hypergeometric (K, M, n)
Sample size: 2304
Block size: 20



All Syntactic constructions

Conclusion and prospects

According to the present study, not only words but also categories at the syntactic level follow Frumkina's law. In all cases (with the exception of the logical direct object), the Negative Binomial distribution could be fitted to the data with good and very good Chi-square values. In all these cases, the Negative Binomial distribution yielded even better test statistics than the Negative Hypergeometric distribution. Only the distribution of the logical direct object differs inasmuch the more general distribution, the Negative Hypergeometric d. with three parameters turns out to be the better model (with a $P(X^2)$ of 0.9627).
If future investigations – of other construction types and of data from other languages – corroborate these results, we can conclude:

that Frumkina's law, which was first found and tested for words, can be generalised (as already supposed by Altmann) to possibly all types of linguistic units,
that the probability of occurrence of syntactic categories in text blocks can be modelled in largely the same way as the probability of words.

However, for words, all the four possible distributions are found in general (the Negative Hypergeometric as well as its limiting cases, the Poisson, the Binomial, and the Negative Binomial distributions). As both distributions found in this study for syntactic constructions are waiting time distributions, a different theoretical approach may be necessary[1].

At present, full interpretation or determination of the parameters is not yet possible. Clearly, block size and the simple probability of the given category have to be taken into account but we do not yet know in which way. Other factors, such as grammatical, distributional, stylistic, and cognitive, are probably also essential.

Another open question concerns the integration of Frumkina's law, which reflects the aggregation tendency of the units under study, into a system of text laws together with other laws of textual information flow.

A potential practical application of our findings is that certain types of computational text processing could profit if specific constructions or categories can be differentiated and found automatically by their particular distributions (or, by the fact that they do not follow expected distributions) – in analogy to text characteristic key words.

[1] I would like to thank Gabriel Altmann for this comment.

References

Altmann, G. (1988): *Wiederholungen in Texten*. Bochum: Brockmeyer.
Altmann, G., & Burdinski, V. (1982): Towards a law of word repetitions in text-blocks. In W. Lehfeldt & U. Strauss (Eds.), *Glottometrika 4* (pp. 146-167), Bochum: Brockmeyer.
Hřebíček, L. (1999): Principle of emergence and text in linguistics. *Journal of Quantitative Linguistics*, 6,1, 41-45.
Sampson, G. (1995). *English for the Computer*. Oxford: Clarendon Press.

Software

Altmann-FITTER (1994): Lüdenscheid: RAM-Verlag.

On Quantitative Characteristics of Corpora Approaching Infinite Size

Jan Králík

In his *Lectures on Text Theory, Lesson 4,* Hřebíček (1997) touches upon an important general problem, which is closely related not only to the future of quantitative linguistics, but also to corpus linguistics, and which inspires further work. The following ideas are considered:

> *"Text is an increasing phenomenon... Then, very important question arises: How are basic language principles affected or changed in the process during which a text increases?"* (p. 90)
> *"Can a text be infinitely long? How long a text can be when in the increasing text its structure becomes broken? Text is an architecture moving in time."* (p. 92)
> *"Let us imagine an infinitely increasing text. As the text increases, the set of hapax legomena diminishes to the degree that finally becomes a zero set."* (p. 93)

Hřebíček opens new perspectives and raises questions which should be considered especially with respect to the new possibilities of corpus linguistics, where the size of text corpora can be very large, and where these extremely large text corpora are generally accessible over the computer net. For Hřebíček's mother tongue, thanks to the Czech Grant Agency of the Czech Republic (Grant No. 405/96/K214), free access has been opened to the Czech National Corpus via the internet address http://ucnk.ff.cuni.cz/cnc.

Nonetheless, is it correct to use the term "text" under such conditions? Does a "set of texts" form a new "text" at all? How might we understand the idea of a text increasing towards infinity in corpus linguistics? Can the mathematical sense of infinity be applied in any sensible way to such linguistic data at all? There is no doubt that text, in its classical sense of a recorded author's work, possesses an architecture and inherent structure, which is realised by means of and within text linearity. That is why architecture and structure appear during and through text growth only. However, neither text architecture, nor text

means of sequential addition only. They originate as the consequence of a uniquely formulated idea or intention, or in communication. The natural text growth ends naturally at the end of the text. Even extremely long natural texts no doubt possess architecture and structure *sui generis*. Thus, a purely sequential ordering of any author's texts (the more sequential ordering of texts within corpora) can hardly form any new architecture or structure at a higher level. A text pool can only summarise. As seen in detail, it cannot summarise the dynamic features of texts. The summarisation is always static. It can concern only such properties, which are not directly bound to the inherent text dynamism, or which are its components only. Therefore, the summarisation cannot concern architecture, nor structure in their whole.

From this point of view it seems, in case of sequential addition of texts, we should speak not about "broken" architecture and structure, but, rather about the neglection of their role. The idea of text architecture and structure being neglected or excluded when a corpus is increased, can be exemplified in the case of a successive inclusion of texts to a pool. From a recent experience, reached within related consequences, a couple of more factual considerations can be supplied, as follows.

Until recently, mainly small linguistic data sets have been the subject of statistical analysis. A sample has usually been defined by the size of 1000, 2000 or 3000 words. In order to be objective, it is necessary to say that also in such cases an integral description of the architecture and structure of the author's invention is impossible, although the reasons are fully opposite to those mentioned above: small sample size offers too *fragmented* information about the whole text. In such cases, therefore, the term "broken" architecture and structure would be fully appropriate, as only *elements* of architecture and structure can be observed only.

As to the architectonic and structural elements, the situation in establishing their quantitative characteristics changes totally, if the sample size rises not only to full *natural units* of text, such as books with an average text length between 20,000 and 30,000 running words, but if the sample extent reaches sets of texts used for the compilation of frequency dictionaries - from 500,000 to 1.5 million words, or more, which now reaches text corpora of hundreds of millions words, however. The latter case means a change not only in quantity, but in quality itself. For such a situation, Hřebíček points out one special example:

> *"With a limitlessly increasing text, the equation turns into paradoxical expression."* (p. 163)

Not only one such case exists, but a more general situation is concerned. It is closely related to limit transitions, as they are known in mathematical analysis.

Up to now, statistical research applied to linguistic material has been based on a proportionality between the full text size and the relative values, as they have been expressed by perfectly understandable and interpretable numerical characteristics. In the case of large text corpora, the statistical observation of events on a level lower than architectonic and structural constructions, turns into observation of the statistical limit behaviour of their elements within conditions, which in mathematics are characterised by the increase of an independent variable beyond all limits. With such an increase of linguistic data, naturally, single quantitative characteristics lose not only their local (interval, author's) contacts, but they also lose their ability to be interpreted intuitively. Such characteristics are simply - and only - summarised as limit averages. In such a case, any chance to describe the real architecture and structure of texts vanishes, and in the same situation, the meaningfulness of such characteristics is nearly lost.

Not only the effect of the Law of Large Numbers applies here. Up to now, for shorter texts, only the first part of dependence curves of empirically stated correlations, coincidences and behaviour of identificators of different characteristics has been sufficient, or, in other words, only the beginning of the half-line has been investigated. In future corpus linguistics, taking the summarised texts into account, however, we shall deal with trends over very large distances, which were never discussed in quantitative linguistics before. We shall deal with the limit behaviour, which was never considered nor analysed, as it belonged to unverifiable hypotheses.

The number of the cohered questions increases: Which form of the well-known dependencies and regularities, as those by Zipf-Mandelbrot, Fuchs, Guiraud, Menzerath-Altmann etc. will be valid if data from corpora are studied? How deeply will the present empirical issues have to be modified, revised, or basically re-built? How will the until now valid knowledge of quantitative characteristics, based on shorter text samples and texts, be influenced by such corrections? Will these corrections confirm the traditional knowledge as a part of some higher or more general phenomenon? Will the until now known trends go on in the same way also in the measures *ad infinitum*? Or, on the contrary, will the traditional knowledge be a particular example which holds under specific quantitative circumstances only? Will the new studies confirm the existential connection of architecture and structure with the natural text size? Or, on the other hand, will it be possible to link the until now valid knowledge to the newly discovered issues within limit situations? And, even more generally, will the statistical prediction be applicable to the same extent in corpus linguistics too?

Attempts to answer such questions in their complexity are premature. However, we can prepare contexts for such issues of research. An example concerning another of Hřebíček's ideas is given here. If classes of linguistic events

for which the category *hapax* does not exist, i.e. closed-class, such as phonemes, syllables, morphological categories etc., will be investigated, it is evident that corpus linguistics can their most precise quantitative characteristics. They are, as already mentioned above, mainly below the distinctive level of the author's style (architecture and structure). When open-ended classes of linguistic events, e.g. lexemes and their forms, word combinations, semantic categories, syntactic structures, dependency trees etc., are investigated, the situation will be completely different especially because of the frequency category *hapax*.

It is natural that the number of *hapax legomena* diminishes as the text increases. In the case of an infinitely increasing text, the proportion of the repeated words must necessarily grow faster than the proportion of new words (hapax legomena). It can be disussed that there will be a moment when the set of potential new words becomes empty, i.e. when there will be no hapax legomena and no new words. It seems to be more likely to find such a situation within a static investigation of texts. From such a point of view, for every given instant also corpora are closed, however: from one side being started by the historical moment of the first use and re-writing of the just invented script, from the other side being ended by the present moment of the research.

However, the quantitative research of text corpora not only opens the phenomenon of the enormous text size, but also a completely new situation: corpora too, increase in time, and language develops in time. New words are invented, and any period of time brings new concepts and terms. Therefore, for a natural language, we can never imagine a moment, in which we could be sure that no new word (form, word combination, semantic category, etc.) arises.

Thus, having been inspired by Hřebíček's ideas about *hapax legomena*, we should, one day, more exactly, think not about the "zero set", but about the "set with the measure zero". Such a set consists of a countable number of elements, where "countable" does not necessarily mean "limited" or "finite". The set with the measure zero is negligible against the size of corpora as a whole. The investigation of the limit behaviour of such a set with measure zero will be one of the most interesting issues of corpus linguistics in the future.

References

Bunge, M. (1967). *Scientific Research.* Berlin, Heidelberg: Springer Verlag.

Králík, J. (1978). On the Dispersion and its Computation. In *Prague Studies in Mathematical Linguistics*, 6 (pp. 149-158), Praha: Academia.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8, 1-15.

Hřebíček, L. (1997). *Lectures on Text Theory*. Prague: Oriental Institute.

The Development of Entropy and Redundancy in Czech from the 13th to the 20th Century: Is there a Linguistic Arrow of Time?

Karel Kučera

Introduction

Soon after it was first formulated by R. Clausius around 1850 the concept of entropy gained a very general dimension: it became acknowledged as a measure of disorder, disorganization, homogeneity, degradation of systems and dissipation of energy, and was associated with irreversibility, since according to the second law of thermodynamics entropy in any closed material system can only grow. Assuming that the universe is a closed system, that is, a system with no exchange of matter and energy across its boundary, it was argued that its ultimate state is the state with maximum entropy, the so-called heat death, characterized by the complete dissipation of ordered matter into disorder and of energy into uniform heat. On the same grounds it was argued that increasing entropy "marks the forward flow of time" or "shows the direction of time", and it was therefore called "the arrow of time". (For discussion, broader context and opposing opinions see e.g. Georgescu-Rogen, 1975; Harrison, 1975; Coveney & Highfield, 1990; Bevensee, 1993.)

Given such a cosmological context, it is hardly surprising that the concept of entropy eventually became almost a household word and spread to a number of fields far removed from its birthplace, the physics of macroscopic thermodynamic phenomena. The broad uses of the label *entropy* were both embraced as "fruitful applications" (Harrison, 1975:41) and dismissed as „the kind of sweeping generality people will clutch like a straw," applying it "in various method-starved studies" (Cherry, 1966:216). To Cherry and others, *entropy of languages* was one of the sweeping generalities: when introduced into the theory of information (and then linguistics) the term completely lost its original meaning, as well as any association with time, and it was rather difficult to see why

the new measure of information proposed by Shannon, sharing with the original physical measure of disorder and homogeneity only the formula, should be termed *entropy* at all. Many, like Rudolf Carnap (1977:72), pointed out that "the general identification of entropy (as a physical concept) with the negative amount of information cannot be maintained" and some put forward new terms such as *negentropy* (Brillouin, 1953) or *intropy* (Peters, 1975) to replace the misnomer *entropy* used in information theory. Carnap, aware of the conceptual confusion, set about constructing "an abstract (i.e., purely mathematical) concept of entropy (...) applicable to any system of N elements of any kind which are characterized by quantitative magnitude" (Carnap, 1977:1), but his efforts met with very little understanding among physicists (ibid., foreword:VII), and the end result, his *Two Essays on Entropy*, published posthumously, at the time when the general interest in entropy was ebbing away, has been all but ignored.

Misleading as it may have been, Shannon's term *entropy* persisted in the domain of information theory. The concept of entropy spread to linguistics, where it has been applied with advantage especially in text theory and analysis (cf., for example, Hřebíček, 1997; Tuldava, 1996; Ejiri, Staeheli & Ooaku, 1994) and probabilistic grammars (cf. Halliday, 1991:35). However, Shannon also applied the informational concept of entropy (and redundancy) to language as a whole, proposed a way of measuring it, and measured the entropy and redundancy of English (Shannon & Weaver, 1949:25ff.). Shannon's calculations appeared to many linguists as a promising way of comparing languages with mathematical exactness, leading perhaps to a discovery of some general, perhaps even universal, language characteristics or distinguishing features. For some time, especially in the 1950s and 1960s, measuring the entropy and redundancy of languages was very popular, but the results and their interpretations were somewhat perplexing. Studies showed that entropies and redundancies in different languages - even in different styles or registers of a language - were more or less different, but before the relevance, consistency and the real linguistic meaning of the differences could be ascertained, the phrases *entropy of a language* and *redundancy of a language* practically disappeared from linguistics, leaving an aftertaste of ambiguity and rather limited usefulness.

In our opinion the state of linguistics in the late 1990s may give a glimmer of new hope to entropy and redundancy as measures of certain characteristics in languages. There are at least two facts that could partly revive the lost interest:

1. Language corpora, in various languages, now encompassing tens or hundreds of millions of running words, offer a reliable ground for verification, and perhaps reinterpretation, of the data, conclusions and statements accumulated during the entropy euphoria of the fifties and sixties.

2. Fluctuations and differences in some kinds of entropy or redundancy may be interesting from a new angle represented by synergetic linguistics.

However, it remains to be seen whether entropy and redundancy, even if resurrected, can bear linguistic fruit despite the fact that they have their roots in physics and information theory.

Experiment

This paper presents the results of an experiment focused on the changes in entropy and redundancy of one language in time - a modest echo of the "arrow of time", which seems to have been completely lost during the journey of entropy from physics to information theory to linguistics. The experiment only dealt with first-order entropy and redundancy (i.e. entropy and redundancy of single graphemes and phonemes, not graphemic or phonemic groups, words, etc.) and was designed to provide answers to three questions that seem relevant from the point of view of both linguistics and the concept of entropy itself:

1. Does first-order entropy (H_1) and redundancy (R_1) fluctuate around a certain value or does it increase or decrease in the history of one language?
2. What changes in the language are reflected by the values of first-order entropy and redundancy?
3. Is first-order graphemic entropy (Hg_1) and redundancy (Rg_1) in the history of one language more or less consistently lower than, or higher than, or equal to, first-order phonemic entropy (Hp_1) and redundancy (Rp_1)?

The language analyzed in the experiment was Czech with its history of seven centuries of textual written records. The values of first-order entropy and redundancy were computed according to the formulas

$$H_1 = -\sum p_i \log_2 p_i \qquad \text{and} \qquad R_1 = 1 - \frac{H_1}{\log_2 N}$$

where p_i are frequencies (or, by extension, probabilities of occurrence) of individual graphemes or phonemes and N is the number of the graphemes (N_g) or phonemes (N_p) (for details see Shannon & Weaver, 1949:19ff). The values of p_i, N_g, and N_p were extracted from 16 samples covering the entire history of written Czech; the samples were randomly chosen from texts written or printed around the middle and the end of the 13th through the 20th centuries (the texts are part of the Czech Diachronic Corpus, a section of the Czech National Corpus being built at Charles University in Prague). The samples were 20,000 to 20,008 characters (letters and spaces only, no punctuation, numbers or symbols) long, the only exception the first, oldest sample from the second half of the 13th century, which, with its 4,294 letters and spaces, represents practically all authentic Czech text pre-

served from the period (for a more detailed description of the samples and their transcription see Kučera, 1998). The 20,000 characters are equal to 2,500 to 3,000 words, a size beyond the threshold of representativeness for most quantitative analysis in Czech (for experiments and discussion see Těšitelová, 1980:40ff., 1985:35ff. and 153ff.). Special software was developed for semi-automatic phonological transcription of the textual samples, the resulting phonological transcripts being projections of the present state of knowledge about Old Czech, Middle Czech and Early Modern Czech phonology as presented in Czech historical grammars and related studies (first and foremost Lamprecht, Šlosar & Bauer, 1986, and Komárek, 1969). The values of N_g, N_p, first-order graphemic entropy (Hg_1) and redundancy (Rg_1), and first-order phonemic entropy (Hp_1) and redundancy (Rp_1) in Czech from about 1250 up to the present day are summarized in Table 1 and represented in Figures 1 and 2.

Table 1

TIME	N_g	Hg_1	Rg_1	N_p	Hp_1	Rp_1
1250	27	4.15	0.1277	48	4.87	0.1287
1300	35	4.34	0.1530	49	4.83	0.1394
1350	27	4.20	0.1159	51	4.80	0.1536
1400	31	4.16	0.1606	39	4.64	0.1220
1450	40	4.36	0.1799	40	4.61	0.1335
1500	34	4.34	0.1460	37	4.58	0.1214
1550	40	4.54	0.1461	36	4.56	0.1186
1600	40	4.53	0.1491	38	4.56	0.1318
1650	38	4.56	0.1305	36	4.55	0.1190
1700	39	4.57	0.1347	36	4.58	0.1148
1750	39	4.52	0.1441	36	4.55	0.1197
1800	37	4.50	0.1357	35	4.51	0.1212
1850	37	4.55	0.1261	35	4.54	0.1140
1900	40	4.58	0.1395	37	4.57	0.1227
1950	37	4.54	0.1292	35	4.52	0.1193
2000	38	4.57	0.1297	36	4.56	0.1176

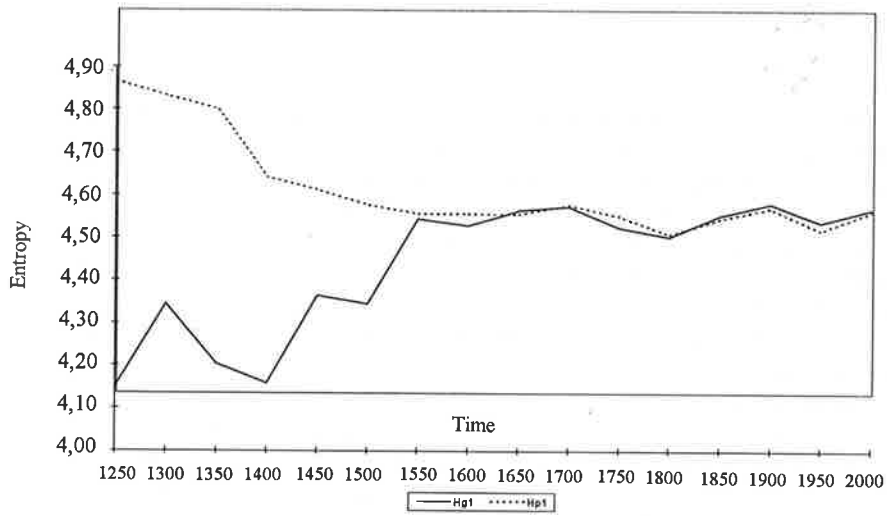


Fig. 1

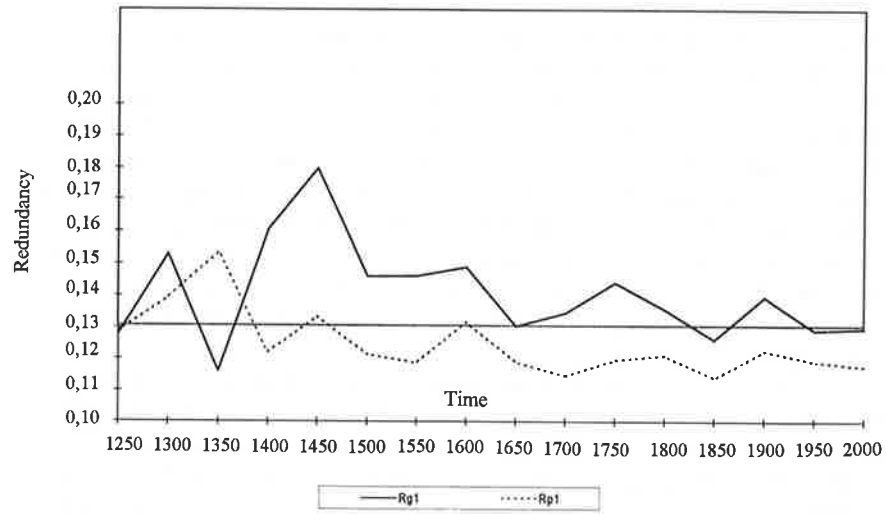


Fig. 2

Discussion

1. First-order entropy

Since 1250, Hg_1 has increased and Hp_1 has decreased in Czech texts, which means that the values of entropy in Czech (as well as many other contemporary languages) gathered in the 1950s and 1960s are generally not so absolute as they may have appeared thirty years ago. It also means that there is no linguistic arrow of time, no linguistic counterpart of the second law of thermodynamics, which would say that the entropy of a language can only increase (or only decrease, or must always remain the same): the results show that at least first-order graphemic and phonemic entropies in at least one language are "arrows" pointing in the opposite directions.

Generally, the quantities of the form $-\sum p_i \log_2 p_i$ primarily measure the *degree of equalization* of frequencies/probabilities of the elements of a system ("any change toward equalization of the probabilities $p_1, p_2, \ldots, p_n$ increases H," Shannon, 1949:21; "entropy is a measure of the degree of homogeneity (...) or equalization of a distribution," Grünbaum, 1975:183), but they do so only in a closed system, which in our case should be interpreted as a system with a constant number of measured elements, i.e. constant N_g, and N_p. If the quantities are used to compare different systems, the reflection of the degree of equalization tends to be blotted out by the much stronger influence of different numbers of their elements. This is usually true when one compares entropies of different languages and this is also true in our case, when we compare entropies of different language states: the decreasing values of Hp_1 have obviously primarily reflected the decreasing number of phonemes in Czech, in particular the gradual disappearance of palatalized labials, sibilants and some other consonants, which took place between 1250 and 1400. After 1400, when the number of Czech phonemes has become constant (although in our text samples N_p has randomly fluctuated between 35 and 40, depending on whether the sample included all the low-frequency phonemes), Hp_1 has remained virtually unchanged, fluctuating between 4.51 and 4.61.

The history of Hg_1 has been influenced by changes of Hp_1 (cf. the drop in both Hp_1 and Hg_1 between 1350 and 1400, and their virtually identical development after 1550), but it has also reflected specific changes in the Czech writing system, which utilized different principles - and different numbers of graphemes - at different times. The low value of Hg_1 in 1250 reflects the so-called primitive writing system, which employed practically only the letters of the medieval Latin alphabet. The number of graphemes, as well as Hg_1, increased around 1300 owing to the introduction of the combinatorial system accompanied by first experiments with diacritics, and decreased again in 1350 when the early diacritics were abandoned. The increase in both the number of graphemes and Hg_1 between 1400 and 1550 was caused by the gradual employment of a new diacritical system which added a number of new graphemes with diacritical marks to the existing repertory of graphemes. The fact that the system, after 1550 firmly established in Czech, was phonological in most of its characteristics is reflected in the above-mentioned virtually identical development of Hg_1 and Hp_1 from 1550 on.

All things considered, one may infer that the history of Hp_1 and Hg_1 found in Czech is language-specific, not universal. Other languages with different numbers of phonemes and graphemes, with a different history of sound changes and changes in their writing systems would no doubt be characterized by different values and histories of Hp_1 and Hg_1. Thus the last of the above three questions can be answered in the negative: evidently, Hg_1 does not have to be consistently lower or higher than, nor consistently equal to, Hp_1 in the history of one language. Our results confirm this inference: although Hg_1 was lower than Hp_1 during most of the history of Czech, at times it was practically equal to and occasionally even slightly higher than Hp_1.

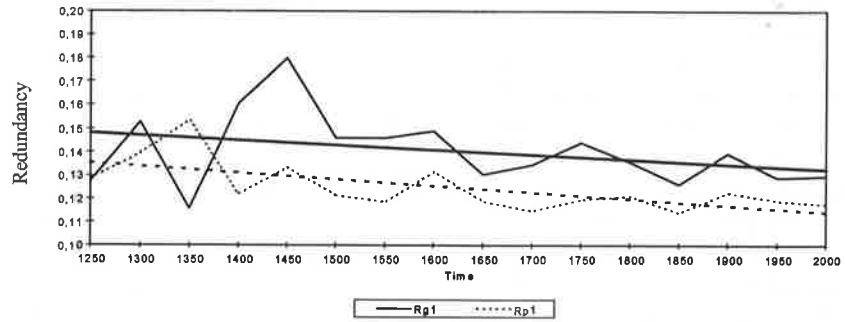
2. First-order redundancy

From 1250 to the present day, Rp_1 and Rg_1 have ranged respectively from 0.11 to 0.15 and from 0.12 to 0.18. One could intuitively expect that the redundancy of the written language would be lower than that of the spoken language, because the latter is obviously confronted with more noise in everyday communication, but contrary to expectations, during the seven centuries phonemic redundancy in Czech has been noticeably and almost consistently lower than graphemic redundancy. Generally higher values of Rg_1 most probably reflect the fact that the Czech writing system has been cultivated and regulated by scribes, printers and later on by linguists to avoid, at the cost of increased redundancy, the homonymy relatively common in the spoken language (virtual absence of such cultivation and regulation in the early stages of the Czech writing system may partly account for the occurrence of the two exceptionally low values of Rg_1 in 1250 and 1350). Since cultivation and regulation (and homophones outnumbering homographs) are typical not only of the Czech writing system, but also of a number of others, generally higher levels of Rg_1 found in Czech may be a more widespread, not strictly language-specific, phenomenon.

The fluctuation of Rp_1 and Rg_1 is larger before 1500 than later on, an obvious result of the extensive changes occurring both in the phonemic and graphemic repertory during the Old Czech period. The fluctuating values of Rp_1 in samples from 1300 and 1500 reflect the impact of a series of major sound changes that were going on at this time (*'u>i, 'ú>í, ó>uo, ý>ej, ú>ou, ie>í*, and *uo>ú* were the most far-reaching) and were causing changes in the degree of equalness of phoneme frequencies (the growing frequency of *í* and the sharply falling frequency of *ó* were among the most conspicuous). The fluctuation of the values of Rg_1 in this period is, of course, influenced by the fluctuation of Rp_1, but it also reflects the rapid, somewhat volatile development in the Czech writing system: Rg_1 is growing when new, diacritical graphemes are being added to the existing

repertory, but being still something of a novelty, they are used only inconsistently in the texts and increase the value of redundancy through their low frequencies (the two peaks reflecting this situation are in 1300 and, above all, 1450). As the novelty wears off and the graphemes with diacritics are used more and more regularly, their frequencies become less exceptional and Rg_1 decreases (cf. the drop between 1450 and 1500; the drop between 1300 and 1350, however, is caused by the above-mentioned fact that the first experiments with diacritics were completely abandoned, so the first, rather rare diacritical graphemes found in the sample from 1300 are completely missing from the sample from 1350).

Considering the fundamental changes in the repertory and use of both the Czech phonemes and graphemes during the seven centuries, it seems rather remarkable that there are no clear-cut increasing or decreasing trends in the history of Rg_1 and Rp_1, or much more sizeable fluctuation: Fig. 3 shows that the values of Rg_1 and Rp_1 lie around two practically parallel, only slightly decreasing lines. (The slight decrease could be, at least partly, explained as a result of the gradually growing number of borrowings causing a gradual increase in frequencies of 'foreign' phonemes and graphemes like *f* or *g*, which were extremely rare in old Czech texts). In our (linguistic rather than informational) interpretation, the values of redundancy resulting from Shannon's formula reflect the limitations placed by a language and its speakers on the use of its elements in a given order: the less equal the frequencies/probabilities of the elements are, the higher is the value of *R* and the less freely the language uses its elements in the given order in units of higher orders. Since redundancy, unlike entropy, avoids the direct overwhelming influence of the number of elements of the system, it seems to be suitable for comparison of different languages or different states of one language and to properly reflect the extent of the above limitations. If we do consider Rg_1 and Rp_1 to be such proper reflections and if we consider the data to indicate that the Czech language is moving neither toward decidedly higher combinatorial freedom of its phonemes and graphemes, nor away from it, we are tempted to say that the fluctuation of Rg_1 and Rp_1 around the two lines may point to the existence of a self-regulating control mechanism participating in the securing of a correct transmission of messages, which is impossible without redundancy (cf. Köhler, 1993:42). The level of redundancy may reflect a dynamic equilibrium between the need to secure the transmission and the force of economy, as well as between stability and adaptation (both the phonemic and graphemic systems seem to have been able to recover the equilibrium relatively soon even after very extensive changes). If Rg_1 and Rp_1 really reflect such an equilibrium, their histories found in Czech may not be completely language-specific.



Conclusion

The data gathered from samples of Czech texts from the 13th to the 20th century showed that the history of Hg_1 and Hp_1 in Czech is language-specific; it does not seem to be characterized by any general or universal trends. Thus, entropy proved not to be a linguistic arrow of time. Changes in Hp_1 and Hg_1 reflected primarily specific changes in the number of Czech graphemes and phonemes at different times; secondarily, they also reflected equalization or differentiation of frequencies of the graphemes and phonemes. In the author's opinion, the employment of entropy, representing a mixture of the two reflections, provided no new insight into the history of the Czech graphemic and phonemic systems.

The history of graphemic and phonemic redundancy in Czech (fluctuation around two practically parallel, only slightly decreasing lines) does not exclude the possibility that Rg_1 and Rp_1 reflect a facet of a self-regulating control mechanism in the language. Also, the fact that in spite of extensive changes in the phonological and writing systems Rp_1 has been almost consistently lower than Rg_1 over the seven centuries raises the question of whether this fact is characteristic only of the Czech language.

References

Bevensee, R M. (1993). *Maximum Entropy Solutions to Scientific Problems.* Englewood Cliffs, N.J.: Prentice Hall.

Brillouin, L. (1953). The negentropy principle of information. *Journal of Applied Physics,* 24, 1152-1163.

Carnap, R. (1977). *Two Essays on Entropy.* Shimony, A. (Ed.), Berkeley, Los Angeles, London: University of California Press.
Cherry, C. (1966). *On Human Communication.* 2nd edition. Cambridge, Massachusetts, London: M.I.T. Press.
Coveney, P., & Highfield, R. (1990). *The Arrow of Time.* London: W. H. Allen.
Ejiri, K., Staeheli, N., & Ooaku, S. (1994). Word Frequency Distribution in Japanese Text. *Journal of Quantitative Linguistics,* 1, 212-223.
Georgescu-Rogen, N. (1975). Bio-economic Aspects of Entropy. In L. Kubát & J. Zeman (Eds.), *Entropy and Information in Science and Philosophy* (pp. 125-142), Amsterdam, Oxford, New York: Elsevier.
Grünbaum, A. (1975). Is the Coarse-grained Entropy of Classical Statistical Mechanics an Antropomorphism? In L. Kubát & J. Zeman (Eds.), *Entropy and Information in Science and Philosophy* (pp. 173-186). Amsterdam, Oxford, New York: Elsevier.
Halliday, M.A.K. (1991). Corpus Studies and Probabilistic Grammar. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics* (pp. 30-43), London, New York: Longman.
Harrison, M.J. (1975). Entropy Concepts in Physics. In L. Kubát & J. Zeman (Eds.), *Entropy and Information in Science and Philosophy* (pp. 41-59). Amsterdam, Oxford, New York: Elsevier.
Hřebíček, L. (1997). *Lectures on Text Theory.* Prague: Oriental Institute.
Komárek, M. (1969). *Historická mluvnice česká I. Hláskosloví.* Praha: SPN.
Köhler, R. (1993). Synergetic Linguistics. In R. Köhler & B. Rieger (Eds.), *Proceedings of the First International Conference on Quantitative Linguistics, QUALICO, Trier 1991* (pp. 41-51). Dordrecht: Kluwer.
Kučera, K. (1998). Vývoj účinnosti a složitosti českého pravopisu od konce 13. do konce 20. století. *Slovo a slovesnost,* 58, 178-199.
Lamprecht, A., Šlosar, D., & Bauer, J. (1986). *Historická mluvnice češtiny.* Praha: SPN.
Peters, J. (1975). Entropy and Information: Conformities and controversies. In L. Kubát & J. Zeman (Eds.), *Entropy and Information in Science and Philosophy* (pp. 61-81), Amsterdam, Oxford, New York: Elsevier.
Shannon, C., & Weaver, W. (1949) *The Mathematical Theory of Communication.* Urbana: University of Illinois Press.
Těšitelová, M. (1980). *Využití statistických metod v gramatice.* Praha: Academia.
Těšitelová, M. (1985). *Kvantitativní charakteristiky současné češtiny.* Praha: Academia.
Tuldava, J. (1996). The Frequency Spectrum of Text and Vocabulary. *Journal of Quantitative Linguistics,* 1, 38-50.

Fractal Structures in Language
The Question of the Imbedding Space

Edda Leopold

Abstract

This contribution deals with the hypothesis of the existence of fractal structures in language which was proposed by Luděk Hřebíček in 1992. I will propose a systems-theoretical point of view in order to define both an imbedding space appropriate to represent linguistic fractal structures and mappings between the observed data and the imbedding space.

Introduction

The hypothesis of fractal structures in language was first formulated by Luděk Hřebíček (see Hřebíček, 1992:91f) as a consequence of his discovery of sign and vehicle aggregations as supra-sentence structures in texts. This ingenious idea, which was derived from the Menzerath-Altmann law, is of great importance for the theory of quantitative linguistics. In recent years Luděk Hřebíček has undertaken further approaches to characterise language phenomena by fractal dimensions. He has considered text as a time-series and analysed it by means of Hurst-exponents (see Hřebíček, 1995, 1997).

In this paper I will examine the hypothesis of fractal structures in language from a mathematical point of view. I shall try to shed light on the implications of the respective mathematical apparatus on linguistic theory and I shall try to answer the question of where fractal structures in language can be found.

Mathematical Characterisation of Fractals

The term fractal is used to refer to a huge class of mathematical objects which cannot be classified as (deformed) lines, surfaces or volumes. These objects do not have an integer valued dimension. They are something in between - neither one, two or three dimensional. Therefore, fractals are characterised by a fractal (i.e. non-integer) dimension.

Usually fractals are subsets of the Euclidean Space $\mathbf{R}^n$. Such a space is called an imbedding space because it imbeds the fractal. (In the final section of this paper I shall show that the fractal interpretation of the Menzerath-Altmann law leads to a fairly abstract imbedding space which does not posses a metric.)

There are several definitions of fractal dimensions. Hausdorff's definition is the oldest and probably the most important. The Hausdorff-dimension has the advantage of being defined for any subset of the imbedding space, which in most cases is $\mathbf{R}^n$. A major disadvantage of the Hausdorff dimension is, that it is hard to calculate in many cases (Falconer, 1990:25).

Let us recall the definition of the Hausdorff-dimension of a fractal set in an imbedding space $\mathbf{R}^n$ (for details see Falconer, 1990:25). The diameter $|U|$ of a non-empty subset U of $\mathbf{R}^n$ is defined as the greatest distance between any pair of points of U:

$$(1) \qquad |U| = \sup\{|x - y| : x, y \in U\}.$$

If $\{U_i\}_{i = 1,2,...}$ is a countable or finite collection of sets of diameter at most δ that cover a set B, we say that $\{U_i\}$ is a δ-cover of B. For any $\delta > 0$ we define

$$(2) \qquad H^s_\delta(B) = \inf\left\{\sum_{i=1}^{\infty} |U_i|^s : \{U_i\} \text{is a } \delta - \text{cover of } B\right\}.$$

Thus we look at all covers of B consisting of sets of diameter at most δ and seek to minimise the sum of the s-th powers of the diameters. As δ decreases the infimum $H_\delta\, s(B)$ increases and so approaches a limit as $\delta \to 0$. We write:

$$(3) \qquad H^s(B) = \lim_{\delta \to 0} H^s_\delta(B).$$

For $\delta < 1$ $H_s(B)$ is a non-increasing function of s. The Hausdorff-dimension of the set B is the critical value D where $H_s(B)$ jumps from ∞ to 0. That is

$$(4) \qquad H^s B = \lim_{\delta \to 0} H^s_\delta = \begin{cases} \infty \text{ if } s < D \\ 0 \text{ if } s > D \end{cases}$$

It is important to mention here that equation (3) contains a limit $\delta \to 0$. So the diameter of the sets U_i, which cover the fractal set, converge to zero. Therefore the imbedding space can never be a disconnected set such as $\mathbf{N}$, because in such a space there is a minimum distance $\delta_{min} > 0$ between its elements, and the limit $\delta \to 0$ is not defined.

From the definition of the Hausdorff-dimension it can be deduced that if a set B is finite or even countable, then its Hausdorff-dimension is zero (Falconer, 1990:29). This means that observed data can never have a Hausdorff-dimension which is different from zero, because we can never make an infinite number of observations. When it is said that some data represents a fractal structure, this is always an idealisation in the sense that the observed structure is extrapolated to the infinitely small.

The central questions one has to ask if the concept of fractal dimension is applied to linguistics are:

1) Is it appropriate to assume that the observed structures can be continued to infinitely small scales?

2) Is it convenient to assume that data is just an empirical manifestation of an idealised phenomenon, which adopts real valued quantities?

Real Model and Mathematical Model

Quantitative raw data on linguistic entities usually consists of finite natural numbers, because data is usually obtained by counting (see Köhler, 1997). When the raw data is aggregated or transformed, for example by calculation of the mean value, the result is no longer integer-valued. This kind of transformation can be interpreted as a mapping from the discrete integer valued space $\mathbf{N}$ into a larger continuous space $\mathbf{R}$.

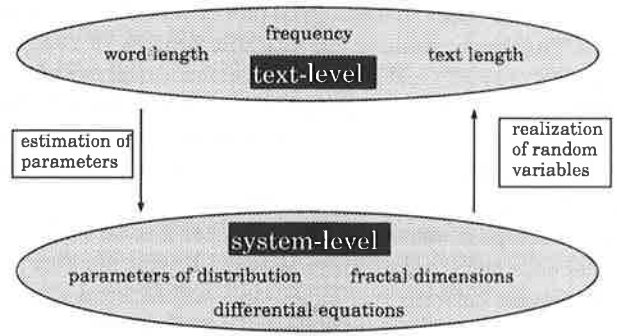


Fig. 1. The relation between the system level and the text level in quantitative linguistic reasoning.

Quantities which can be obtained by counting linguistic entities are elements of the text level. These quantities can be used in order to estimate abstract quantities on the system level, as for instance parameters of probability distributions. The probability distributions on the system level govern the behaviour of random variables which are realised in texts.

The mapping from **N** to **R** does not seem problematic at first sight, but it becomes important if one uses mathematical notions or operations which require topological properties of **R** which are not properties of **N**. This is the case if the fractal dimension is calculated. Due to the limit in equation (3), the notion of a fractal dimension does not make sense if the imbedding space is totally disconnected as in the case of **N**.

Another operation which is also maldefined on the raw data is the derivation, which also involves a limit $\delta \to 0$:

$$(5) \qquad f'(x) = \lim_{\delta \to 0} \frac{f(x+\delta) - f(x)}{\delta}$$

So differential equations set up in quantitative linguistics tacitly imply that linguistic quantities are real valued phenomena. This assumption is obviously not fulfilled if the raw linguistic data itself is considered, but it makes sense if one considers abstract real valued quantities (say parameters of probability distributions) which are estimated by some transformation of the raw data. That is, differential equations cannot directly refer to the text-level. They can only be defined on an abstract system-level - as for example in the paradigm of synergetic linguistics - where the required topological properties can be sensibly assumed.

Maki and Thompson examine the scientific process of model construction. They distinguish between a real model and a mathematical model. The real model is constructed by idealisation and construction of the real world. The mathematical model is obtained by abstraction and formalisation of the real model. The application of mathematical formalisms in the mathematical model leads to conclusions which are compared with the real world.

Maki and Thompson admit that in many cases it is difficult to decide where the real model ends and the mathematical model begins, but they point out that a failure to distinguish between real model and mathematical model is confusing and can lead to wrong conclusions (Maki & Thompson, 1973:4).

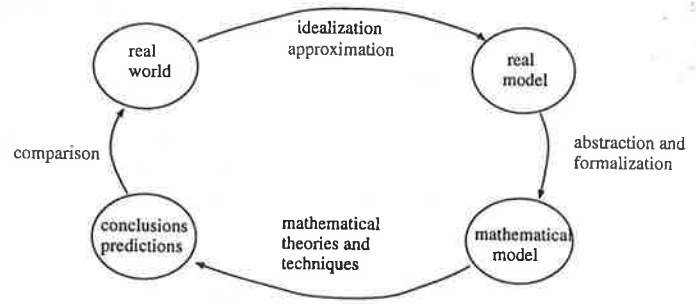


Fig. 2. The relation between real model and mathematical model.

I will draw the distinction between the real model and the mathematical as follows. Quantities obtained by simply counting linguistic entities are elements of the real model. The real model however does not consist of mathematical objects, and mathematical operations do not exist on this level of model construction. These operations are defined only on the level of the mathematical model.

The reason for this distinction is that the set of positive integers **N** is not closed under division and subtraction. The application of division and subtraction of elements of **N** generates the set of rational numbers **Q** which has completely different topological properties from **N**. **N** is totally disconnected in contrast to *Q*.

The distinction between real model and mathematical model corresponds to the distinction between the text-level, which consists of texts and of linguistic entities used in a text, and the system-level, where processes and forces postulated in synergetic linguistics are operating (for details see Leopold, 1998b:12). The raw data obtained by counting linguistic entities in texts belongs to the text level, while mathematical transformations of these data are estimators of abstract theoretical quantities on the system level.

In most cases one may think of an abstract quantity on the system level as the expected value of the respective numbers on the text level. Although the quantities on the text level are always positive integer valued numbers, the quantities on the system level can assume non-integer values. (An example: the points of a dice are integer valued, whereas the expected value (3.5) is not. If one tosses the dice various times the mean value will usually differ from 3.5, but the law of large numbers ensures that it converges to the expected value.)

I will give an example for the difference of the text level and the system level:

Figure 3 displays empirical data on Japanese Kanji (The data was collected by Claudia Prün). Each point represents a Sino-Japanese grapheme with its frequency on the horizontal axis and the number of strokes it consists of on the vertical axis. The number of strokes is always a positive integer. So points of graphemes with coinciding number of strokes and different frequencies form a horizontal line.

Figure 4 characterises the situation on the system level. It was obtained from Figure 3 by adding a uniformly distributed ($U[0;1]$) random variable to each data point. Figure 4 can be interpreted as a two dimensional probability density function. The more points lie in a given area, the more probable is the respective combination of length and frequency in the observed text. The marginal density in the vertical direction represents a grapheme-complexity density for each frequency F. Note that the number of strokes in Figure 4 is a real valued variable L, which does not represent observable values, but the inclination of the language system to adopt a value near L, when a text is produced.

In Figure 5 the mean number of strokes is calculated for each frequency. The displayed numbers do not represent quantities on the text-level because mathematical operations such as summation and division have been involved in their calculation. For each frequency a point in Figure 5 is an estimator of the expected value of the respective marginal grapheme-complexity density.

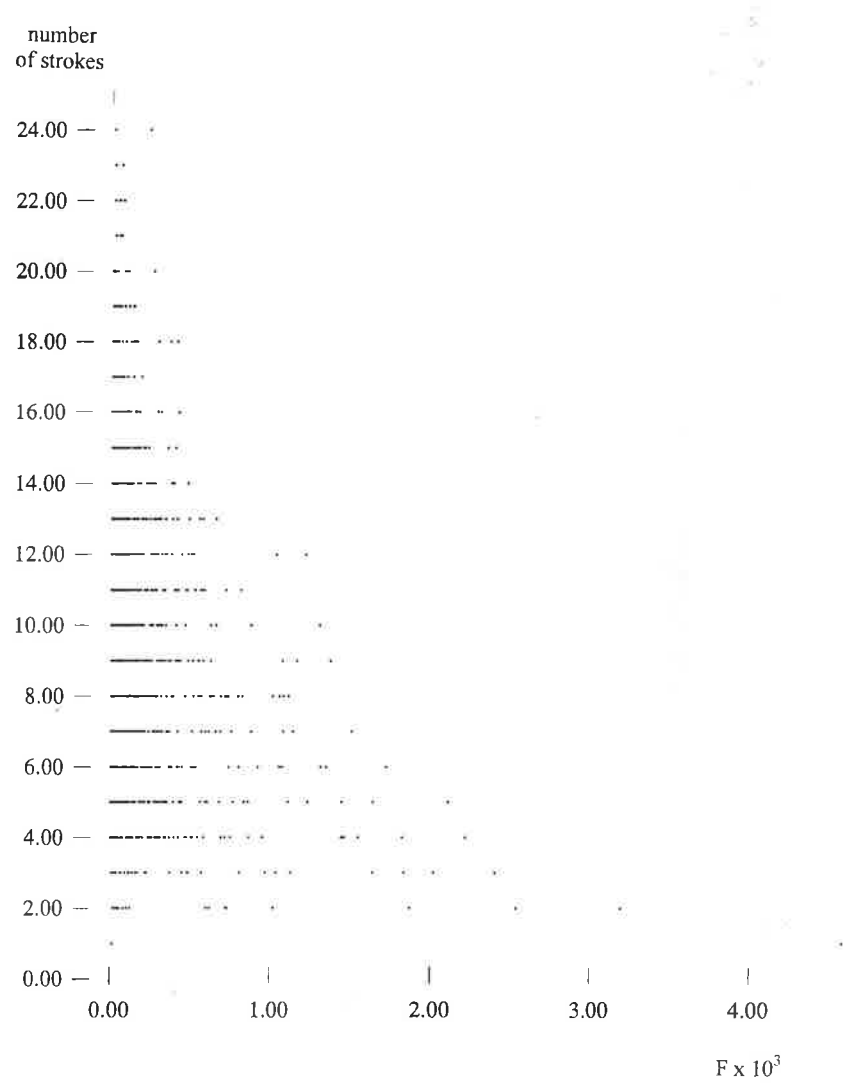


Fig. 3. Grapheme-complexity versus frequency of Japanese Kanji (collected data). Every point in this figure represents a Sino-Japanese grapheme with its frequency and number of strokes.

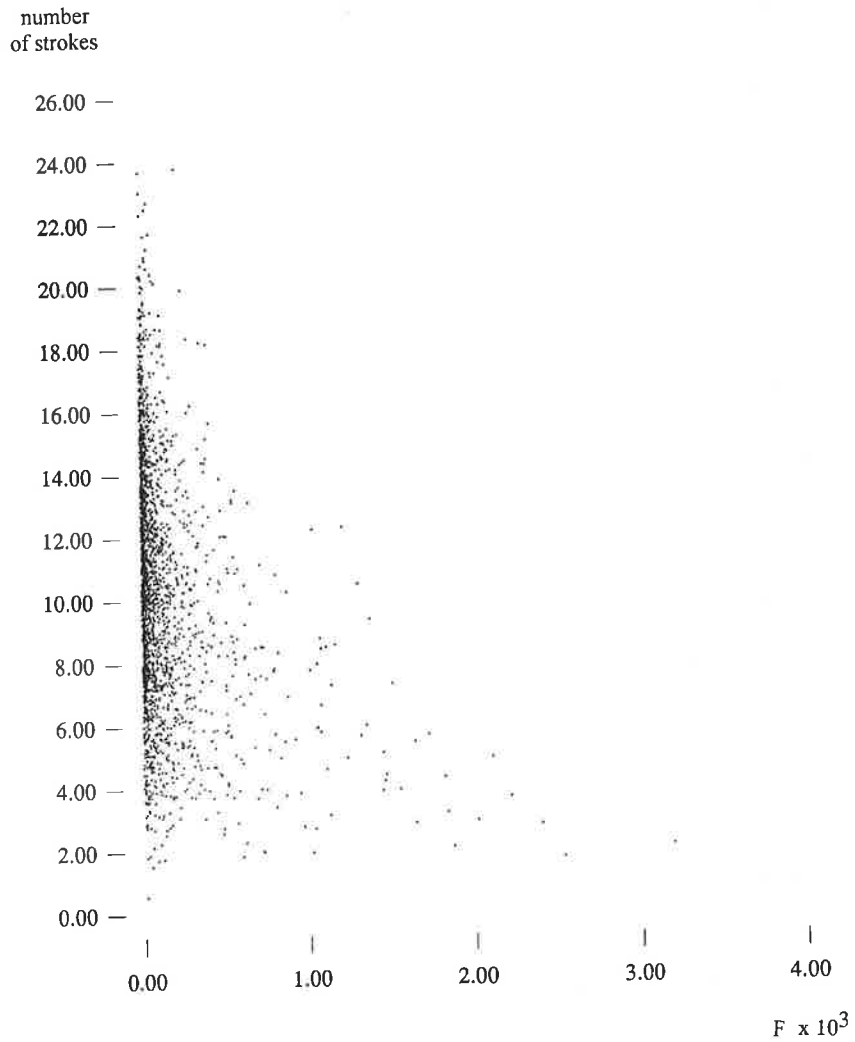


Fig. 4. Grapheme-complexity versus frequency of Japanese Kanji (disturbed data). A uniformly distributed random variable was added to each data point of Figure 3. Figure 4 represents the inclination of the language system to adopt different combinations of frequency and complexity.

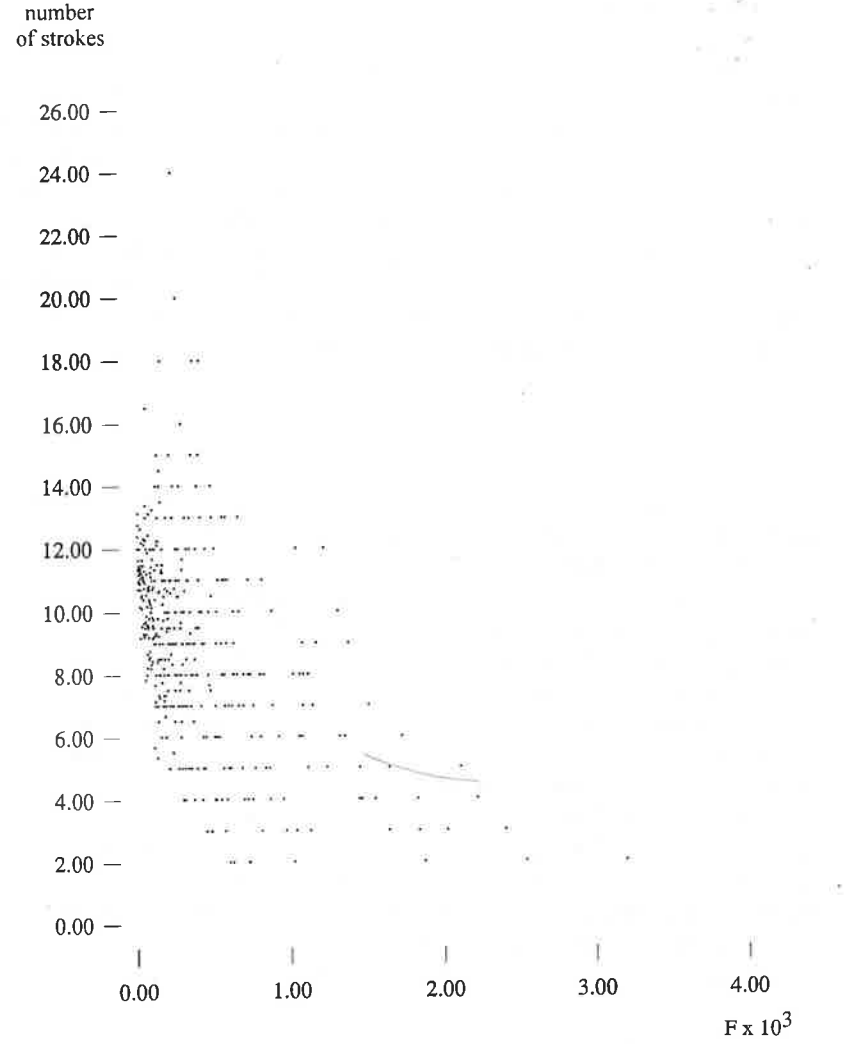


Fig. 5. Grapheme-complexity versus frequency of Japanese Kanji (mean values). For each frequency the average number of strokes is presented on the vertical axis.

Dynamic Systems as a Means of Modelling

If the interrelation of two or more linguistic variables is considered, it is convenient to examine the joint distribution of the respective product space. Figure 4 displays such a distribution. One approach is to analyse the marginal distributions. So in the case of frequency and length one can analyse the frequency spectra for each word length separately, as was done in (Leopold, 1998a) or one can examine the word-length distribution for e.g. all hapaxes.

One may ask if the possible (or more exactly the most probable) combinations of different interrelated variables can be described by means of a dynamic system. So it is possible to set up a partial differential equation which describes the interrelation of frequency and length as follows

$$(6) \qquad \begin{aligned} \dot{\varphi} &= \alpha \cdot \lambda \cdot \varphi \\ \dot{\lambda} &= r \cdot \lambda \end{aligned}$$

Note that the variables λ and φ in this equation do not represent frequency and length themselves. They symbolise the expected values of the probability distributions of length and frequency.

(In the case of frequency the situation is somewhat more complicated, because frequency is dependent on text size. Therefore φ denotes the intensity of a non-stationary Poisson-process. (c.f. Leopold, 1998b:48))

As long as one is dealing with partial differential equations in two dimensions no fractal attractor can occur. For in the plane the range of attractors for continuous systems is rather limited. The only attractors for continuous systems are isolated points or closed loops. (Falconer, 1990:184)

But when three dimensional partial differential equations are considered, it is possible that they exhibit chaotic dynamics and thus converge to strange attractors of fractal dimension. So one could consider frequency, length and polysemy, and extend equation 6 to three dimensions. A famous example of a three dimensional system of partial differential equations is the Lorentz attractor which applies to hydrodynamic problems and is defined by

$$(7) \qquad \begin{aligned} \dot{x} &= \sigma(y - x) \\ \dot{y} &= rx - y - xz \\ \dot{z} &= xy - bz \end{aligned}$$

At present, each continuous dynamic system must be studied individually since there is little general theory available. Attractors of continuous systems are well suited to computer study, and mathematicians are frequently challenged to explain 'strange' attractors that are observed on computer screens (Falconer, 1990:188).

Fractal Dimensions Derived from the Menzerath-Altmann Law

Hřebíček's idea that the validity of the Menzerath-Altmann law indicates fractal structures in language has arisen from the following formula:

$$(8) \qquad x_1 = \left(\frac{A_1}{\left(\frac{A_2}{\left(\frac{A_3}{x_4} \right)^{\frac{1}{b3}}} \right)^{\frac{1}{b2}}} \right)^{\frac{1}{b1}}$$

Hřebíček emphasises the self-similar structure of equation (8). He points out:

> "*In this formula, for example, m = 1 corresponds to phonemes, m = 2 to morphemes, m = 3 to words, and m = 4 to sentences. The fractal character of the sets of constructs and constituents is evident from the shape of the formula which is a formation similar to Japanese puppets.*" (Hřebíček, 1997:104)

The relation between the different levels of analysis in the Menzerath-Altmann law is reminiscent of the generator of the Cantor-dust (see for example Hřebíček, 1995:107). The existence of fractal structures in language seems to be an obvious consequence of the Menzerath-Altmann law, but it is difficult to grasp the hypothesis exactly. So one has to answer the question: what is the imbedding space the fractals are defined on? and what kind of metric or topology is defined on this space?

The problem we have to face in order to derive a fractal dimension from the Menzerath-Altmann law, is neither the absence of a (physical) dimension nor the fact that measurements in quantitative linguistics usually arise from counting procedures (as Köhler, 1997 argued), because one could consider discrete numbers as realisations of random variables and estimate the (real-valued) parameters of their distribution as described above (see Fig. 1). What is needed to derive a fractal dimension from the Menzerath-Altmann Law is a continuous scale of levels of analysis. So we should be able to proceed continuously [!] from the level of

sounds to the level of syllables and further on to the levels of morphs, words, clauses, sentences and supra-sentence structures. Furthermore, if δ denotes the level of analysis in this continuous lattice, then the limiting level of analysis for $\delta \to 0$ has to be defined. From a linguistic perspective this is of course a rather strange idea but it seems to me that this is not too far from Hřebíček's vision when he wrote:

> *"In our empirical argumentation two neighbouring levels are characterised by parameters with values which are only their estimates; they change when a new level is inserted between the two former neighbours. The scheme of linguistic levels in an arbitrary form is nothing but a classification of language units. Any classification represents a relationship between the classifier and its knowledge about linguistic units and their relations."* (Hřebíček, 1995:111)

If the above conditions on the lattice of levels of analysis are fulfilled, the definition of the Hausdorff dimension can be adapted to Hřebíček's idea of fractal structures in texts. Note that in the following presentation the entities become smaller when δ decreases in contast to the notation of Hřebíček which denotes the largest unit by x_1 and the smaller by x_2, x_3, and so on.

Let T be a text. We want to calculate or merely define the fractal dimension of T. Let S be the imbedding space of T, i.e. $T \subset S$. Clearly S cannot be the Euclidean space $\mathbf{R}^n$. So let us define S as the set of all possible texts in a given environment. Each element of S consists of a stream of pre-theoretical physical events $h(t)$ which are produced at a (physical) instant of time t. So $h(t)$ adopts values in a space X of physical events. Let $[0;t_{max}]$ be the time-span of text production. Then S can be written as the product space $S = X \times [0; t_{max}]$. If we assume X to be a metric space then it is not difficult to define a metric on S. But the usual definitions of distances on spaces like S are not useful for our purposes.

Therefore I skip the first step (equation (1)) in the definition of the Hausdorff dimension and replace it directly by a definition of what is meant by a δ-cover $\{U_i\}_{i=1,2,...}$ of a text:

A δ-cover $\{U_i\}$ of a text is the collection of all entities at the δ-level of analysis, which can be found in a text. I call these entities δ-level-entities for short. Clearly the diameter of a δ-level-entity is δ, and the unit of all δ-level-entities covers the whole text T.

Now the definition of the Hausdorff dimension can be applied in a straightforward manner. For any level of analysis we define

$$(10) \qquad H_\delta^s(T) = \inf\left\{\sum_{i=1}^{\infty} |U_i|^s : \{U_i\} \text{is a } \delta\text{ - level - entity of } T\right\}$$

This can be reduced to

$$(11) \qquad H_\delta^s(T) = n_\delta \cdot \delta^s$$

where n_δ denotes the number of δ-level-entities in the text T. The Hausdorff measure of a text is therefore

$$(12) \qquad H^s(T) = \lim_{\delta \to 0} H_\delta^s(T) = n_\delta \cdot \delta^s$$

Finally the Hausdorff dimension of a text is that number $D > 0$ which ensures that

$$(13) \qquad n_\delta \cdot \delta^s \to \left\{\begin{matrix} \infty \text{ if } s < D \\ 0 \text{ if } s > D \end{matrix}\right\}$$

as $\delta \to 0$. One can say more simply: The Hausdorff dimension of a text is that positive real number D where the number of δ-level-entities increases with exactly the same speed as δ^D decreases if δ approaches zero.

Now the notion of a fractal dimension of a text is mathematically well-defined. The only question left is: How do we quantify the step from one level of analysis (say clauses) to another (say words)? Does it correspond to a multiplication by 0.5, or 0.1, or even 0.01? Every choice of this factor is inevitably eclectic, but it has a considerable effect on the result of the calculation of the fractal dimension.

References

Falconer, K. (1990). Fractal Geometry. Chichester et al.: Wiley & Sons.

Feder, J. (1988). Fractals. New York, London: Plemum.

Hřebíček, L. (1992). Text in Communication: Supra-Sentence Structures. Bochum: Brockmeyer.

Hřebíček, L. (1995). Text Levels Language Constructs, Constituents and the Menzerath-Altmann Law. Trier: Wissenschaftlicher Verlag Trier.

Hřebíček, L. (1996). Word Associations and Text. In P. Schmidt (Ed.), Glottometrika 15 (pp. 96-101), Trier: Wissenschaftlicher Verlag Trier.

Hřebíček, L. (1997). Persistence and Other Aspects of Sentence-Length Series. Journal of Quantitative Linguistics, 4, 103-109.

Köhler, R. (1997). Are there Fractal Structures in Language? Units of Measurement and Dimensions in Linguistics. Journal of Quantitative Linguistics, 4, 122-125.

Leopold, E. (1998a). Frequency Spectra within Word Length Classes. Journal of Quantitative Linguistics, 5, pp.224-231.

Leopold, E. (1998b). Stochastische Modellierung lexikalischer Evolutionsprozesse. Hamburg: Kova.

Maki, D.P., & Thompson, M. (1973). Mathematical Models and Applications. Englewood Cliffs (N.J.): Prentice Hall.

Sentence Length and Sentence Structure as Statistical Characteristics of Style in Prose

Viktor V. Levickij, Oksana O. Pavlyčko, Tatyana G. Semenyuk

Introduction

The quantitative characteristics of the authorial and functional styles have been in the focus of linguists' attention for a long time (see Yule, 1939; Lesskis, 1963; literature survey is in Perebijnis, 1967; Golovin, 1971, 1974). Based on the works of the four authors (H. Böll, H. Kant, Th. Mann and E.-M. Remarque) this paper analyses two characteristics of the authors' styles: sentence length and the frequency of different types of sentences.

It was assumed that these two characteristics would be able to differentiate the peculiarities of style of the authors examined. One thousand sentences were written out from the works of each author by the method of consecutive selection to study sentence length, together with six hundred complex sentences to study the frequency of subordinate clause use.

Sentence length

The study of sentence length can be found in L. Hřebíček's works (Hřebíček, 1992; Hřebíček, 1995). L. Hřebíček is right in stating that the correlation between sentence length and some other text characteristics is regulated by the so-called Menzerath-Altmann law (see Hřebíček, 1997:103).

G.U. Yule (Yule, 1939) found that sentence length is an important characteristic of an author's style. It is also evident that sentence length depends, in its turn, on sentence structure. It has been proved that there is a direct dependence between sentence length and the number of complex sentences in the text (Lesskis, 1963:106). Therefore to compare the length of simple and complex sentences and also their frequency correlation in the text, it is necessary, first, to make a list of the main types of syntactic structures called sentences. In accordance with the classifications of sentences made by different German linguists

which can be found in many German grammar books all the sentences examined belong to the following types:

a) simple (e.g. *Das Parkett drehte sich langsam um uns. Die Geige und das Cello erhoben sich zu einer sanften Kantilene über das raunende Orchester.* (Remarque));

b) compound (e.g. *Es war schon kühl abends, und in den Fenstern waren dicke Bündel Silberfüchse und warme Mäntel für den Winter ausgestellt.* (Remarque));

c) complex (e.g. *Und hinaus rauschte sie, indem sie die Schultern ein wenig emporzog und den Kopf zurückwarf.* (Th. Mann) *Das ganze Leben macht ihm keinen Spaß mehr, wenn er diese Sache verlassen muß.* (Remarque));

d) compound-complex (e.g. *Es fehlte nicht an einer Flasche Rotwein, welche vor dem Hausherrn stand, denn Herr Grünlich frühstückte warm.* (Th. Mann)).

The data on the average number of the word-forms in each of the four above-mentioned types of sentences are presented in Table 1.

Table 1
The average length of sentences of different types

Types of sentences	*Authors*				
	Böll	Kant	Mann	Remarque	Average
Simple	8.2	8.9	8.7	6.4	8.05
Compound	13.9	20.3	17.6	13.5	16.3
Complex	21.2	19.9	24.9	13.1	19.3
Compound-Complex	38.6	35.2	37.9	29.1	35.2
Average	20.5	21.1	22.3	15.5	19.9

The data show that sentences having the longest average length most frequently occur in the works by Mann, followed by the works by Kant and Böll. Remarque's style is characterized by the use of the shortest sentences. The length of the sentences in Böll's works is the closest to the average, whereas in the works by Remarque and Mann it is the farthest from the average.

Frequency distribution of the sentences of different types

The frequency of use of syntactic structures (sentences) of different types is shown in Table 2.

Table 2
Frequency distribution of sentences of different types

Sentence type	*Authors*				
	Böll	Kant	Mann	Remarque	Total
Simple	348	324	490	696	1858
Compound	228	224	162	150	764
Complex	168	186	224	104	682
Compound-Complex	256	266	124	50	696
Total	1000	1000	1000	1000	4000

First of all we want to find whether the difference between the frequencies in Table 2 is significant. This can be done with the help of the χ^2-test using the formula

$$(1) \qquad \chi^2 = \Sigma \frac{(O-E)^2}{E}$$

where O is the observed frequency and E the expected frequency. The results of the analysis are: $\chi^2_9 = 448.7$; $\chi^2_{0,01;9} = 21.67$, which shows that the difference between the frequencies in Table 2 is significant.

The data also show that there is no significant difference between the frequencies of the compound and compound-complex sentences as they vary within the range of 700, whereas the frequency use of the simple sentences is much higher than that of each of the other three types of sentences. The hypothesis about the frequency divergence of simple and other types of sentences can be proved or rejected by the χ^2-test. For this purpose it is necessary to re-arrange the data in Table 2 in order to compare the frequencies of simple sentences with those of the other types (see Table 3).

Table 3
Frequency of different types of sentences

Types of sentences	Authors				
	Böll	Kant	Mann	Remarque	Total
Simple	348	324	490	696	58
Others	652	676	510	304	2142
Total	1000	1000	1000	1000	4000

The results of the analysis are: χ^2_3=351.99; $\chi^2_{0.01;3}$=11.34. As is well-known, the significant value of χ^2 only shows that the divergence between the frequencies analysed is significant. But the value of χ^2 cannot measure either the connection or the differences between the investigated features. This can be measured with the help of the Chuprov contingency coefficient *K* (see Urbach, 1964:359, 360) as its formula contains the value of χ^2. The following formula is used for multi-cell tables:

$$(2) \qquad K = \frac{\sqrt{\chi^2}}{\sqrt{N\sqrt{(r-1)(c-1)}}}$$

where *N* is the number of sentences studied (in Table 2 *N* = 4 000), *r* the number of rows, and c the number of columns.

With the help of this formula, it is possible to find in which table the contingency of features is higher and in which of them the divergence between the empirical frequencies is greater. For Table 2, *K* = 0.193; and for Table 3, *K* = 0.225. The results of another (additional) analysis of frequency distribution of each of the three types of sentences (compound, complex and compound-complex) in Table 2 in comparison with all the others (*N* in all cases = 4,000) has shown that χ^2 = 108.43; df = 6; *K* = 0.144 and that in the comparison [compound sentences] - [the others] - χ^2_3 = 0.892; [complex sentences] - [the others] - χ^2_3 = 0.974 and [compound-complex sentences] - [the others] - χ^2_3 = 2.42. Thus, the major difference between the styles of the four writers lies in the use of simple and composite sentences. Evidently each author favours certain types of syntactic constructions. This hypothesis can also be verified with the help of the χ^2- test and a contingency coefficient. However, the procedure of the statistical analyses must be changed. Using the multi-cell Table 2, it is necessary to make alternative four-cell tables of frequency distribution (see Table 4), the coefficient Φ for which be calculated by formula (3)

$$(3) \qquad \Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}},$$

where *a*, *b*, *c* and *d* are empirical frequencies in a four-cell table.

J.Tuldava (1988:157-158) has shown that while comparing the data of different experiments (*N* being the same for all of them) it is more expedient to use the so-called *normalizing coefficient* Φ norm., using the formula

$$(4) \qquad \Phi_{norm} = \frac{ad - bc}{n * \min(b,c) + (ad - bc)}$$

than the coefficient Φ calculated by formula (3).

As our task is to find a significant relation between the semantic elements in the alternative tables, but not to compare the value of this relation, formula (2) can be used for this purpose. The results of the statistical analysis are presented in Table 5. The statistically significant values are given in Table 5 (the significance is calculated by the the χ^2 value).

Table 4
Frequency distribution of different types of sentences in the works by Remarque and the other authors

Types of sentences	*Authors*		
	Remarque	Others	Total
Simple	696	1162	1858
Others	304	1838	2142
Total	1000	3000	4000

χ^2_1 = 287.27; Φ = 0.268

The distribution of all the frequencies in Table 2 was studied in an analogous way. The value of the contingency coefficient Φ is determined by the χ^2 value. Statistically significant Φ values are presented in Table 5 (Note: only the cases in which actual frequencies exceeded theoretically expected frequencies were taken into consideration).

Table 5
Statistically significant values of contingency coefficients Φ

Types of setences	*Authors*			
	Böll	Kant	Mann	Remarque
Simple				0.268
Compound	0.054	0.048		
Complex			0.082	
Compound-Complex	0.125	0.14		

From Table 5 it follows that Remarque's and Mann's styles are characterized by one of the four syntactic features: Remarque's by simple sentences, and Mann's by complex ones, whereas the typical feature of Böll's and Kant's styles is the usage of compound and compound-complex sentences. In other words, Remarque prefers to use simple sentences, Mann prefers complex, whereas Böll and Kant give preference to compound and compound-complex sentences.

To study the degree of stylistic divergence of each of the four authors from a certain *average norm*, the χ^2-test can be used by comparing the frequencies in the works of this or that author to the average values obtained by using the data in Table 2 (see Table 6 as a model).

Table 6
Frequency distribution of different types of sentences in Remarque's works compared to the average values.

Types of sentences	*Authors*	
	Remarque	Average values
Simple	696	464.5
Complex	150	191
Compound	104	170.5
Compound-complex	50	174
Total	1000	1000

The results of the statistical analysis are as follows:

Remarque: $\chi^2_3 = 135.86$; $K = 0.198$;

Böll: $\chi^2 = 35.62$; $K = 0.10$;

Kant: $\chi^2 = 47.57$; $K = 0.117$;

Mann: $\chi^2 = 18.65$; $K = 0.07$; N in all the cases is 2,000.

Based on the frequency use of sentences of different types, we have come to the conclusion that the most original style is that of Remarque ($K = 0.198$), whereas Mann's style is the least original of all the four authors ($K = 0.07$).

Frequency distribution of different types of subordinate clauses

It can be shown that the length of a sentence is a function of its structure. As the complex sentence possesses the greatest variety of different structural-semantic types, it is expedient to examine the frequency distribution of different types of subordinate clauses in the works of the four authors. The results of the analysis are presented in Tables 7 and 8.

Table 7
Frequency use of different types of subordinate clauses

Types of clauses	*Authors*				
	Böll	Kant	Mann	Remarque	Total
Subjektsätze	16	12	6	16	50
Prädikativsätze	2	4	2	24	32
Objektsätze	124	146	96	134	500
Attributsätze	194	158	282	166	800
Temporalsätze	122	82	92	74	370
Lokalsätze	42	6	18	8	74
Kausalsätze	18	30	12	28	88
Finalsätze	4	6	2	8	20
Vergleichssätze	20	42	20	28	110
Bedingungssätze	28	84	20	66	198
Modalsätze	10	2	18	12	42
Konzessivsätze	6	8	16	8	38
Konsekutivsätze	14	20	16	28	78
Total	600	600	600	600	2400

The data in Table 7 show the frequency distribution of the thirteen types of subordinate clauses examined. As was expected, the most frequently used types of subordinate clauses are attributive and objective ones. On the whole, the frequency distribution in Table 7 is disproportionate, as shown by the χ^2-test: $\chi^2_{36} = 272.86$; $\chi^2_{0.01;36} = 58.6$.

Table 8
Statistically significant values of contingency coefficients Φ

Types of clauses	*Authors*			
	Böll	*Kant*	*Mann*	*Remarque*
Subjektsätze				
Prädikativsätze				0.134
Objektsätze		0.05		
Attributsätze			0.167	
Temporalsätze	0.079			
Lokalsätze	0.131			
Kausalsätze		0.04		
Finalsätze				
Vergleichssätze		0.067		
Bedingungssätze		0.121		0.058
Modalsätze			0.055	
Konzessivsätze				
Konsekutivsätze				0.046

The data in Table 8 show that each of the four authors gives preference to different types of subordinate clauses: Mann to subordinate clauses of time and place, and Kant to those of comparison, condition, cause and objective. The frequency of predicative subordinate clauses in the works by Remarque is much higher than in the works of the rest of the authors studied.

The results of the statistical analysis presented in Table 7 show not only the distinctions but also the similarities in the styles of the four authors. To study the degree of this similarity a correlational analysis was used. The results of the analysis are in Table 9.

Table 9
The values of the correlational coefficients

Authors	Böll	Kant	Mann	Remarque	Total
Böll		0.898	0.842	0.925	0.976
Kant			0.830	0.982	0.950
Mann				0.891	0.957
Remarque					0.977

$df = 11$; $r_{11;0.01} = 0.68$

The data in Table 9 show that the greatest similarity in frequency is in different types of subordinate clauses between Remarque's and Kant's styles ($r = 0.982$) and those of Böll and Remarque ($r = 0.925$).

If we make the column *Total* a certain «norm», then it would be possible to calculate the correlation between this and the frequencies in each of the four authors' styles. The values of the corresponding correlation coefficients are presented in Column 6 of Table 9. These values are approximately equal in Böll and Remarque's works. The lowest coefficient value recorded in Kant's works.

Thus, the degree of similarity of the four authors' styles by the frequency use of different types of subordinate clauses use does not coincide with the degree of similarity by the frequency of simple and complex sentences use (see Section 3).

Conclusions

Having studied a vast text corpus in the works of many Russian writers, G. Lesskis in the earlier-mentioned article came to the conclusion that sentence structure is a function of the author's style. "It seems possible", he writes, "to determine the dependence between sentence length and the content of the text." (Lesskis, 1963:106, 107). If the term *content* means the same as the term *content* traditionally used in the study of literature, than G. Lesskis's statement seems to be too audacious. In fact, as is seen from G. Lesskis's article, (p.109ff), by the term *content of the text* he meant only the manner of writing.

The results of the research also make it possible for us to come to certain conclusions about the manner of expressing *content* in the works of the four authors.

The characteristic feature of Böll's style is a very high frequency of elaborated sentences, with a preference for compound and compound-complex sentences (see Table 5). The elaborated sentences in his writings are very long (the average length (of a sentence) is 38.6 word-forms). The individual parameter of H. Böll's style is the usage of subordinate clauses of place and time (see Table 8).

In H. Kant's works the frequency of compound sentence use is higher than that of complex sentences. Most compound sentences are very long (20.3 word-forms). The most frequently used subordinate clauses are those of comparison and condition (see Table 8).

Th. Mann's writings abound in long complex sentences mainly with attributive and subordinate clauses (see Tables 5 and 8).

In contrast to Kant and Mann, Remarque prefers to use simple and short sentences, as the action in his works is very dynamic and fast developing. It is not by chance that the observed frequency of predicative clause use in Remarque's works is higher than that theoretically expected (see Table 8).

References

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

Golovin, B.N. (1971). *Jazyk i statistika.* Moscow: Prosveščenie.

Golovin, B.N. (1974) (Ed.). *Voprosy statističeskoj stilistiki.* Kiev: Naukova Dumka.

Hřebíček, L. (1992). *Text in Communication: Supra-Sentence Structures.* Bochum: Brockmeyer.

Hrebícek, L. (1995). *Text Levels. Language Constructs, Constituents and the Menzerath-Altmann Law.* Trier: WVT.

Hrebícek, L. (1997). Persistence and Other Aspects of Sentence-Length Series. *Journal of Quantitative Linguistics*, 4,103-109.

Lesskis, G. A. (1963). O zavisimosti meždu razmerom predloženia i charakterom teksta. *Voprosy jazykoznania*, 3, 92-112.

Perebijnis, V.S. (1967) (Ed.). *Statystyčni parametry styliv.* Kyiv: Naukova Dumka.

Tuldava, J. (1988). O primenenii koefficientov sopražonnosti v lingvistike i avtomatičeskij analiz teksta. In *Quantitative linguistics and automatic text analysis*, Nr. 827 (pp. 146-162), Tartu.

Urbach, V.Ju. (1964). *Biometričeskie metody.* Moscow: Nauka.

Yule, G. U. (1939). On a Sentence Length as a Statistical Characteristics of Style in Prose. *Biometrika*, 30, 363-390.

Texts

Böll, H. (1958). Und sagte kein einziges Wort. Erzählungen. Moskau: Verlag für fremdsprachige Literatur.

Kant, H. (1987). Bronzezeit. Berlin: Aufbau-Verlag.

Kant, H. (1989). Schöne Elise. Leipzig.

Mann, Th. (1959). Buddenbrooks: Verfall einer Familie. Moskau: Verlag für fremdsprachige Literatur.

Mann, Th. (1973). Lotte in Weimar. Berlin und Weimar.

Remarque, E.-M. (1975). Drei Kameraden. Moskau: Verlag für fremdsprachige Literatur.

Remarque, E.-M. (1981). Im Westen nichts Neues. Moskau: Verlag für fremdsprachige Literatur.

Word Length as an Indicator of Semantic Complexity

Jaan Mikk, Heli Uibo, Jaanus Elts

The topic of the present study arises from a crucial problem in the field of readability formula application. Most of the formulae include word length as an independent variable to predict text complexity. Advocates of the formulae are convinced that the length of a word indicates its semantic complexity (Klare, 1988). On the other hand, the formulae are questioned by many critics who suggest that word length is too superficial a characteristic of text complexity.

There is some evidence that word length is a valid indicator of the word's semantic complexity. Tuldava (1998) has recently found a standardised regression coefficient -0.949 between a word's age and its length. According to his data, the total correlation coefficient between frequency and length was -0.683. Word frequency is a well-known indicator of semantic complexity - more frequent words in a language are, as a rule, better known to people. In another investigation, we have found a correlation coefficient of 0.96 between the abstractness of Russian nouns and their length (Elts, 1995). The correlation coefficient between the nouns' average tendency to be used as a technical term and their length was 0.86 in this study. Abstract words and terms are semantically complex.

The aim of our paper is to investigate the relationship between word length and semantic complexity in Estonian. We will study the problem on two levels: on the word level and on the text level.

Experiment 1

Method

Here we describe semantic complexity of nouns by indices of abstractness, terminological use, and frequency. The first two indices are the main characteristics, the third is an indirect indicator of noun complexity. Frequency is related to the familiarity of a word - an important characteristic of word complexity.

Abstractness of nouns was evaluated by human experts on the following three-stage scale:

1. nouns signifying directly observable objects (e.g. *bicycle, horse*),
2. nouns signifying observable activities and phenomena (e.g. *run, sunshine*),
3. nouns signifying non-observable notions (e.g. *atom, subject*).

Terminological use was also assessed by human experts according to the following scale (Elts, 1992):

1. nouns in everyday use which are not technical terms (e.g. *window, breakfast*),
2. technical terms which have the same meaning in everyday language (e.g. *velocity, biotechnology*),
3. nouns, which are not used in everyday language (e.g. *transcription, electron*).

The frequency of the nouns was taken from the frequency dictionary of Estonian fiction (Kaasik et al., 1977). We grouped the nouns under consideration into the following three categories:

0. frequent words (frequency over 5, approximately 2000 of the most frequent Estonian words),
1. words with frequency 3 to 5, approximately 2500 additional words fromthe dictionary,
2. infrequent words (frequency 2 or less in the dictionary).

The 454 nouns under consideration were taken from Estonian physics and biology textbooks for middle grades.

Results and discussion

The statistical characteristics of the nouns in our sample are given in Table 1.

Table 1
Mean values of the characteristics of the analysed nouns

Characteristic	*Mean*	*Standard deviation*
Noun length in letters	6.62	2.59
Abstractness	1.98	0.82
Terminological use	1.19	0.49
Frequency	0.89	0.89

From the table it can be seen that average values of abstractness and frequency were near the centre of the scale (2 and 1 accordingly). The index of terminological use was near to the lowest point of the scale. Most of the analysed words (85%) were not terms.

We used linear correlation coefficients as indicators of the relationship between noun length and semantic complexity. The coefficients are shown in Table 2.

Table 2
Correlation coefficients between noun length and mean semantic complexity.

	Length	*Abstractness*	*Terminological use*
Abstractness	0.60		
Terminological use	0.20	-0.04	
Frequency	-0.95	-0.66	-0.10

The table shows that noun length has a very high correlation with the mean frequency index. The words which are more frequent are shorter. Frequency explains about 90% of the variety in noun length. Frequent nouns are more familiar than rare nouns and therefore semantically less complex. As a consequence, semantic complexity, which is related to noun frequency, can be described by noun length.

The correlation between noun abstractness and length is statistically significant at the 0.07 level: Length increases with abstractness and the abstractness explains about 36% of the variety in length. Abstract nouns are more complex and the complexity is to some degree described by noun length.

The third indicator of the noun's semantic complexity - its terminological use - had no statistically significant correlation with word length or other indicators of semantic complexity (Table 2). This result can be explained by the fact that the standard deviation of the characteristic was much smaller than the standard deviation of other characteristics (Table 1).

In conclusion, we can say that in this sample of Estonian nouns, length had a very high correlation (-0.95) with mean frequency. The correlation between noun length and mean abstractness was moderate. Consequently, word length is a good indicator of semantic complexity.

Experiment 2

Method

In this experiment, 30 students studied 40 texts on biology and answered questions on the content of the texts. On the basis of the resulting data, several readability formulae were elaborated.

The texts for the experiment were taken from popular scientific texts in Estonian. The average length of the texts was about 1500 characters. There were no illustrations in the text.

The testees were seventh and eighth grade students (14-16 years old). All testees had to study all the texts, but the total number of answers was about 900 instead of 1200.

The testees studied the texts independently. They had no possibility of consulting their peers or teachers. After studying, they gave the texts to the teacher and received a set of content questions. There were eight free response questions in a set and four different sets for every text. The percentage of correct answers (post-testscore) was considered as an indicator of text simplicity. The post-test score was used in regression analysis as a dependent variable.

All the texts were computer-analysed using the Estonian morphological analysis program and the computer dictionary which included indices of abstractness and of terminological use of nouns (Mikk, 1991; Uibo, 1995). Scales for the assessment of abstractness and terminological use were the same as described in the first part of the paper. Here we will use the percentage of abstract nouns instead of the mean abstractness of nouns since the first had a somewhat higher predicting value. For the same reason we will use the percentage of words with more than seven letters instead of mean word length. The percentage of long words is a better predictor of text readability than mean word length (Elts & Mikk, 1996).

A student's post-test score heavily depends on his/her ability level, including general intelligence, the knowledge of the subject, reading ability, etc. To characterise the level of abilities, we used the mean post-test score for every testee.

Results and discussion

We had about 900 cases of sets of test results. Every case formed a row in our data table including characteristics of the student, the text, and of study effectiveness: altogether about 90 characteristics. We calculated linear correlation coefficients between the characteristics, using SPSS 7.0 for Windows. Some of the correlations are shown in Table 3.

Table 3
Correlations between post-test score and its predictors

	Post-test score	*Ability*	*Words7*	*Abstr3*	*WL 2000*
Ability	0.42				
Words 7	-0.39	-0.02			
Abstr3	-0.36	0.01	0.63		
WL2000	0.37	0.02	-0.77	-0.36	
MTerm	-0.25	-0.01	0.37	0.13	-0.35

Ability - mean post-test score of the student (his/her level of abilities)
Words7 - percentage of words of 7 or more letters
Abstr3 - percentage of nouns with abstractness 3
WL2000 - percentage of words from the wordlist of the 2000 most frequent Estonian words
Mterm - mean tendency of terminological use of the nouns.

In Table 3 we see that word length (Words7) has high correlations with word frequency (WL2000) and the percentage of abstract nouns. The percentage of words with high frequency in the language explains 59% of the percentage of long words in the text. The percentage of abstract words explains 40% of the variation in word length. Mean word length in the text is highly related to the semantic complexity of the text.

Our main question is whether word length can replace the other indicators of semantic complexity in the readability formula. The other indicators we used are word frequency, noun abstractness and terminological use. The indicators have high content validity - rare words in language, abstract nouns, and terms are as a rule difficult for readers. In Table 3 we saw that word frequency and noun abstractness are highly related to the indicator of word length in text, however the correlations are far from maximum - acorrelations[1] are 0.6 and 0.8 correspondingly. Consequently, in spite of high correlations between word length in text and semantic complexity, the question remains as to whether mean word length is as good an indicator of semantic complexity of a text as the other indicators under consideration.

To solve this problem, we used a stepwise multiple regression analysis. The results of the analysis are shown in Table 4.

[1] Acorrelation is a measure of lack of connexion, calculated by the formula $K = \sqrt{(1 - r^2)}$, where r is coefficient of correlation. See Sepetliev (1968:242).

Table 4
Standardised coefficients in regression analysis for predicting the post-test score

Model No	*Predictor variables*					*R*	R^2	*SE*
	Ability	Word7	Abstr3	WL2000	MTerm			
2	0.422	-0.408				0.589	0.347	19.8
4	0.427	-0.024	-0.278	0.262		0.636	0.405	18.9
5	0.427		-0.287	0.277		0.636	0.404	18.9
6	0.426		-0.291	0.230	-0.132	0.648	0.420	18.7

R - coefficient of multiple correlation
SE - standard error of estimate

The first two variables included in the formula were the testees' ability level and the percentage of long words. The next two variables were the percentage of abstract nouns and the percentage of frequent words. After including the last two variables (Model 4), the coefficient of the indicator of word length (-0.024) was statistically insignificant, and the program removed the predictor variable from the model (Model 5). Then the program included the last indicator of semantic complexity - mean terminological use of nouns - in the formula.

It can be seen that the percentage of long words is the best single indicator of the semantic complexity of text. If the other indicators of semantic complexity are included in the formula, then word length has no influence on text readability. The whole influence of mean word length can be reduced to semantic complexity.

Is mean word length as good an indicator of semantic complexity as the others used in our analysis? To answer the question, we will compare models 2 and 6 (Table 4). Model 2 includes only one indicator of semantic complexity - the percentage of words of 7 or more letters. Model 6 includes three indicators of semantic complexity - the percentage of abstract nouns, the percentage of frequent words in the language,and the terminological use of nouns. The multiple correlation coefficients are 0.589 for the first formula, and 0.648 for the second. Is the second formula statistically significantly more powerful in measuring text readability than the first?

To compare the effectiveness of the two formulae, we can compare the confidence intervals of the two multiple correlations. The results of the calculations are given in Figure 1.

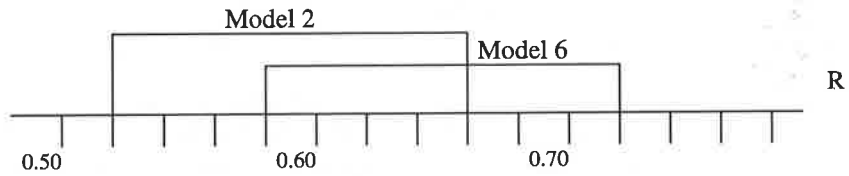


Fig. 1. Confidence intervals of the multiple correlations of the two models from Table 4.

In Figure 1 we see that the confidence intervals of the two multiple correlation coefficients overlap considerably. The two models are statistically equally effective. Word length (Model 2) is as effective as the percentage of abstract nouns, the percentage of frequent nouns, and terminological use in corpora.

The data in Table 4 support the idea that the percentage of long words in a text is a very good indicator of its semantic complexity. At the same time we see in the table that R-square is 0.42 for Model 6 and 0.35 for Model 2. The description of post-test results is about one fifth better if we use word frequency, noun abstractness and terminological use instead of word length.

This last idea is supported by the formulae developed on the data of other experiments. We consider here readability formulae which treat the indices of word length and other indices of semantic complexity simultaneously. In some of these investigations noun abstractness, word frequency, or other indices of semantic complexity were included in the formula whereas word length was not included (Mikk, J. 1991; Kukemelk, & Mikk, 1993). In other cases word length and some other indicators of semantic complexity were included in the formula (Elts, 1992). The results seem to support the idea that word length is not a perfect indicator of semantic complexity. Semantic complexity is obviously a multi-sided phenomenon and can not be perfectly described by a single indicator.

General discussion and conclusion

To compare the findings from the investigations above, we will give the main results in Table 5. We will also include the data from our earlier research in the table.

Table 5
Correlations between word length and other indices of semantic complexity

Content of texts	*Language*	*Number of cases*	*Correlation coefficient with*		
			abstract-ness	*Terminological use*	*frequency*
Biology and Physics	E	10	0.60	0.20*	-0.95
Biology	E	900	0.63	0.37	-0.77
Popular-scientific	E	30	0.32	-	0.49**
Biology	R	19	0.96	0.86	-
Biology	R	40	-	-	-0.60

* Statistically nonsignificant correlation.
** An assessment of unknown words was used in this investigation.

The correlations in the rows of Table 5 differ considerably from each other. The differences may be explained by differences between the text samples, the languages and by the small number of cases. For example, in the first line of Table 5, 90% confidence intervals of the correlation coefficient 0.60 are 0.0 - 0.9. If the number of cases is large, the confidence intervals do not differ so much. Nevertheless, the average relationship between the observed characteristics can be found as a mean tendency in many investigations.

In Table 5 we see that all the indices of semantic complexity had statistically significant correlations with word length in all the investigations except one case. The correlations were found in the investigations which used word groups as units and texts as units in data sets for correlation analysis. High correlations were found between word frequency in language and length, and noun abstractness and length. Word length is a good indicator of semantic complexity.

Returning to Table 4, we can say that the percentage of long words is as good an indicator of semantic complexity as abstractness, terminological use and frequency of words in corpora as far as readability formulae are concerned. Some other research supports the finding but still other readability formulae include word length beside other indicators of semantic complexity of text.

We can conclude that mean word length is a surprisingly good indicator of the semantic complexity of a text. Nevertheless other indicators of semantic complexity should be investigated and included in the readability formula. The other indicators are easily computable if computer programs have access to sources of information such as frequency dictionaries, dictionaries of word abstractness, terminological use.

References

Elts, J. (1992). A readability formula for texts on biology. In V. Rimsa (Ed.), *Psychological problems of reading* (pp. 42-44), Vilnius: Martynas Mazvydas National Library of Lithuania.

Elts, J. (1995). Word length and its semantic complexity. In J. Mikk (Ed.), *Family and Textbooks* (pp. 115-126), Tartu: University of Tartu.

Elts, J., Mikk, J. (1996). Determination of optimal values of text characteristics. *Journal of Quantitative Linguistics*, 3, 144-151.

Kaasik, Ü., Tuldava, J., Villup, A., & Ääremaa, K. (1977). Eesti tänapäeva ilukirjandusproosa autorikõne lekseemide sagedussõnastik. (Frequency dictionary of Estonian fiction (Words in non-conversational material)). *Acta et Commentationes Universitatis Tartuensis*, 413, 5-139.

Klare, G. R. (1988). The formative years. In B.L. Zakaluk & S.J. Samules (Eds.), *Readability, its past, present and future* (pp. 14-34), Newark, Delaware: IRA.

Kukemelk, H., Mikk, J. (1993). The prognosticating effectivity of learning a text in physics. In G. Altmann (Ed.), *Glottometrika 14* (pp. 82-103), Trier: WVT.

Mikk, E. (1991). A morphological analysis program for the Estonian language. *Acta et Commentationes Universitatis Tartuensis*, 926, 103-111.

Mikk, J. (1991). Studies on teaching material readability. In *Problems of textbook effectivity* (pp. 34-50), Tartu: University of Tartu.

Sepetliev, D. (1968). *Statistical methods in medical scientific research*. Moscow: Medicina.

Tuldava, J. (1998). Investigating causal relations in language with the help of Path Analysis. *Journal of Quantitative Linguistics*, 5,3, 256-261

Uibo, H. (1995). Computer readability analysis of Estonian texts. In *Family and Textbooks* (pp. 96-114), Tartu: University of Tartu.

Die Satzlängenverteilung in literarischen Prosatexten der Gegenwart

Brigitta Niehaus

0. Modellierung von Satzlängen

Bei der Erstellung von Texten sind für den Autor neben einigen Vorüberlegungen (Thema, Gattung, Länge, intendierter Hörerkreis,...) vor allem die grammatikalisch richtige Struktur und inhaltliche Aspekte wichtig. Die Gestaltung der Satzlänge erweist sich dagegen als weitgehend unbewußter Prozeß. Obwohl der Autor der Satzlänge in der Regel keine Beachtung schenkt, ergeben sich dennoch „anständige Häufigkeitsverteilungen", die sich dadurch erklären lassen, daß Sprache ein selbstregulierendes System ist und „sich die Länge im Laufe der Texterzeugung selbst organisiert" (Köhler & Altmann, Manuskript:8).

Ausgehend von der Annahme, daß alle Spracheigenschaften eines Textes gesetzmäßig verteilt sind und diese Gesetze sich aus dem Zusammenwirken bestimmter Faktoren der Texterzeugung ableiten lassen, wurden Modelle für die Satzlängenverteilung entwickelt (Altmann, 1988a:148).

Diese Modelle sind u.a. von der Wahl der Einheiten abhängig, in denen die Satzlänge gemessen werden soll. Dies geschah bisher in der Regel entweder mit Hilfe der unmittelbaren Konstituenten, der Clauses (Altmann, 1988a; Niehaus, 1997; Strehlow, 1997; Wittek, 1995), oder durch mittelbare Konstituenten, die Anzahl der Wörter (Altmann, 1988a; Best, 1998); denkbar wären aber auch Messungen der Satzlängen mit noch kleineren Einheiten wie Silbe oder Morphem. So ist es bei der Messung der Satzlänge in Clauses in deutschen Texten vor allem die 1-verschobene Hyperpoisson-Verteilung (Strehlow, 1997), die die Satzlänge gut modelliert, aber auch die positive negative Binomialverteilung (Niehaus, 1997) hat sich als durchaus geeignet erwiesen.

Bei der Verteilung der Satzlänge, gemessen in der Anzahl der Wörter pro Satz, haben sich bislang hauptsächlich zwei Modelle herauskristallisiert, die weiter überprüft werden müssen.

Altmann begründet zunächst die Hyperpascal-Verteilung als geeignetes Modell, da die Störung, die womöglich dadurch entsteht, daß die Länge der Sätze

damit nach ihren indirekten Konstituenten gemessen wird, hier Berücksichtigung findet (Altmann, 1988b:63). Die vorläufige Bestätigung dieser Hypothese erfolgte anhand von 245 Textdateien, zumeist aus dem Altgriechischen, an die sich in der überwiegenden Mehrzahl die Hyperpascal-Verteilung gut anpassen ließ (Altmann, 1988a:162 ff). Die Häufigkeitsklassen wurden dabei in Intervallen von 1-5, 6-10,...Wörtern pro Satz zusammengefaßt.

Best (1998c) dagegen zeigt, daß sich die Hyperpascal-Verteilung im Deutschen nicht besonders gut eignet; ihre Überprüfung an 25 Texten lieferte insgesamt keine guten Ergebnisse. Best schlägt deshalb seinerseits für das Deutsche die 1-verschobene negative Binomialverteilung als vorläufiges Modell vor. Im Unterschied zu Altmann wertet Best zunächst alle Textdateien ohne Zusammenfassung der einzelnen Längenklassen aus und prüft erst im zweiten Schritt die zusammengefaßten Daten, wiederum in Intervallen von 1-5, 6-10,...Wörtern pro Satz.

Bei der Anpassung der negativen Binomialverteilung an Textdateien ohne Zusammenfassung der einzelnen Längenklassen weichen nur zwei Texte von der von ihm erwarteten Verteilung ab; die Anpassung der negativen Binomialverteilung an die Dateien mit zusammengefaßten Längenklassen liefert ein ähnlich gutes Ergebnis.

Ein Ziel dieser Arbeit ist es nun, die Datenbasis der deutschen Texte zu vergrößern, um die Ergebnisse Bests für das Deutsche weiter zu testen. Es soll aber auch überprüft werden, ob die Hyperpascal-Verteilung für dieses weitere Korpus nicht doch brauchbare Ergebnisse liefert.

Ausgewählt wurden 20 abgeschlossene literarische Texte der Gegenwart, definiert als Zeitraum nach 1945. Elf Texte sind der Kinder- und Jugendliteratur zuzurechnen, die übrigen neun gehören zur Kurzprosa für Erwachsene.

Die hier untersuchten Texte wurden bereits als Datenbasis für die Modellierung der Satzlänge nach der Anzahl der Teilsätze verwendet (Niehaus, 1997). Damit ergibt sich im Gegensatz zu den bisherigen Untersuchungen die Möglichkeit des Vergleichs. Es bietet sich die Chance, eventuelle Zusammenhänge zwischen der Verteilung der Satzlänge in Teilsätzen und der Satzlänge in Wörtern ansatzweise aufzeigen zu können. Denkbar ist beispielsweise, daß sich für alle Texte, die bei der Verteilung der Satzlängen, gemessen nach der Zahl ihrer Clauses, einen hohen P-Wert bezüglich der positiven negativen Binomialverteilung erreicht haben, auch bei dem entsprechenden Modell der Satzlängen, gemessen nach der Zahl ihrer Wörter, ein hoher P-Wert ergibt. Denkbar ist auch, daß an alle Texte, die bei der Bestimmung der Satzlängen in Clauses der positiven negativen Binomialverteilung folgen, wieder das analoge Modell bei der Messung in Wörtern angepaßt werden kann, bzw. umgekehrt, daß alle Texte, die sich nicht gemäß der positiven negativen Binomialverteilung verhalten, von dem erwarteten Modell abweichen.

1. Datenerhebung

Um die angesprochene Vergleichbarkeit der Ergebnisse dieser Untersuchung mit der Studie von 1997 zu ermöglichen, bleiben die definitorischen Abgrenzungen dieselben wie in der Untersuchung zuvor. Dies erscheint auch deshalb sinnvoll, da Best (1998c) sich ebenfalls auf diese Definitionen stützt und somit eine gleichartige Datenerhebung bei der Bestimmung der Satzlängen in Wörtern für das Deutsche erfolgen kann.

Für den Vergleich der Auszählungsergebnisse dieser Studie mit der von 1997 ist zu erwarten, daß es bei einigen Texten geringfügige Abweichungen bei der Anzahl der Sätze pro Text geben wird. Solche Schwankungen können sich durch Fehler beim Auszählen sowie durch eine andere Bearbeitung der Zweifelsfälle in der jetzigen Studie ergeben. Toleriert werden Abweichungen von bis zu zwei Prozent.

„Satz" wird definiert als Text zwischen zwei satzabschließenden Zeichen, zu denen der Punkt, das Ausrufezeichen und das Fragezeichen gehören. Der Doppelpunkt nimmt eine Sonderstellung ein, denn er gilt nur dann als satzbeendendes Zeichen, wenn das erste Graphem des folgenden Wortes groß geschrieben wird, bzw. ein vollständiger Satz folgt.

Bei der direkten Redewiedergabe werden die redeeinleitenden Hauptsätze als eigenständige Sätze gewertet, unabhängig davon, ob sie voran- bzw. nachgestellt oder eingeschoben sind. Ähnliche Fälle bei der indirekten Rede dagegen werden als ein Satz gewertet.

Die Wortdefinition erfolgt analog zu den Untersuchungen des „Göttinger Projekts zur Quantitativen Linguistik" und berücksichtigt den graphematischen Aspekt (Best, 1998c:2). Demnach sind Wörter durch Spatien und Interpunktionszeichen voneinander getrennt und lassen sich dadurch eindeutig identifizieren.

Abkürzungen werden entsprechend ihrer lautlichen Realisierung ausgewertet (*z.B.*: „zum Beispiel" = 2 Wörter; *NATO* = 1 Wort).

Bei der Auswertung der Texte findet lediglich der laufende Text Beachtung; Überschriften bleiben unberücksichtigt, da sie zumeist nicht satzwertig realisiert sind. Die Datenerhebung erfolgt so, daß alle Texte nacheinander vollständig ausgezählt werden und keine Stichproben erhoben werden, um Inhomogenitäten zu vermeiden. Im ersten Schritt erfolgt eine Erhebung der Satzlängen nach 1-Wort-Sätzen, 2-Wort-Sätzen, ...; im Anschluß daran werden die Satzlängen jeweils in Gruppen, bestehend aus 5 Längenklassen, zusammengefaßt.

Ausgewertet wurden die folgenden Texte:

Text 1: Ole Schultheis, Zirkus auf dem Bauernhof
Text 2: Margret Rettich, Geschichte ohne Ende
Text 3: Margret Rettich, Trines Spuk
Text 4: Achim Bröger: Moritz lernt Schwimmen
Text 5: Margret Rettich, Michels Kaninchen
Text 6: Ursula Wölfel, Das Miststück
Text 7: Ursula Wölfel, In einem solchen Land
Text 8: Achim Bröger, Moritz tauscht
Text 9: Gert Prokop, Die Maus im Fenster
Text 10: Gudrun Pausewang, Sascha und Elisabeth
Text 11: Monika Pelz, Der Wind in der Krottenbachstraße
Text 12: Christa Reinig, Die Wölfin
Text 13: Urs Widmer, Tod und Sehnsucht
Text 14: Ilse Aichinger, Die geöffnete Order
Text 15: Siegfried Lenz: Ein Haus aus lauter Liebe
Text 16: Gerd Gaiser, Der Schlangenkönig
Text 17: Hans Bender, Mein Onkel aus Amerika
Text 18: Hans Bender, In der Gondel
Text 19: Gisela Elsner, Die Mieterhöhung
Text 20: Gabriele Wohmann, Wiedersehen in Venedig

2. Überprüfung des Modells

Es folgen die Anpassungen der 1-verschobenen negativen Binomialverteilung an die Dateien der 20 ausgewählten Texte mit Hilfe des Altmann-Fitters (1994); die Formel dieser Verteilung lautet:

$$P_x = \binom{k+x-2}{x-1} p^k q^{x-1}, \quad x = 1, 2, \ldots$$

Zunächst werden die Textdateien mit den einzelnen Satzlängen aufgeführt; im Anschluß daran erfolgt eine Zusammenfassung der Satzlängen zu Klassen von je 5 Satzlängen und eine erneute Anpassung des Modells an diese „geglätteten" Dateien.

Als Kriterium der Güte der Anpassung der Texte an das Modell wird der Chiquadrat-Test verwendet. Akzeptiert werden Anpassungen, wenn die Wahrscheinlichkeit P des berechneten Chiquadrats > 0.01 ist, gute Anpassungen lassen sich aber erst für $P > 0.05$ erzielen. Sind die P-Werte sehr klein, so wird zusätzlich der Diskrepanzkoeffizient $C = X^2 / N$ betrachtet. N gibt dabei die Anzahl der Sätze pro Text an. Das Ergebnis wird dann bei $C < 0.02$ akzeptiert, sollte aber möglichst $C < 0.01$ erfüllen.

In den Tabellen werden folgende Informationen gegeben:

x	: Satzlängenklasse;
$n(x)$	: Anzahl der Sätze mit x Wörtern im Text;
$NP(x)$	: theoretische Anzahl der Sätze mit x Wörtern;
k, p	: Parameter der 1-verschobenen negativen Binomialverteilung;
X^2	: Chiquadrat;
FG	: Freiheitsgrade;
P	: Wahrscheinlichkeit des Chiquadrats;
C	: Diskrepanzkoeffizient.

2.1 Anpassungen der 1-verschobenen negativen Binomialverteilung an Textdateien ohne Zusammenfassung der einzelnen Längenklassen

	Text 1		Text 2		Text 3		Text 4	
x	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$
1	0	0.90	4	4.01	1	2.36	5	3.22
2	3	3.58	6	8.16	11	5.68	6	7.82
3	11	7.97	23	11.38	6	8.83	13	11.93
4	13	13.13	11	13.46	9	11.21	11	14.63
5	23	17.85	10	14.49	18	12.66	19	15.73
6	20	21.20	16	14.67	18	13.21	17	15.49
7	19	22.72	16	14.23	9	13.04	18	14.32
8	16	22.49	10	13.38	5	12.35	7	12.63
9	25	20.87	8	12.27	12	11.32	9	10.73
10	20	18.36	11	11.03	4	10.11	7	8.85
11	18	15.46	10	9.77	9	8.84	8	7.12
12	8	12.52	10	8.53	7	7.59	5	5.61
13	13	9.82	10	7.37	9	6.42	7	4.34
14	3	7.48	7	6.31	4	5.36	4	3.31
15	7	5.56	8	5.35	4	4.42	2	2.49
16	4	4.03	2	4.51	6	3.61	2	1.85
17	4	2.87	2	3.78	2	2.93	1	1.36
18	3	2.01	5	3.15	3	2.35	2	1.00
19	1	1.38	2	2.61	5	1.88	1	0.72
20	1	0.93	1	2.15	3	1.49	0	0.52
21	1	1.87	3	1.76	4	1.18	1	1.33
22			1	1.44	0	0.92		
23			2	1.18	1	0.72		
24			2	5.01	0	0.56		
25					0	0.43		
26					1	1.53		
k =	7.9615		2.6803		3.4103		3.8912	
p =	0.5038		0.2421		0.2957		0.3759	
X^2 =	14.100		26.884		32.773		9.663	
FG =	16		21		20		16	
P =	0.5912		0.1747		0.0357		0.8836	

	Text 5		Text 6		Text 7		Text 8	
x	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$
1	2	2.35	0	3.69	6	9.18	1	4.09
2	7	6.24	10	7.65	18	16.67	16	9.39
3	10	10.36	11	10.73	16	20.51	13	13.84
4	14	13.79	19	12.64	29	21.24	19	16.60
5	13	16.08	15	13.49	21	19.93	12	17.64
6	21	17.15	8	13.48	19	17.54	21	17.30
7	14	17.16	16	12.87	11	14.75	13	16.01
8	9	16.37	5	11.88	11	11.99	17	14.19
9	14	15.03	15	10.67	11	9.50	8	12.16
10	15	13.39	9	9.39	5	7.37	6	10.15
11	18	11.63	5	8.13	8	5.62	6	8.28
12	12	9.90	6	6.93	2	4.23	12	6.63
13	9	8.27	5	5.84	3	3.14	6	5.23
14	7	6.81	7	4.87	3	2.31	3	4.07
15	4	5.53	5	4.03	3	1.69	3	3.13
16	4	4.44	4	3.30	0	1.22	3	2.38
17	5	3.52	3	2.69	2	0.88	3	1.79
18	1	2.77	0	2.18	0	0.63	5	1.34
19	4	2.16	3	1.76	1	0.44	0	1.00
20	0	1.68	1	1.41	0	0.31	1	2.78
21	2	1.29	3	1.12	0	0.22		
22	0	0.99	0	0.89	0	0.15		
23	0	0.75	0	0.70	0	0.11		
24	0	0.57	0	0.55	0	0.07		
25	3	0.43	2	0.44	1	0.30		
26	1	0.32	1	1.67				
27	0	0.24						
28	0	0.17						
29	0	0.13						
30	0	0.09						
31	0	0.07						
32	0	0.05						
33	1	0.27						
k =	3.9507		2.8367		2.8081		3.5224	
p =	0.3291		0.2690		0.3538		0.3483	
X^2 =	16.609		23.203		10.663		24.611	
FG =	20		20		15		16	
P =	0.6782		0.2790		0.7761		0.0770	

	Text 9		Text 10		Text 11		Text 12	
x	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$
1	1	6.92	5	13.52	1	0.69	0	1.90
2	6	8.20	26	19.94	2	2.15	7	5.26
3	19	8.55	24	22.25	3	4.07	6	8.97
4	12	8.50	25	22.15	6	6.11	15	12.15
5	8	8.25	19	20.73	7	7.98	16	14.33
6	7	7.88	19	18.65	11	9.46	14	15.38
7	5	7.46	10	16.33	15	10.47	16	15.43
8	7	7.01	14	14.02	15	11.01	10	14.70
9	5	6.55	11	11.86	8	11.10	12	13.46
10	7	6.10	12	9.92	5	10.83	13	11.92
11	6	5.66	8	8.21	14	10.29	5	10.28
12	6	5.23	10	6.74	8	9.55	10	8.67
13	6	4.83	2	5.50	6	8.69	12	7.17
14	4	4.44	5	4.46	7	7.78	4	5.83
15	2	4.08	4	3.60	5	6.86	6	4.68
16	4	3.75	2	2.89	7	5.97	6	3.71
17	1	3.44	2	2.32	5	5.14	4	2.90
18	2	3.15	4	1.85	8	4.38	3	2.22
19	2	2.88	1	1.47	3	3.70	2	1.73
20	1	2.63	0	1.17	3	3.10	2	1.32
21	2	2.40	0	0.92	3	2.57	0	1.00
22	3	2.19	4	0.73	2	2.12	1	0.75
23	3	2.00	1	0.57	3	1.74	0	0.56
24	1	1.82	0	0.45	2	1.42	0	0.41
25	2	1.66	0	0.35	0	1.15	0	0.30
26	3	1.51	0	0.28	1	0.93	0	0.22
27	4	1.37	0	0.22	0	0.75	1	0.16
28	1	1.25	0	0.17	1	0.60	0	0.12
29	2	1.13	0	0.13	1	0.48	0	0.08
30	1	1.03	0	0.10	0	0.38	0	0.06
31	0	0.93	1	0.08	0	0.30	0	0.04
32	0	0.85	0	0.06	0	0.23	0	0.03
33	0	0.77	1	0.05	1	1.00	1	0.23
34	1	0.70	0	0.03				
35	1	0.63	0	0.03				
36	0	0.57	1	0.02				
37	0	0.52	0	0.01				
38	2	0.47	0	0.01				
39	1	0.43	0	0.01				
40	2	0.39	0	0.00				
41	0	0.35	0	0.00				
42	1	0.32	0	0.00				
43	0	0.29	1	0.20				
44	0	0.26						
45	0	0.23						
46	0	0.21						

x	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$
47	0	0.19						
48	0	0.17						
49	0	0.16						
50	0	0.14						
51	0	0.13						
52	0	0.11						
53	0	0.10						
54	0	0.09						
55	0	0.08						
56	1	1.04						
k =	1.3184		1.9507		4.3145		4.2387	
p =	0.1012		0.2440		0.2868		0.3485	
X^2 =	41.294		28.047		16.276		16.465	
FG =	34		20		24		19	
P =	0.1820		0.1083		0.8779		0.6261	

	Text 13		Text 14		Text 15		Text 16	
x	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$
1	3	2.62	0	3.33	2	8.83	0	3.32
2	5	6.14	7	6.37	19	8.00	5	5.44
3	8	9.53	15	8.71	3	7.41	9	6.84
4	8	12.30	11	10.35	5	6.92	12	7.70
5	19	14.26	14	11.37	8	6.48	11	8.19
6	28	15.41	9	11.86	10	6.09	6	8.39
7	11	15.84	7	11.95	11	5.73	8	8.37
8	15	15.69	10	11.74	2	5.39	8	8.21
9	13	15.10	16	11.30	2	5.08	7	7.93
10	9	14.20	10	10.70	3	4.79	7	7.58
11	14	13.10	7	10.01	1	4.52	5	7.18
12	7	11.90	7	9.26	2	4.27	5	6.75
13	12	10.67	8	8.50	3	4.03	6	6.31
14	8	9.46	9	7.74	3	3.80	6	5.87
15	13	8.30	11	7.00	2	3.59	5	5.43
16	7	7.23	6	6.29	0	3.39	7	5.00
17	10	6.24	9	5.63	5	3.21	5	4.59
18	8	5.35	7	5.01	8	3.03	7	4.20
19	3	4.57	1	4.45	3	2.86	4	3.84
20	1	3.87	4	3.93	5	2.71	4	3.50
21	2	3.27	5	3.46	3	2.56	1	3.18
22	3	2.75	3	3.04	2	2.42	3	2.88
23	5	2.30	1	2.67	5	2.29	1	2.61
24	0	1.91	4	2.33	1	2.16	3	2.36
25	1	1.59	0	2.03	0	2.05	4	2.13
26	1	1.32	1	1.77	1	1.94	3	1.92
27	1	1.09	2	1.54	2	1.83	0	1.72
28	1	0.89	1	1.33	2	1.73	3	1.55

x	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$
29	0	0.73	2	1.15	0	1.64	1	1.39
30	2	0.60	0	1.00	4	1.55	1	1.25
31	1	0.49	0	0.86	2	1.47	1	1.12
32	1	0.40	2	0.74	2	1.39	1	1.00
33	0	0.32	0	0.64	3	1.31	0	0.89
34	0	0.26	1	0.55	3	1.24	2	0.80
35	0	0.21	0	0.47	2	1.18	0	0.71
36	0	0.17	0	0.40	0	1.11	0	0.63
37	0	0.14	0	0.34	2	1.05	2	0.56
38	0	0.11	2	2.18	1	1.00	0	0.50
39	1	0.67			2	0.94	0	0.45
40					0	0.89	0	0.40
41					0	0.84	0	0.35
42					1	0.80	0	0.31
43					0	0.76	1	0.28
44					2	0.71	0	0.25
45					1	0.68	0	0.22
46					1	0.64	0	0.19
47					0	0.61	0	0.17
48					2	0.57	1	0.15
49					0	0.54	0	0.13
50					1	0.51	0	0.12
51					0	0.49	1	1.14
52					1	0.46		
53					0	0.44		
54					1	0.41		
55					0	0.39		
56					1	0.37		
57					0	0.35		
58					1	0.33		
59					0	0.31		
60					0	0.30		
61					0	0.28		
62					0	0.26		
63					0	0.25		
64					0	0.24		
65					1	0.22		
66					0	0.21		
67					0	0.20		
68					0	0.19		
69					0	0.18		
70					1	3.58		
k=	3.0592		2.3091		0.9557		1.8844	
p =	0.2347		0.1729		0.0539		0.1297	
X^2 =	35.340		29.915		79.465		22.026	
FG =	27		30		46		35	
P =	0.1305		0.4700		0.0016		0.9569	

	Text 17		Text 18		Text 19		Text 20	
x	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$
1	0	2.83	7	17.78	2	7.37	1	5.03
2	4	5.40	23	17.87	10	8.43	15	7.88
3	12	7.34	13	16.37	18	8.70	8	9.58
4	7	8.63	22	14.51	10	8.65	6	10.52
5	12	9.36	17	12.64	6	8.44	14	10.94
6	10	9.64	15	10.91	13	8.14	12	10.99
7	10	9.58	12	9.34	8	7.79	9	10.79
8	9	9.26	7	7.96	7	7.42	13	10.41
9	9	8.78	6	6.76	8	7.03	10	9.91
10	6	8.18	2	5.73	7	6.64	6	9.34
11	11	7.53	3	4.84	7	6.26	9	8.72
12	4	6.85	4	4.08	3	5.88	8	8.10
13	5	6.18	3	3.43	2	5.52	10	7.47
14	3	5.53	2	2.88	3	5.17	5	6.86
15	7	4.92	0	2.42	4	4.84	6	6.28
16	3	4.34	0	2.03	4	4.52	6	5.72
17	5	3.82	1	1.70	3	4.22	6	5.19
18	6	3.34	3	1.42	3	3.94	3	4.70
19	3	2.91	0	1.19	3	3.67	4	4.25
20	4	2.53	2	1.00	1	3.42	5	3.83
21	1	2.19	1	0.83	1	3.19	8	3.44
22	2	1.89	0	0.69	1	2.96	3	3.09
23	2	1.62	0	0.58	5	2.76	0	2.77
24	0	1.39	2	0.48	2	2.56	4	2.48
25	1	1.19	0	0.40	2	2.38	1	2.21
26	0	1.02	0	0.33	3	2.21	4	1.98
27	1	0.87	0	0.28	6	2.05	1	1.76
28	1	0.74	1	0.23	1	1.90	2	1.57
29	1	0.63	0	0.19	3	1.77	0	1.39
30	0	0.53	1	0.16	0	1.64	1	1.24
31	0	0.45	1	0.13	2	1.52	2	1.10
32	1	0.38	0	0.11	1	1.41	1	0.97
33	0	0.32	0	0.09	1	1.30	0	0.86
34	1	0.27	0	0.07	0	1.21	0	0.76
35	0	0.23	1	0.06	2	1.12	1	0.67
36	0	0.19	0	0.05	2	1.04	0	0.60
37	0	0.16	0	0.04	0	0.96	0	0.53
38	0	0.13	0	0.03	0	0.89	2	0.46
39	0	0.11	0	0.03	1	0.82	0	0.41
40	0	0.09	0	0.02	1	0.76	0	0.36
41	0	0.08	1	0.34	2	0.70	0	0.32
42	0	0.06			1	0.65	0	0.28
43	1	0.51			1	0.60	1	0.25
44					1	0.56	0	0.22
45					1	0.51	0	0.19
46					1	0.47	0	0.17
47					0	0.44	0	0.15

x	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$
48					0	0.41	1	1.26
49					0	0.37		
50					1	0.35		
51					2	0.32		
52					1	0.29		
53					0	0.27		
54					0	0.25		
55					0	0.23		
56					0	0.21		
57					0	0.20		
58					1	0.18		
59					1	0.17		
60					0	0.15		
61					0	0.14		
62					0	0.13		
63					1	1.90		
k =	2.3519		1.2151		1.2435		1.8121	
p =	0.1894		0.1730		0.0802		0.1356	
X^2 =	19.603		34.395		50.802		34.312	
FG =	27		20		42		34	
P =	0.8470		0.0236		0.1655		0.4528	

Da die von Altmann vorgeschlagene 1-verschobene Hyperpascal-Verteilung bei zwei Drittel dieser Textdateien keine akzeptable Anpassung ermöglicht, wird auf eine Darstellung der Ergebnisse verzichtet.

2.2 Anpassungen der 1-verschobenen Binomialverteilung an Textdateien mit Zusammenfassung von je 5 Längenklassen

An alle Texte läßt sich zudem die 1-verschobene Hyperpascal-Verteilung anpassen. Der entsprechende P-Wert ist deshalb in diesen Fällen zusätzlich in Klammern angegeben.

	Text 1		Text 2		Text 3		Text 4	
x	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$
1	50	70.45	54	53.03	45	42.23	54	60.01
2	100	73.84	61	63.88	48	52.23	58	50.18
3	49	42.51	45	39.38	33	33.81	26	23.55
4	13	17.78	12	16.65	19	15.24	6	8.17
5	1	8.42	8	7.15	5	5.37	1	3.09
6					1	2.12		
k =	10.146		42.425		21.275		8.1719	
p =	0.8967		0.9716		0.9419		0.8977	
X^2 =	23.998		2.308		2.067		4.057	
FG =	2		2		3		2	
P =	0.0000	(0.0179)	0.3154		0.5587		0.1315	
C =	0.1127		(0.0342)		(0.1315)		(0.1714)	

	Text 5		Text 6		Text 7		Text 8	
x	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$
1	46	51.05	55	53.73	90	90.55	63	67.52
2	73	67.26	51	52.20	57	56.80	65	59.07
3	50	44.40	28	28.41	19	18.04	30	28.88
4	14	19.58	11	11.41	3	3.87	12	14.53
5	5	6.49	5	3.77	1	0.74		
6	2	1.72	1	1.48				
7	1	0.50						
k =	447.408		8.3003		78.4203		8.4953	
p =	0.9971		0.8829		0.9920		0.8970	
X^2 =	3.926		0.619		0.131		1.375	
FG =	3		3		1		1	
P =	0.2696	(0.2969)	0.8922	(0.5253)	0.7173	(0.3859)	0.2410	
C =								(0.002)

Bei Text 8 muß statt P das Prüfkriterium C für die Güte der Anpassung der Hyperpascal-Verteilung angegeben werden.

	Text 9		Text 10		Text 11		Text 12	
x	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$
1	46	41.11	99	103.74	19	28.33	44	47.34
2	31	32.20	66	57.34	54	45.36	65	59.29
3	24	22.79	29	28.70	40	38.56	37	37.23
4	10	15.56	13	13.86	26	23.14	17	15.62
5	11	10.42	5	6.58	10	10.98	1	4.93
6	11	6.90	0	3.08	3	4.39	1	1.24
7	2	4.54	2	1.43	1	2.24	1	0.35
8	5	2.97	1	0.66				
9	1	1.93	0	0.30				
10	0	1.25	1	0.31				
11	0	0.81						
12	1	1.52						
k =	1.2384		1.2325		16.0653		371.123	
p =	0.3676		0.5516		0.9004		0.9966	
X^2 =	10.613		5.743		6.321		4.160	
FG =	8		5		4		3	
P =	0.2246	(0.1768)	0.3320	(0.1554)	0.1764	(0.2749)	0.2447	(0.0871)

	Text 13		Text 14		Text 15		Text 16	
x	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$
1	43	46.66	47	44.20	38	31.34	37	34.88
2	76	69.34	52	55.46	28	26.95	36	38.36
3	54	54.56	42	42.46	11	21.53	27	29.89
4	29	30.21	27	25.58	21	16.75	27	20.10
5	11	13.21	13	13.33	11	12.86	12	12.44
6	5	4.85	6	6.29	9	9.79	5	7.30
7	2	1.55	3	2.76	12	7.42	3	4.13
8	1	0.62	2	1.92	5	5.60	2	2.27
9					4	4.21	1	1.22
10					4	3.16	1	0.65
11					2	2.37	1	0.76
12					2	1.77		
13					1	1.32		
14					1	3.93		
k =	16.9707		4.5415		1.1671		2.3965	
p =	0.9124		0.7237		0.2629		0.5411	
X^2 =	1.685		0.525		13.437		4.365	
FG =	4		5		11		7	
P =	0.7935	(0.8270)	0.9912	(0.8562)	0.2657	(0.2233)	0.7370	(0.5180)

	Text 17		Text 18		Text 19		Text 20	
x	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$	$n(x)$	$NP(x)$
1	35	35.76	82	86.05	46	41.73	44	42.71
2	44	42.14	42	33.51	31	29.92	50	50.07
3	30	30.80	12	15.43	19	21.99	38	39.14
4	21	17.92	6	7.47	14	16.29	24	25.50
5	6	9.09	3	3.70	11	12.12	16	14.96
6	3	4.20	2	1.86	13	9.04	9	8.19
7	3	1.82	2	0.94	6	6.75	4	4.27
8	0	0.74	0	0.48	4	5.04	2	2.14
9	1	0.53	1	0.56	6	3.77	1	1.05
10					2	2.83	1	0.97
11					3	2.12		
12					2	1.59		
13					1	4.81		
k =	4.151		0.7326		0.9523		2.9957	
p =	0.713		0.4684		0.2471		0.6087	
X^2 =	2.851		4.030		8.354		0.341	
FG =	5		4		10		6	
P =	0.7230	(0.5023)	0.4020	(0.0307)	0.5943	(0.4974)	0.9993	(0.9930)

3. Interpretation der Ergebnisse

a. Satzlängenverteilung ohne Zusammenfassung von Längenklassen

Bei den Anpassungen der beiden Modelle, jeweils in ihrer 1-verschobenen Form, da es keine Null-Wort-Sätze gibt, hat sich gezeigt, daß die Hyperpascal-Verteilung bei der Betrachtung der Textdateien mit den einzeln gelassenen Längenklassen keine akzeptablen Ergebnisse liefert; die Mehrzahl der Texte folgte dem Modell nicht. Zusammen mit den Ergebnissen von Best (1998c) läßt sich vermuten, daß diese Verteilung für die Satzlängen im Deutschen ohne Zusammenfassung von Längenklassen nicht geeignet ist; zusätzliche Untersuchungen sind aber nötig, um diese These weiter zu stützen.

Die von Best vorgeschlagene 1-verschobene Binomialverteilung erweist sich auch hier als das wesentlich bessere Modell. Allerdings ergeben sich nicht für die gesamte Datenbasis gute Anpassungen. Während 17 Texte (ohne Zusammenfassung von Längenklassen) einen P-Wert aufweisen, der größer als 0.05 ist, weshalb die Verteilung für diese Texte akzeptiert werden kann, gibt es auch drei Texte, bei denen die Anpassung nur schwach (Text 3, 18: $0.05<P<0.01$) bzw. gar nicht (Text 15: $P<0.01$) möglich ist. Eine zusätzliche Betrachtung des C-Wertes brachte keine Verbesserung des Ergebnisses, da nur Werte mit $C<0.02$ akzeptiert werden können. Text 15 läßt sich mittels dieser theoretischen Verteilung gar nicht modellieren, da sowohl der P- als auch der C-Wert inakzeptabel sind.

Gründe für die schlechten Anpassungen lassen sich nur schwer angeben. Lediglich bei Text 18 ist eine Erklärung für die Abweichung vom Modell denkbar.

Die Geschichte „In der Gondel" ist stark dialogisch aufgebaut, und es ist zu vermuten, daß solche Texte einer anderen Verteilung folgen als überwiegend erzählende Texte. Es wäre z.B. daran zu denken, daß für die dialogischen Textpassagen womöglich eine andere Verteilung als für den Rest des Textes nötig wäre, da es sich bei derartigen Texten um eine Mischung von (simulierter) gesprochener mit geschriebener Sprache handelt.

b. Satzlängenverteilungen mit Zusammenfassung der Längenklassen

Für die Textdateien mit den zusammengefaßten Satzlängenklassen zeigt die tabellarische Auswertung, daß die 1-verschobene negative Binomialverteilung mit einer Ausnahme (für Text 1 ist keine Anpassung möglich) ein gutes Modell darstellt. Zusammen mit den Untersuchungen von Best (1998c) braucht die negative Binomialverteilung deshalb bis auf weiteres nicht als Modell verworfen zu werden.

Im Gegensatz zur Untersuchung Bests erweist sich für diese „geglätteten" Dateien aber auch die Hyperpascal-Verteilung als einigermaßen geeignet, da für alle 20 Texte eine Anpassung möglich ist, wenn diese auch in drei Fällen (Texte 1, 2 und 18) nur schwache *P*-Werte aufweist.

Insgesamt ergibt sich durch die Anpassung der Hyperpascal-Verteilung aber keine Verbesserung, da bei 15 Texten die negative Binomialverteilung einen teilweise allerdings nur geringfügig besseren *P*-Wert liefert.

Vergleicht man die *P*-Werte, die sich bei der 1-verschobenen negativen Binomialverteilung mit und ohne Zusammenfassungen der Längenklassen ergeben, so stellt man fest, daß sich bei immerhin acht Texten eine bessere Modellierung ohne die Zusammenfassungen ergibt. Dieses Ergebnis überrascht insofern, als zu vermuten ist, daß durch die Gruppierungen zu Klassen Schwankungen bei den einzelnen Längen besser ausgeglichen werden können, woraus eine verbesserte Anpassung der theoretischen Verteilung resultieren müßte; das ist aber offensichtlich nicht immer der Fall.

Insgesamt kann man feststellen, daß die Hyperpascal-Verteilung nur eingeschränkt geeignet erscheint, um die Satzlänge nach der Anzahl der Wörter im Deutschen zu modellieren, dann nämlich, wenn man die Satzlängenklassen zu Gruppen zusammenfaßt. Die 1-verschobene negative Binomialverteilung erweist sich dagegen als ein recht geeignetes Modell, und zwar sowohl für Dateien ohne als auch für solche mit Zusammenfassung von Längenklassen. Die Ergebnisse untermauern damit die Studie Bests (1998c), die zu einem ähnlichen Schluß kommt.

Überraschend ist dieses Ergebnis insofern, als zwei Formen der negativen Binomialverteilung, die positive und die 1-verschobene, sich damit bislang als relativ geeignet erweisen, sowohl die Satzlängen, gemessen nach der Zahl der Clauses als auch gemessen nach der Zahl der Wörter pro Satz, angemessen zu modellieren. Wenn sich dieses Ergebnis auch bei weiteren Untersuchungen einstellen sollte, muß man daraus schließen, daß die intervenierende Ebene bei der Satz-Wort-Variante im Deutschen offenbar keine Störungen hervorruft, wie dies noch von Altmann generell vermutet wurde (Altmann 1988b:63).

Zu prüfen wäre in diesem Zusammenhang, ob die negative Binomialverteilung auch noch für die Modellierung der Satzlänge geeignet ist, wenn diese durch die Anzahl der Silben oder Morph(em)e bestimmt wird.

4. Vergleich der Ergebnisse mit der Satzlängenstudie Niehaus (1997)

Im zweiten Teil der Auswertung soll überprüft werden, ob ein Zusammenhang zwischen der Verteilung der Satzlänge in Clauses (Satz/Clause-Verteilung) und der Verteilung der Satzlänge in Wörtern (Satz/Wort-Verteilung) erkennbar ist, wenn die negative Binomialverteilung in ihren beiden Formen an die ermittelten Daten angepaßt wird.

Die folgende Tabelle zeigt die Wahrscheinlichkeit des Chiquadrats bezüglich der einzelnen Untersuchungen an. Die erste Spalte gibt dabei den *P*-Wert der Satz/Wort - Verteilung („Wort") an, die zweite Spalte den P-Wert der zu Intervallen zusammengefaßten Klassen („Wort 2") und die dritte Spalte („Clause") die *P*-Werte der entsprechenden Texte, deren Satzlänge nach der Zahl der Teilsätze bestimmt wurde (Niehaus 1997:238ff).

	P-Wert Wort	*P*-Wert Wort 2	*P*-Wert Clause
Text 1	0.5912	0.0000	C=0.0198
Text 2	0.1747	0.3154	0.12
Text 3	0.0357	0.5587	C=0.0020
Text 4	0.8836	0.1315	C=0.0305
Text 5	0.6782	0.2696	0.06
Text 6	0.2790	0.8922	C=0.07
Text 7	0.7761	0.7173	0.06
Text 8	0.0770	0.2410	C=0.0184
Text 9	0.1820	0.2246	0.85
Text 10	0.1083	0.3320	0.77
Text 11	0.8779	0.1764	0.77
Text 12	0.6261	0.2447	0.28
Text 13	0.1305	0.7935	0.63
Text 14	0.4700	0.9912	0.49
Text 15	0.0016	0.2257	0.14
Text 16	0.9569	0.7370	0.30
Text 17	0.8470	0.7230	0.37
Text 18	0.0236	0.4020	0.37
Text 19	0.1655	0.5943	0.63
Text 20	0.4528	0.9993	0.21

Die Tabelle zeigt, daß die oben angeführten Hypothesen für diese Texte nicht bestätigt werden können. So gibt es Texte, die bezüglich der Satz/Wort-Verteilung dem Modell der negativen Binomialverteilung folgen, bei denen aber bezüglich der Satz/Clause-Variante keine Anpassung möglich ist (Vgl. Texte 4 und 6). Umgekehrt ergibt sich dies auch bei Text 1. Hier kann keine Anpassung der negativen Binomialverteilung an die zusammengefaßten Daten erfolgen, dennoch ist aber eine Modellierung der Satz/Clause-Verteilung möglich.

In diesem Zusammenhang zeigt sich auch, daß die *P*-Werte der einzelnen Untersuchungen nicht immer gleich gut sind. So kann für diese Texte nicht davon ausgegangen werden, daß hohe *P*-Werte bei der Satz/Wort-Verteilung auf hohe *P*-Werte bei der Satz/Clause-Variante oder umgekehrt hindeuten (Texte 5, 7, 9).

Aus diesem Ergebnis läßt sich aber nicht folgern, daß alle Textdateien der Satz/Wort-Variante, an die das Modell der negativen Binomialverteilung gut angepaßt werden konnten, immer einen geringen *P*-Wert bei der Satz/Clause-Verteilung liefern. So ergeben sich für Text 11 bei dem Vergleich der für alle Längen einzeln bestimmten Satz/Wort-Verteilung mit der Satz/Clause-Verteilung zwei recht hohe *P*-Werte; bei zusammengefaßten Längenklassen liefert Text 18 einen etwa gleich guten *P*-Wert wie bei der Satzlängenbestimmung nach Clauses.

Zusammenfassend läßt sich für diese Untersuchung feststellen, daß zunächst keine klaren Zusammenhänge zwischen den verschiedenen Arten der Satzlängenmessung und der Güte der Anpassung der negativen Binomialverteilung zu bestehen scheinen. Ob dies tatsächlich so ist, müßte anhand weiterer Untersuchungen geprüft werden, da die Datenmenge doch noch zu gering ist, um abschließende Aussagen machen zu können.

Bearbeitete Texte

Aichinger, I. (1981). Die geöffnete Order. In Aichinger, I., *Meine Sprache und ich. Erzählungen* (S. 20-26), Frankfurt.

Bender, H. (1962). In der Gondel. In Bender, H., *Mit dem Postschiff. 24 Geschichten* (S. 10-14), München.

Bender, H. (1984). Mein Onkel aus Amerika. In Bender, H., *Der Hund von Torcello. 32 Geschichten* (S. 164-168), Frankfurt.

Bröger, A. (1983). Moritz tauscht. In Bröger, A., *Moritzgeschichten* (S. 72-78), Ravensburg.

Elsner, G. (1983). Die Mieterhöhung. In *Erzählungen seit 1960 aus der Bundesrepublik Deutschland, aus Österreich und aus der Schweiz* (Hg. H. Vormweg) (S. 206-215), Stuttgart: Reclam.

Gaiser, G. (1983). Der Schlangenkönig. In Gaiser, G., *Mittagsgesicht. Erzählungen* (S. 125-132), Ostfildern.

Lenz, S. (1986). Ein Haus aus lauter Liebe. In Lenz, S., *Die Erzählungen (1949 - 1984).* Bd. 1 (1949 - 1958) (S. 65-70), München.

Pausewang, G. (1987). Sascha und Elisabeth. In *Die schönsten Freundschaftsgeschichten* (Hg. H. Westhoff) (S. 11-18), Ravensburg.

Pelz, M. (1988). Der Wind in der Krottenbachstraße. In *Meine Welt. Geschichten zum Lesen und Vorlesen* (Hg. I. Ryssel) (S. 11-18), Gütersloh.

Prokop, G. (1982). Die Maus im Fenster. In Prokop, G., *Die Maus im Fenster. Gute-Nacht-Geschichten* (S. 7-18), Zürich, Köln.

Reinig, C. (1986). Die Wölfin. In Reinig, C., *Gesammelte Erzählungen* (S. 288-293). Darmstadt und Neuwied (Sammlung Luchterhand).

Rettich, M. (1986). Geschichte ohne Ende. In Rettich, M., *Allerlei von früher, jetzt und irgendwo* (S. 97-105), Hamburg.

Rettich, M. (1988). Michels Kaninchen. In Rettich, M., *Seidenhund und Lumpenköter. Geschichten von besonderen Tieren* (S. 22-30), Wien, München.

Rettich, M. (1988). Trines Spuk. In Rettich, M., *Seidenhund und Lumpenköter. Geschichten von besonderen Tieren* (S. 126-132), Wien, München.

Schultheis, O. (1988). Zirkus auf dem Bauernhof. In *Ich hör' so gern Geschichten: kleine Geschichten zum Vorlesen* (Hg. U. Schultheis) (S. 37-46), München.

Widmer, U. (1983). Tod und Sehnsucht. In *Erzählungen seit 1960 aus der Bundesrepublik Deutschland, aus Österreich und der Schweiz* (Hg. H. Vormweg) (S. 194-201), Stuttgart: Reclam.

Wölfel, U. (1970). Das Miststück. In Wölfel, U., *Die grauen und die grünen Felder. Wahre Geschichten* (S. 79-86), Mühlheim/Ruhr.

Wölfel, U. (1970). In einem solchen Land. In Wölfel, U., *Die grauen und die grünen Felder. Wahre Geschichten* (S. 51-57), Mühlheim/Ruhr.

Wohmann, G. (1979). Wiedersehen in Venedig. In Wohmann, G., *Ausgewählte Erzählungen aus zwanzig Jahren* (S. 25-33), Bd. 1. Darmstadt und Neuwied (Sammlung Luchterhand).

Literatur

Altmann, G. (1988a). Verteilungen der Satzlängen. In K.-P. Schulz (Hrsg.), *Glottometrika 13* (S. 147-169), Bochum: Brockmeyer.

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

Best, K.-H. (1998a). *Quantitative Linguistik: Entwicklung, Stand und Perspektive.* Manuskript.

Best, K.-H. (1998b): Results and perspectives of the Göttingen project on quantitative linguistics. *Journal of Quantitative Linguistics*, 5, 1-8.

Best, K.-H. (1998c). Wieviele Wörter enthalten Sätze im Deutschen. In K.-H. Best (Hrsg.), *Verteilungen in Texten* (in Arbeit).

Köhler, R., & Altmann, G. *Einführung in die quantitative Linguistik* (Arbeitstitel). Manuskript, Kap. 2, 1-18).

Niehaus, B. (1997). Untersuchung zur Satzlängenhäufigkeit im Deutschen. In K.-H. Best (Hrsg.), *Glottometrika 16* (S. 213-275), Trier: WVT.

Strehlow, M. (1997). *Satzlängen in pädagogischen Fachartikeln des 19. Jahrhunderts.* Staatsexamensarbeit, Göttingen.

Wittek, M. (1995) *Zur Entwicklung der Satzkomplexität im gegenwärtigen Deutschen.* Staatsexamensarbeit, Göttingen.

Software

Altmann-FITTER (1994). Lüdenscheid: RAM - Verlag.

Comparing Word Length in Different Languages

Antero Niemikorpi

Finnish is usually regarded as a language characterised by long words. One gets this impression when examining handbooks and textbooks which describe the Finnish language. Is this impression in fact quite correct? Is Finnish, gauged in different ways, characterized by long words in every sense in relation to the languages with which it is most commonly compared? I will begin by considering some earlier studies and will then attempt to shed some light on the question by presenting a few new ways of calculating word length.

Finnish words are polysyllabic

What are the grounds for the conception, embodied in so many general descriptions, that long words are typical of the Finnish language? Lauri Hakulinen in his frequently revised work on the structure and development of Finnish (*Suomen kielen rakenne ja kehitys*) characterised Finnish words for instance in the following manner: 1) there are only some 50 monosyllabic word stems in Finnish, which is very little compared with the Indo-European languages (e.g. German has some 2000 monosyllables), 2) Finnish words are on the average comparatively polysyllabic, and 3) the syllables of Finnish words are "lighter" than those of other languages. Hakulinen refers in support of his statement among other things to a calculation based on the New Testament, where, in the Gospel according to Saint Matthew, which constituted the material for comparison, the number of syllables in the Finnish version is about 40500, in the Swedish 35000, in the German 33000 and in the English 29000 (1979:30-32).

Calculations based on the number of syllables as a measure of the terseness of expression of a language have not been considered wholly reliable, for they lead to a quite different result than a comparison of texts in their entirety (cf. Alhoniemi, 1968:92, 95).

An observation worth noting in Hakulinen's characterisation above is the reference to the lightness of the syllables of Finnish words. The syllables seem to

follow the principle referred to as Menzerath's law: the longer the linguistic units, the shorter their subunits (cf. for instance Altmann, 1980:5-10). In this respect Finnish is not exceptional: calculations concerning the relation between the number of letters and syllables in German words also indicate a similar ratio (Meier, 1967:290). According to calculations carried out by Kalevi Wiik (1977:275) on the basis of Finnish fictional texts, the most frequent syllable type of the more than 3000 segmentally distinct types of syllables actually occurring in Finnish is CV (C = Consonant, V = Vowel). Fred Karlsson for his part notes that one-mora[1] V and CV syllables are the lexically most frequent types. Of the total of 8 possible V syllables and 104 possible CV syllables, all are used. In the set of two-mora syllables occasional gaps occur involving 10% of the possible syllables; the proportion of gaps with regard to three-mora syllables rises to 65-90%. The CV syllable as a type is, according to Karlsson, the optimal syllable in Finnish (1983:135-139).

In Finnish the two-syllable word is the most favoured, and one-syllable words are relatively rare. There are only 13 VV, CV and VC word types in all, and the one-syllable basic forms total no more than 85. This is a fairly small number in proportion to the theoretically possible number and in comparison with the Indo-European languages, as it has been estimated that for instance in English such forms amount to about 7000 (cf. also Karlsson, 1975:18). On the other hand, 25% of the 100 most frequent words in Finnish belong to this category; in other words, the relative proportion of monosyllables in a text is much larger than in the lexicon representing the language system. According to some calculations the frequency of occurrence of one-syllable words in English (a Byron text) is 70-80% of all the words, in German about 50% but in Finnish only about 15%. Although the data are hardly quite comparable statistically and with regard to text genre, the differences between the languages seem nevertheless to be indisputable.

The view that syllables become lighter when their number in a word increases is also supported by my calculation, which was based on the 200 most frequent words in the Finnish frequency word-list. The number of these words is small in proportion to the whole lexicon, but their importance when estimating the weight of the syllables of the running words in a text (the tokens) is quite considerable if one takes into account that these words cover 43% of all the words in a text.

According to the computation, the proportion of syllables of different lengths in the words in question was as follows:

[1] Mora is a quantitative unit, a short sound bigger than a phone; a syllable may consist of one or more morae.

Length of syllables in graphemes	*Proportion %*
1	9.3
2	48.6
3	35.7
4	6.4

The average length of words with a different number of syllables was as follows:

Syllables per word	*Average length of syllables in graphemes*
1	3.1
2	2.4
3	2.2
4	2.7

There are only four four-syllable words in the category, so the deviation from the regularity of the syllable-length distribution is probably accidental. Otherwise the result is in agreement with Menzerath's law (Reinhard Köhler among others also speaks of the Menzerath-Altmann law, see 1997:122). Below I will focus in more detail on length calculations based on the number of graphemes of the words.

Different estimates of the lengths of textual words in Finnish

Calculating the average length of word-types and word-tokens in a running text is unproblematic even when it is a question of a single language. The calculations carried out with reference to Finnish are illustrative in this respect as well. Vilho Setälä based his calculation on the length of syllables in a Finnish translation of the New Testament and reached the conclusion that the average length of the running words in the Finnish text was 6.2 letters and stated that this figure represented the average length of the "Finnish lexeme" (= word-token) (1967:371). In my own calculations based on the Oulu corpus I arrived at a more than one letter higher figure (cf. Niemikorpi, 1991:86). The difference in our calculations was that Setälä's corpus – just like the samples of several other earlier calculations – represented only one text genre with rather short words, whereas my own calculations were based on a statistical sample representing a

versatile selection of different text genres of Finnish written language. Perhaps the best way of establishing the length of words in some language would be to indicate their maximum and minimum lengths, in which case it would be possible to state that the average length of word-tokens in Finnish written language varies between ca 5.8 and 8.5 letters depending on text genre. Another factor complicating an unambiguous interpretation covering all the text genres of a language is the fact that the length of word-types or dictionary words and naturally also the word-tokens or running words in texts belonging to different parts of speech as well as those of varying frequency varies considerably depending on text genre (cf. Niemikorpi, 1991:153-171).

The average length and distribution of word-tokens in different languages

Next I will examine some calculations concerning the length of word-tokens in different languages. The average length of Finnish word-tokens calculated on the basis of the Oulu corpus is 7.4 graphemes or, as was mentioned above, about one letter more than indicated by certain earlier presented calculations chiefly based on fictional or religious texts. The dependence of word length on text genre has been shown to hold good in other languages as well (as far as German is concerned, see for instance Fucks & Leuter, 1971:113). Word length has also been found to be connected with the abundance of inflections of a language. Of the languages included in this comparison, English has the shortest words and is morphologically the simplest, while Finnish has the longest words and is the most complex morphologically. According to Karlsson, it is possible to form at least 2112 derivatives from the nouns and as many as 12000 from the verbs (1983:357).

The average length of English word-tokens, as calculated on the basis of the so-called Brown corpus, is 4.74 graphemes (Kučera & Francis, 1970:366). The corresponding figure for German is 5.04 graphemes, calculated on the basis of the Meier corpus (1967:2909), and for Swedish 5.4 graphemes calculated on the basis of the big Swedish frequency dictionary (Allén, 1970:89). Although the above-mentioned calculations are based on corpuses of different age and type, they nevertheless probably represent each written language fairly adequately for a comparison, and the order of the languages in terms of word length seems to be clear: Finnish, Swedish, German, English.

The differences in the length of words in the languages is also illustrated by the distribution of word-tokens into different length classes. If the words are divided depending on their length into three categories of practically the same size in Finnish – short, medium-long and long – the distribution looks like this:

	short words 1-5 graphemes	*medium-long words 6-8 graphemes*	*long words 9 graphemes*
Finnish	35.1%	30.9%	34.0%
Swedish	62.0%	20.9%	17.1%
German	67.0%	19.6%	13.4%
English	67.8%	21.9%	10.3%

This comparison shows that Finnish has the longest average word-tokens. Finnish uses roughly speaking only a little more than half as many short words in texts as English or German, and at the same time the proportion of long words in Finnish texts is twice as big or even three times as big as that of the other languages.

Word length in the light of different dictionaries

Comparing the length of words between languages by means of dictionaries is not unproblematic either. The difficulty arises in particular from the interpretation of phrases and compound words, for the orthography and word configuration principles differ in this respect considerably among languages (I will later return to their different structural differences). Below, in my calculations based on the dictionaries of different languages, I have been content to note only solid words and word equivalents in the illustrations or samples. An exception is made for a few words in Branch et al.'s student's glossary, where it was necessary to include phrases in the calculation in order to get the sum total of the counted words totally.

The first sample to be examined was a multilingual glossary compiled for teaching purposes on a frequency basis. It was compiled for the purpose of teaching foreigners Finnish. The glossary is based on the fiction and newspaper text samples of the so-called Oulu corpus and thus represents well Finnish standard language (for more details, see Branch et al., 1980:16-18). From the top of the rank list of the glossary I computed the average length of the word types of the five languages concerned in such a manner that I first estimated the word length cumulatively increasing the number of words in accordance with this order of frequency, at intervals of 50 words, till I reached number 500; and in addition in the same way from the running number 1 800 to the end of the glossary. The parameters arrived at in this way are enough to show the change in word length occurring in the early part of the glossary and the relation between the languages. In addition to the languages involved in the earlier comparison, French is now included as well (Fr).

Word order	*Fi*	*Sw*	*G*	*Fr*	*Eng*
1-50	4.2	3.8	4.4	4.7	3.5
1-100	4.6	4.1	4.8	5.3	4.0
1-150	4.9	4.2	4.9	5.4	4.3
1-200	5.0	4.5	5.2	5.7	4.6
1-250	5.2	4.6	5.3	5.9	4.8
1-300	5.4	4.8	5.5	6.0	5.0
1-350	5.4	4.9	5.6	6.0	5.0
1-400	5.5	5.1	5.8	6.1	5.1
1-450	5.6	5.2	5.9	6.2	5.2
1-500	5.6	5.3	6.0	6.2	5.3
---	---	---	---	---	---
1800-1850	7.0	6.4	7.2	6.7	6.3
1800-1900	7.4	6.8	7.7	7.6	6.9
1800-1950	7.2	6.7	7.6	7.5	6.8
1800-2000	7.3	7.0	7.7	7.5	6.7

As far as Finnish is concerned, the above calculations can first of all be compared with those I have presented above, carried out on slightly different principles, on the basis of the Finnish frequency word-list and thus on the whole corpus. (Niemikorpi, 1991:153-155; in the table below the words from the running number 2943 upwards with the same frequency number have been combined into one group, and the last is the group of hapax legomena with the running number 25422.)

Order in the frequency list	*Average word length in graphemes*
1-50	4.4
51-100	5.5
101-150	5.5
151-200	5.9
201-250	6.4
251-550	6.7
551-1 050	6.2
1051-1550	6.8
1551-2050	7.4
2051-2550	8.3
2551-2943	8.2
3740	8.3
4315	7.9

Order in the frequency list	*Average word length in graphemes*
4701	8.8
5204	8.2
5842	9.7
6684	9.3
7844	9.1
9466	9.2
11536	10.5
25422	10.9

The difference between the two tables above with respect to Finnish are due to the method of calculating: the former is calculated cumulatively, the latter again as sections representing different frequency groups. The most conspicuous difference is that the latter method more clearly displays the dependence of word length on frequency whereas the former shows that the average length of the most frequent two thousand words does not yet reach the parameter of the whole corpus.

In the frequency word-list of Finnish, the words occurring in the group with the running number 11532 (with items occurring twice) represent with regard to their length more or less the average of the word-types in the whole corpus. It is natural that average length is to be found in the region of the most infrequent words when one takes into consideration that the proportion of words that occur once and twice is 73.1% of all words (Saukkonen et al., 1979:19). The proportion of these words of the total of word-tokens is, however, only 9.4%, and their quantitative weight in texts in relation to their lexical proportion is small. In any selection of dictionary words, however, their proportion is considerable. Of the words in the biggest Finnish dictionary of written language (*Nykysuomen sanakirja*), the number of compound words amounts to 64.8% while 26.6% are derivatives; in other words, the majority of these are precisely the most infrequent words, and only 8.6% are basic, underived words in modern Finnish. In the 43670 words of the list that forms the basis of the frequency list of Finnish words, these basic words constitute an even smaller proportion. (Niemikorpi, 1991:154.)

All in all it can be noted that the frequency of words in all languages consistently influences their length. Another essential observation based on the just presented parameters and serving the purpose of a final assessment of the relation between languages indicates that Finnish is not in this comparison the language with the longest words, but German and French surpass Finnish in this respect; nor are Swedish words much shorter than Finnish. The biggest difference between Finnish and the other languages with which Finnish has been compared on

the basis of the Finnish frequency list and other frequency lists arises from the fact that generally in other languages the top ranks of the frequency list are taken up by articles and prepositions, which are wholly missing from this examination (because the structure of the Finnish frequency word-list determined the selection of the material for comparison). Whether the syntactic structure of the sentences is analytical or synthetic is something that this comparison does not reveal even indirectly. This means in practice that in the present calculation practically only individual lexical elements carrying substantial meaning are included. The grammatical elements which in the description of Finnish are listed and described in morphology and which in other languages are included in the lexicon are thus left out of account here, and the comparison here concerns what could be called "purified" vocabularies. Apart from this, the top ranks of the frequency word lists of different languages have proved to be qualitatively similar (cf. also Niemikorpi, 1991:94, 335), whereas the result arrived at here reflects the real length relations between semantically equivalent words in different languages.

A comparison between bilingual dictionaries on the basis of samples

The other set of material on which the lengths of lexical words was compared consisted of bilingual dictionaries in which one language was Finnish and the other language one of the languages the study concerns. The dictionaries contained 25000-35000 words and from each a sample of some 30 pages was selected by picking words at regular intervals, the total bulk of each sample amounting to 500-600 words. The comparison thus in each case involved a Finnish word and its counterpart in the other language, and phrasal items were not included in this case either. The average lengths of the words calculated in graphemes were the following:

Fi	*Sw*	*G*	*Fr*	*Eng*
10.0	10.1	10.8	8.1	8.3

As regards Finnish, the figure is somewhat smaller than the average length, 10.5 graphemes, of word-types based on the frequency word-list. The result may also vary with respect to the other languages for reasons concerned with the different ways in which the dictionaries were compiled, but the figures are nevertheless strongly indicative and for their part also indicate that German words are longer than Finnish words and Swedish words more or less equal to Finnish words in length, while French and English words here prove to be shorter than the rest. With regard to French and English the result seems natural, for in the

dictionaries the English and French equivalents of Finnish word entries were often offered in the form of phrases (e.g. Fi *etsintäkuulutettu* – Eng *wanted by the police*; Fi *osastopäällikkö* – Fr *chef de département*).

Conclusion

The length of a word is difficult, not to say impossible, to fully define by means of one parameter in comparisons between languages, and even in the case of a single language the definition of word length by means of a parameter will lead to an inadequate result with regard to both the system and the variety of a language. Word length varies with the different variants of a language: in fiction both types and tokens are the shortest, in scientific literature the longest. This is basically a result of syntactic differences between these textual genres: in fiction the sentences are short and contain few long nominal forms, whereas in scientific literature the sentences are long, containing a great number of modifications and long noun phrases.

When comparing languages, several different structural as well as lexical aspects should be considered.

From the standpoint of lexicon and usage there are two ways of studying and comparing word length in different languages:

1) First of all the length of word-tokens should be separately examined. It reflects more widely the structure and syntax of a language and a text (see e.g. Niemikorpi, 1997:193-196; cf. Uhlířová, 1997:266-275). In addition, a comparison of word-tokens between languages also reveals differences in the actual vocabulary and the morphosyntactic structure of the languages (e.g. Fi *talo+ssa+ni* – Eng *in my house*. Finnish has in a wider perspective been found to have fewer full clauses and more clause equivalents and embeddings than the languages with which Finnish has been compared in this context; cf. Ingo, 1990:225-230).

2) The length of lexemes in dictionaries should be separately examined. Here one decisive factor is the way in which different languages, for various reasons, favour certain basic words, derivations and compounds.

Word length has also been found to be connected with the phonological structure of a language: the tendency is that the more phonemes there are in a language and the fewer restrictions on combining the phonemes, the shorter are the words of the language (Grotjahn, 1982:74). According to certain sources Finnish standard language is considered to have 8 vowels and 13 consonants, and especially the number of consonants is small compared for instance with 19 in Swedish and 24 in English. Besides, phonotactic factors tend to restrict the number of word types in Finnish: vowel harmony, debarring word initial consonant

clusters, limitations on word final vowels and consonants etc. In accordance with these principles, Finnish is therefore structurally destined to be a language with long words. Still, on the basis of this study it can be stated that the number of long words included in the dictionaries of all the languages examined is, perhaps surprisingly, very similar and that even the lexical words in these different languages are closer to each other as regards word length than has usually been assumed.

Word length thus cannot be defined by means of one parameter, nor can comparisons be directly applied for instance to describe the terseness of expression of languages: about the same amount of space is needed by a text written in different languages (according to some recent comparisons between translated texts, Finnish even requires less space than the languages with which Finnish has been compared above). Calculations of sentence length based on the frequency list corpuses between the languages compared here indicate that Finnish and English in terms of the sentence length of written language computed in graphemes are fairly close to each other (Fi 95 graphemes and Eng 92 graphemes / sentence) and that an average German sentence contains twice as many graphemes as Finnish and English (G 115 graphemes / sentence). If the word-tokens in Finnish are longer than those of the languages with which Finnish has been compared, in terms of the number of words Finnish sentences are correspondingly on the average shorter. Finnish is thus not as uneconomical as one might be led to believe on the basis of general descriptions of the language.

References

Alhoniemi, A. (1986). Suomen kirjakielen luonteenomaiset piirteet. In O. Ikola (Ed.), *Suomen kielen käsikirja* (pp. 92-95) [Characteristic features of written Finnish]. Tapiola: Weilin & Göös.

Allén, S. (1970-75). *Nusvensk frekvensordbok baserad på tidningstext - 4.* [A modern Swedish frequency dictionary based on newspaper texts]. Stockholm: A & W.

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

Branch, M., Niemikorpi, A., & Saukkonen, P. (1980). *A Student's Glossary of Finnish.* Porvoo: WSOY.

Fucks, W., & Leuter, J. (1970). Mathematische Analyse des literarischen Stils. In R. Gunzenheuser & H. Kreuzer (Eds.), *Mathematik und Dichtung* (pp. 107-122). München: Nymphenburger Verlagsbuchhandlung.

Grotjahn, R. (1982). Ein statistisches Modell für die Verteilung der Wortlänge. *Zeitschrift für Sprachwissenschaft,* 1, 44-75.

Hakulinen, L. (1979). *Suomen kielen rakenne ja kehitys.* [The structure and development of Finnish]. 4. korjattu ja lisätty painos. Helsinki: Otava.

Ingo, R. (1990). *Lähtökielestä kohdekieleen.* [From source language to target language] Porvoo: WSOY.

Kučera, H., & Francis, W. N. (1970). *Computational Analysis of Present-day American English.* Rhode Island: Brown University Press.

Köhler, R. (1997). Are there Fractal Structures in Language? Units of Measurement and Dimensions in Linguistics. *Journal of Quantitative Linguistics,* 4, 122-125.

Meier, H. (1967). *Deutsche Sprachstatistik.* Hildesheim: Olms.

Niemikorpi, A. (1991). *Suomen kielen sanaston dynamiikkaa.* (Summary: Dynamics of Finnish.) Acta Wasaensia.

Niemikorpi, A. (1997). Equilibrium of Words in the Finnish Frequency Dictionary. *Journal of Quantitative Linguistics,* 4, 26-275.

Saukkonen, P., Haipus, M., Niemikorpi, A., & Sulkala, H. (1979). *Suomen kielen taajuussanasto.* (Summary: A Frequency Dictionary of Finnish.) Porvoo: WSOY.

Setälä, V. (1967). Tilastollisia tietoja Uuden testamentin suomennoksen sanastosta. (Resumo: Vortara statistiko pri la Nova Testamento en finna traduko.). *Virittäjä,* 4, 368-372.

Setälä, V. (1972). Suomen kielen dynamiikkaa. (Resumo: Dinamiko de finna linvo.). *Suomi,* 116, 3. Helsinki: SKS.

Uhlířová, L. (1997). Length vs Order: Word Length and Clause Length from the Perspective of Word Order. *Journal of Quantitative Linguistics,* 4, 266-275.

Wiik, K. (1977). Suomen tavuista. (Summary: On Finnish syllables.) *Virittäjä,* 265-278.

Sequential Modelling of Text Structure and its Application in Linguistic Typology

Adam Pawłowski

Introduction

Quantitative measures in linguistic typology are inseparably associated with the person of Joseph Harold Greenberg, the American linguist and anthropologist. In 1954, he proposed a list of numerical indices (so called *Greenberg indices*), which serve the purpose of quantitative classification of languages.[1] Since the beginning, Greenberg's method, as well as the very question of linguistic taxonomy, has been regarded as a controversial issue.[2] In particular, one can hardly determine which set of indices can be considered as complete and satisfactory. It can also be shown that different methods of multidimensional scaling may produce different clusterings of the same set of data.[3] However, despite these critiques, new studies on the quantitative typology of languages continue to be published (Batagelj, Pisanski, & Keržić, 1992).

The present paper *is not* meant to be one more voice in the discussion over the (non-)scientific character of taxonomies or the ontological status of language universals. Instead, we intend to carry out a series of empirical tests in order to gain new insights into the issue of sequential modelling of text and its possible applications (e.g. in typology). In particular, we will try to answer the following questions:

- is there sequential structure in text segmented in words?
- can this structure be modelled by means of mathematical tools?
- is this structure dependent on the morphosyntactical characteristics of language?

[1] The description of these indices can be found in the article Greenberg, 1960. After its publication, the original Greenberg list was on frequent occasions modified (e.g. Silnitsky, 1993).

[2] "Some of the most obvious and frequently mentioned syntactical differences do not easily lend themselves to this technique (...) all these [language characteristics – AP], and many more like them, are difficult to reduce to a meaningful number." (Householder, 1960:195).

[3] Taxonomy as an object of QL is criticised by Gabriel Altmann: "At early stages of explorative research, one usually *classifies* texts, languages or particular phenomena in order to obtain a map of the scope of taxa. (...) One can gain useful impulses but one observes that empirical taxonomies lead quickly to a dead end." (Altmann, 1997:15). Cf. also Altmann & Lehfeld, 1973.

Language in the mass vs language in the line

Greenberg's indices, originating in previous typological concepts, especially those of Edward Sapir (Greenberg,1960:180-184), are constructed as simple numerical relations (the number of units having a relevant feature divided by the number of all the units). When defining numerical indices in this way, we deal with language "in the mass", without taking into consideration the arrangement (order) of linguistic units in the sample. However, for language users, continuous text is not a random, meaningless sequence of words. A different kind of modelling is possible which takes into account the order of linguistic units in text (in this case we'll speak of "language in the line").[4]

The problem

Even superficial analysis of frequency lists of analytic languages[5] reveals that a very limited number of highly frequent words cover a great part of the text. For instance, the ten most frequent words in Italian cover as much as 32.5% of the text; the analogous value for Spanish is 33.5% and for French it is 30.5% (500,000 word samples).[6] In the case of synthetic languages, the corresponding values are much lower. This can be seen in the example of Slavic languages: the ten most frequent Russian words cover only 18% of the text (1,000,000 word sample); in Polish the analogous quantity is also 18% (500,000 words sample); for Ukrainian it is 17% (500,000 word sample, literary prose language) and for

[4] The expressions "language in the mass" and "language in the line" have been introduced to the vocabulary of QL by Gustav Herdan (Herdan, 1966:423). We use them to clearly distinguish two complementary investigative approaches. When choosing the former, we admit that the order of linguistic units in text is non-relevant; when choosing the latter, we consider this order as a significant feature under investigation (cf. Pawłowski, 1998:50-53). See also Altmann, 1997:16-17; Hřebíček & Altmann, 1993:13-15; Hřebíček, & Altmann, 1996:45-50; Köhler & Galle, 1993.

[5] All natural languages combine both analytic and synthetic characteristics. When calling a language *analytic* or *synthetic*, we always mean analytic or synthetic tendency in language.

[6] Cf. Juilland, Brodin, & Davidovitch, 1971; Juilland, & Chang-Rodriguez, 1964; Bortolini, Tagliavini, & Zampolli, 1971.

Czech it is 18.5% (1,623,527 words sample).[7] In spite of different sample sizes, the disproportion is striking.

The question arises whether or not such a great contrast between the lowest and the highest word frequencies in some languages is reflected in the sequential structure of text and if so, which mathematical tool could be used to model this phenomenon. A closer look at the morphology and syntax of languages under investigation should clear up the question. Most Slavic languages can be characterised by a relatively loose word order in a sentence and a low number of free grammatical morphemes (prepositions and pronouns), the lack of "positional" information being compensated by a rich inflectional system. In analytic languages the reverse can be observed: a relatively rigid word order in a sentence, as well as a greater number of free grammatical morphemes (articles, prepositions and pronouns), make up for the poor inflectional system. In the syntagmatic (linear) perspective, the relevant feature of analytic languages is thus a more or less regular appearance of words of very high frequencies (grammatical morphemes), separated by words of low frequencies (lexical morphemes).

Considering these facts, we advance the hypothesis that the *linear arrangement of words in texts, represented by a quantitative measure related to their frequency, depends on the morphosyntactical structure of a language and, in some cases, is not random*. In languages of synthetic tendency and inflectional syntax, no sequential regularity should be detected whereas in languages of analytic tendency and positional syntax, we expect to find weak stochastic processes, due to the alternate appearance of grammatical (very frequent) and lexical (rare) morphemes. The aim of the present paper is to verify this hypothesis by making the sequential structure of text explicit and building appropriate mathematical models. The parameters of these models are claimed to depend on morphosyntactical types of languages under investigation.

Method of research and quantification

It is assumed that the hitherto existing classifications of languages representing analytic or synthetic tendency are correct.[8] A new method of sequential modelling will be applied to the samples of these languages. If the result obtained confirms this morphosyntactical distinction, i.e. if stochastic processes are discovered in the sequences of words in analytic languages and, on the other hand, no correlation is discovered in the sequences of words in synthetic languages, the proposed method of sequential modelling will be provisionally corroborated and will be susceptible to further tests on other languages.

The other important assumption of the analysis is to consider text as a *time-series*, i. e. a series of units (numbers) arranged along an independent axis (for instance a time axis or, by analogy, text line). Respecting the tradition of structural linguistics, this linear axis will be called *syntagmatic time*.[9] The key feature of this approach is to ignore *proportions* (thus, ultimately, statistical distributions) of linguistic units in the mass of text and to consider as relevant their linear (sequential) arrangement.

The next stage of the procedure is the *quantification of data,* i.e. the conversion of text, being a sequence of categorial (qualitative) units, into a numerical series. This stage is crucial to the analysis. These considerations might suggest that the best method of quantification would be to replace text units (words) by their frequencies (absolute or relative). This solution is adequate but not perfect. The argument of the dependence of word frequencies on sample size could be raised here but in fact this is not a sufficient obstacle.[10] The basic problem is that word frequencies are only numbers which lack a convincing and linguistically meaningful interpretation. The measure fulfilling the criterion of linguistic interpretability is, instead, the *quantity of information* conveyed by successive linguistic units. On the one hand, information is a central notion of many theories of discourse analysis (which consider as relevant the order of units in text),[11] on the other, Claude Shannon's theory assures a satisfying formal representation to the notion of information.[12]

In practice, the proposed method of quantification consists in replacing the successive words of text with their quantities of information (I_n), computed according to the Shannon's formula:

$$(1) \qquad I_n = -\log_2 p_n$$

where the probability p_n is the relative frequency of a unit n within some linguistic universe.

[7] Cf. Засорина, 1977; Kurcz (et al.), 1990; Орлова, & Перебийніс, 1981; Jelínek, Bečka, & Tešitelová, 1961.

[8] We are aware of the objections and hesitations evoked by the question of typology. Still, in QL no experiment could be run without some well determined *initial conditions* and the above assumption should be considered as such.

[9] The notion of *syntagmatic time*, substituted for the notion *real time*, was introduced in our previous studies (Pawłowski, 1997:206-207 and Pawłowski, 1998:4). Other scholars notice the ahistorical character of time in the modelling of social processes: "This practice constructs 'time' to be a linear organising device for a sequence of events or an incremental counter 'marking time' in equal units as in clock or calendar time." (Isaac & Griffin, 1989:875).

[10] One could be concerned because the same linguistic units are assigned different numbers, according to the size of the sample. However, in the proposed method this fact has no importance.

[11] Adam, 1992; Atkinson & Heritage, 1984; Dijk, 1980.

[12] Shannon, 1948; Pawłowski, 1998:87-89.

An interesting problem is the method of calculating the values p_n which should estimate the probability of appearance of a given linguistic unit in the text. In most cases, one must rely on the representative method and approximate them by relative frequencies of linguistic units in a representative corpus of texts. However, in some cases, it is possible to avoid this limitation. For instance, if we consider as a complete and closed entity (thus linguistic universe) the totality of writings by an author, we can say that the values p_n are no longer approximations.[13]

In the present study, the quantities of information *per word* are computed on the basis of the data drawn from *frequency dictionaries* of Polish and Italian.[14] We consider both dictionaries sufficiently representative for processing contemporary texts in these languages. During quantification, we respected all the morphological, lexical and typographic norms used by their authors. The percentage of words encountered in texts which do not appear in the frequency dictionaries is very low (on the average 6%). These are rare lexical words or certain proper names. As the lowest frequency shown in the dictionaries of Polish and Italian is 4, we assign these absent words the frequency 3. The table below includes some illustrative quantities of information conveyed by words having different frequencies in a 500,000 word dictionary:

Table 1

Word frequency	3	5	10	25	50	250	900	1500	2500	5000	9000
Information (bits)	17.35	16.61	15.61	14.29	13.29	10.97	9.12	8.38	7.64	6.64	5.80

The last step of the procedure is the choice of a numerical technique to be applied. As we intend to treat long series of words supposed to convey rather stochastic processes, we suggest the application of the Box and Jenkins ARIMA method of time-series analysis. While it has been worked out for industrial and economical applications, it has also been successfully tested in the fields of social

[13] The experimental method consisting in guessing the unknown (hidden) words of a sentence and finding in this way the quantity of information conveyed by successive words on the 1st, 2nd etc. position cannot be applied here. The object of our research is not a single sentence (or a set of sentences) but the entire text.

[14] Cf. Kurcz (et al.), 1990 and Bortolini, Tagliavini, & Zampolli, 1971.

sciences and linguistics.[15] Among its different uses, the ARIMA method allows modelling of time-series by means of autoregression equations (time-domain modelling), as well as spectrograms (frequency-domain modelling). When dealing with strongly stochastic series (which we expect to find in a quantified text), time-domain analysis will be preferred. Its output is much clearer and convincing, and seems more suitable for linguistic interpretations.

We wish to stress here that the choice of the ARIMA method should not be considered as the only possible method for other quantitative studies of sequential structure of text. We believe that other numerical techniques would yield similar results.

Data

As it mentioned before, we compare samples of languages which are considered as analytic or synthetic, and in this way check whether or not sequential structure of text reflects language analytic or synthetic tendency. Twenty samples of Polish (synthetic) and ten samples of Italian (analytic) are treated.[16] Our findings are then compared with the results for French, English and Italian, obtained in previous studies (Pawłowski, 1998:96-111; Corduas, 1995). Samples treated are composed of about one hundred words of continuous text.

Results

In order to verify the above hypothesis, we transformed text data into numerical time-series, replacing words by the quantity of information they convey. For instance the sentence in Italian "*Io alzai le spalle ed uscii in punta di piedi.*" is represented by the sequence "6.00 11.95 3.19 12.36 5.30 11.19 5.82 11.05 4.30 11.49". As frequencies of words (and, consequently, quantities of information *per word*) remain stable in a given corpus of texts, the time-series obtained in this

[15] A sort of "canonical" reference to the ARIMA method is *Time series analysis: forecasting and control* by G. Box and G. Jenkins (Box & Jenkins, 1967). Since then, numerous studies have been published for the use in social and human sciences (Glass et al., 1975; Whiteley, 1980; Gottman, 1984; Cryer, 1986; Stier, 1989). The first comprehensive study on the application of the ARIMA method in linguistics is the monograph Pawłowski, 1998. All these sources being generally accessible, we do not go into the details of the ARIMA method in the present paper.

[16] Alberto Moravia (1963). *Nuovi Racconti Romani di Moravia.* Bompiani (samples: p.158 (1), 188 (2), 212 (3), 230 (4), 236 (5), 316 (6), 348 (7), 386 (8), 476 (9), 530 (10)). Andrzej Szczypioski (1992). *I ominęli Emaus.* Poznań: Kantor Wydawniczy SAWW (samples: p.16 (1), 35 (2), 66 (3), 72 (4), 88 (5), 112 (6), 130 (7), 146 (8), 156 (9), 172 (10)). Tadeusz Konwicki (1982). *Wniebowstąpienie.* Warszawa: Iskry (samples: p.19 (1), 33 (2), 42 (3), 54 (4), 68 (5), 72 (6), 80 (7), 134 (8), 152 (9), 176 (10)).

way are stationary in a broader sense.[17] This allows modelling the series as a simple (or mixed) autoregressive and/or moving-average stationary process (AR, MA or ARMA process).

The first and, in fact, the most revealing step of the procedure is the computation of empirical autocorrelations (ACF) for the series. In table 2, we present the values of the ACF for lags 1 and 2. It can be noticed that the series of quantities of information *per word* in Polish is not autocorrelated and turns out to be a random sequence.

Table 2
Autocorrelation coefficients for lags 1 and 2 (r_1 and r_2)

		1	2	3	4	5	6	7	8	9	10	*Mean*
Italian	r_1	-0.11	-0.33	-0.34	-0.31	-0.28	-0.27	-0.21	-0.26	-0.23	-0.31	**-0.27**
(M)	r_2	0.10	0.02	0.07	-0.03	0.15	0.06	-0.07	-0.01	0.01	0.05	**0.04**
Polish	r_1	-0.03	-0.09	-0.27	-0.17	-0.07	0.04	-0.02	-0.09	-0.13	-0.06	**-0.09**
(S)	r_2	0.02	-0.07	-0.04	0.18	0.15	-0.08	-0.02	0.14	0.06	-0.03	**0.03**
Polish	r_1	-0.03	0.04	0.08	0.10	0.02	0.15	0.15	-0.26	-0.13	-0.22	**-0.01**
(K)	r_2	-0.06	0.05	-0.07	0.01	0.12	0.02	0.04	-0.12	0.14	-0.09	**0.00**

(M) = Moravia (S) = Szczypiorski (K) = Konwicki

The analysis of the ACF for Italian yields a quite different result: while r_2 is null, r_1 seems to be significant. An approximate standard deviation for the ACF estimate is $1/\sqrt{N-k}$, where N is the length of the series and k is the lag. The confidence interval for the ACF at the level 0.05 (2SD) can thus be defined as $\pm 2/\sqrt{N-k}$.[18] As the average length of the series in Italian is 110 words, the confidence interval for r_1 is not larger than (-0.2; 0.2). Consequently, we conclude that the coefficients r_1 for all the Italian samples (except the first one) differ significantly from zero.

After checking the form of the ACF function, we proceed in building a model for the time-series. In case of a stationary process with one significant value of the ACF (r_1) and a large proportion of noise in data, two possible choices are recommended: *moving-average model* MA(1) and *autoregressive model* AR(1). Despite some hesitations, we choose MA(1) as the best model for time-series

[17] Cf. Priestley, 1981:112.
[18] Cf. Gottman, 1981:67; Pawłowski, 1998:17-18.

based on Italian texts.[19] As was shown, in Polish texts the time-series obtained are not autocorrelated (cf. Tab.2) and, consequently, there is no stochastic process to be modelled. The proposed model for Italian has the form:

(2) $x_t = e_t - b_1 e_{t-1}$

where x_t – the value of the series at the moment t;
b_1 – model coefficient;
e_t – white noise normally distributed $N(0,1)$.

In Table 3, we present the values of b_1 coefficient for Italian. Except the first sample, all others are statistically significant.

Table 3
Model coefficients and 2SD for Italian texts

		1	2	3	4	5	6	7	8	9	10
Italian	b_1	0.11	0.33	0.37	0.28	0.28	0.30	0.20	0.27	0.19	0.27
(Moravia)	2SD	0.18	0.18	0.17	0.18	0.18	0.18	0.18	0.18	0.17	0.17

In order to estimate the goodness of fit of the model (2), we find how much of the original variance in the series it explains. In the table below, three parameters for each sample are presented: σ_R^2 is the residual variance (variance unexplained by the model), σ_T^2 is the original (total) variance in the series and the third value is the percentage of the original variance explained by the model, defined as $100*(1-\sigma_R^2/\sigma_T^2)$. Although the time-series generated from Polish texts are random and there is no deterministic element to be modelled, data for all the samples are presented in Table 4 in order to show that there always exist some exceptional arrangements of words due to chance.

[19] For this purpose we analyse the form of the partial autocorrelation function (cf. Gottman, 1981:142).

Table 4
Percentage of the original variance explained by the model MA(1)

		1	2	3	4	5	6	7	8	9	10	*Mean*
Italian	σ_R^2	19.8	17.1	16.7	17.3	16.4	15.0	18.8	14.4	16.4	17.7	
(Moravia)	σ_T^2	20.0	19.0	19.2	18.7	17.7	16.3	19.5	15.3	17.1	19.2	
Variance explained		1%	10%	13%	8%	7%	8%	4%	6%	4%	8%	**6,9%**
Polish	σ_r	11.3	13.9	13.9	17.4	15.7	14.2	9.6	17.0	15.2	18.4	
(Szczypiorski)	σ_T	11.3	13.9	14.7	17.8	15.7	14.2	9.6	17.0	15.3	18.4	
Variance explained		0%	0%	5%	2%	0%	0%	0%	0%	1%	0%	**0,8%**
Polish	σ_r	15.5	15.1	15.2	13.5	16.6	10.7	13.3	17.1	14.1	15.4	
(Konwicki)	σ_T	15.5	15.1	15.2	13.5	16.6	10.7	13.3	18.0	14.1	16.0	
Variance explained		0%	0%	0%	0%	0%	0%	0%	5%	0%	4%	**0,9%**

Again, the difference between samples in Polish and Italian is apparent. While the percentage of the original variance explained by the model in Polish is almost zero, in Italian, the MA(1) model explains on the average 6.9% of the original variance of the series. Compared with some more deterministic phenomena in language, this value is far from being impressive.[20] However, as it appears very regularly and is statistically significant, we conclude that it confirms the initial hypothesis.

An important argument supporting our findings is the result of other tests carried out on Italian, French and English texts (analytic tendency and positional syntax). In the case of French and English, the results obtained were very similar. It was shown beyond doubt that the series of "quantities of information *per word*" were negatively correlated at the lag 1 and the best fit to the data was obtained with the MA(1) model.[21]

In the study of Italian, the method of quantification was different: instead of Shannon's quantity of information, word lengths (in letters) were substituted for words of text (Corduas, 1995). However, by virtue of Zipf's laws both measures strongly correlated (the most frequent words are usually the shortest ones) and also in this case one should expect similar results. Actually, the form of the ACF presented in the cited study is identical (significant but weak negative autocorrelation at the lag 1). As far as estimated stochastic models are concerned, Corduas proposed not only simple AR(1) but also mixed models (ARMA(4,2), ARCH).[22]

[20] A good example is the rhythm of accentuated and non-accentuated syllables (Pawłowski, 1977).

[21] Ca. 400 samples of French and 40 samples of English were analysed, cf. Pawłowski, 1998:103-104.

[22] *AutoRegressive Moving-Average* model and *AutoRegressive Conditional Heteroscedastic* model.

This fact does not invalidate our result: firstly, we deliberately avoided mixed models which inevitably increase the number of parameters without a substantial improvement of model efficiency; secondly, in the case of some very "noisy" series (like those generated from texts), the difference between AR(1) and MA(1) is not clear cut and some hesitations are fully legitimate.

Conclusion

No quantitative study makes sense without an underlying linguistic hypothesis. The source of the phenomenon described here lies in the morphosyntactical structure of language. Both in analytic and synthetic languages, continuous text is a sequence of words. However, the rules of positional syntax impose on the former a regular use of grammatical morphemes – articles, prepositions and pronouns. For instance, the sequences *preposition–noun*, *article-noun*, *pronoun–verb*, translated into numbers representing information (in bits), produce in most cases more or less regular oscillations. In spite of a very complex structure of text, this *morphosyntactical rhythm* generates a stochastic moving-average process which accounts for ca. 7% of the original variance. This explanation has a strong experimental background: when in a French text grammatical morphemes were omitted, the autocorrelation of the time-series of quantities of information *per word* decreased to zero (Pawłowski, 1998:102). Inversely, the lack of positional constraints in Polish (synthetic, loose word order), results in the random character of time-series generated from texts in this language.

Although the research is based on a limited number of samples, it clearly indicates a general tendency which further tests on multilingual corpora should, we hope, confirm. Coefficients of sequential models could be then applied in linguistic typology.

References

Adam, J.-M. (1992). *Les textes: types et prototypes: récit, description argumentation, explication et dialogue*. Paris: Nathan Université.

Altmann, G. (1997). The Art of Quantitative Linguistics. *Journal of Quantitative Linguistics*, 4, 13-22.

Altmann, G., & Lehfeldt, W. (1973). *Allgemeine Sprachtypologie*. München: Fink.

Atkinson, J.M., & Heritage, J. (Eds.) (1984). *Structure of Social Action. Studies in Conversational Analysis*. London, Paris: Oxford University Press, Maison des Sciences de l'homme.

Batagelj V., Pisanski T., & Keržić, D. (1992). Automatic Clustering of Languages. *Computational Linguistics,* 18, 339-352.

Bortolini, U., Tagliavini, C., & Zampolli, A. (1971). *Lessico di frequenza della lingua italiana contemporanea*. Milano: Garzanti.

Box G., & Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco (etc.): Holden-Day.

Corduas, M. (1995). La struttura dinamica dei dati testuali. In S. Bolasco, L. Lebart & A. Salem (Eds.), *Analisi Statistica dei Dati Testuali, III Journées Internationales d'Analyse Statistique des Données Textuelles*, Rome 11-13 XII, 345-352.

Cryer, J. (1986). *Time series analysis*. Boston: Duxbury Press.

Dijk, van T.A. (1980). *Macrostructures. An Interdisciplinary Study of Global Structures in Discourse, Interaction and Cognition*. Hillsdale: Lawrence Erlbaum Association.

Glass, G.V., Wilson, V.L., & Gottman, J.M. (1975). *Design and Analysis of Time-Series Experiments*. Colorado: Colorado Associated University Press.

Gottman, J.M. (1984). *Time-Series Analysis: a comprehensive Introduction for Social Scientists.* Cambridge: Cambridge University Press.

Greenberg, J.H. (1960). A Quantitative Approach to the Morphological Typology of Language. *International Journal of American Linguistics,* 26, 178-194.

Herdan, G. (1966). *Advanced Theory of Language as Choice and Chance*. Berlin: Springer-Verlag.

Householder, F.W. (1960). First Thoughts on Syntactic Indices. *International Journal of American Linguistics,* 26, 195-203.

Hřebíček, L., & Altmann, G. (1993). Prospects of Text Linguistics. In L. Hřebíček & G. Altmann (Eds.), *Quantitative Text Analysis* (pp. 1-28), Trier: WVT.

Hřebíček, L., & Altmann, G. (1996). The levels of Order in Language. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 38-61), Trier: WVT.

Isaac, L.W., & Griffin, L.J. (1989). Ahistoricism in Time-Series Analyses of Historical Process: Critique, Redirection and Illustrations from U.S. Labor History. *American Sociological Review,* 54, 873-890.

Jelínek, J., Bečka, J.V., Tešitelová, M. (1961). *Frekvence slov, slovních druhů a tvarů v českém jazyce*. Praha: Státní pedagogické nakladatelství.

Juilland, A., Brodin, D., & Davidovitch, C. (1971). *Frequency Dictionary of French Words*. The Hague: Mouton.

Juilland, A., & Chang-Rodriguez, E. (1964). *Frequency Dictionary of Spanish Words*. The Hague: Mouton.

Köhler, R., & Galle, M. (1993). Dynamic Aspects of Text Characteristics. In L.Hřebíček & G. Altmann (Eds.), *Quantitative Text Analysis* (pp. 46-53), Trier: WVT.

Kurcz, I., Lewicki, A., Sambor, J., Szafran, K., & Woronczak, J. (1990). *Słownik frekwencyjny polszczyzny.* Kraków: Polska Akademia Nauk, Instytut Języka Polskiego.

Орлова, Л.В., & Перебийніс, В.С. (ред.) (1981). *Частотний словник сучасної української художньої прози.* Киев: Видавництво Наукова Думка.

Pawłowski, A. (1997). Time-Series Analysis in Linguistics. Application of the ARIMA Method to Some Cases of Spoken Polish. *Journal of Quantitative Linguistics,* 4, 203-221.

Pawłowski, A. (1998). Séries temporelles en linguistique. Avec application à l'attribution de textes: Romain Gary et Émile Ajar. *Travaux de linguistique quantitative,* 62, Paris, Genève: Champion-Slatkine.

Priestley, M.B. (1981). *Spectral Analysis and Time Series*. London: Academic Press.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal,* XXVII, 379-423.

Silnitsky, G. (1993). Typological Indices and Language Classes: A Quantitative Study. In G. Altmann (Ed.), *Glottometrika 14* (pp. 139-160), Trier: WVT.

Stier, W. (1989). Basic Concepts and New Methods of Time Series Analysis in Historical Research. *Historical Social Research/Historische Sozialforschung*, 14, 3-24.

Whiteley, P. (1980). Time Series Analysis. *Quality and Quantity,* 14, 225-247.

Засорина, Л. Н. (ред.) (1977). *Частотный словарь руского языка*. Москва: Издателство русский язык.

Psychiatric Linguistics and Synergetics

Rajmund Piotrowski

Introduction

Although linguists of the last two centuries, beginning from K. Bekker and W. Humboldt, never cease to claim that language is a well arranged organism, the mechanisms providing for the normal functioning of a language in society have not yet been revealed. As has been stressed by Professor Hřebíček, the abstract system of language is

> *"a non-explicit inductive generalisation from the real individual... systems existing in human minds. The idea of individual personal... systems is evidently more realistic... We cannot look into the brain and investigate these systems. One possible way to obtain some information is the psychological approach based on word association. The other way is offered by... text linguistics"* (Hřebíček, 1996:97).

However in a neurophysiologically normal state, human verbal-mental activity (VMA) functions in an astonishingly unified and harmonious way, shutting tightly the "windows" through which an experimenter could observe mysterious "language forces" (Altmann & Köhler, 1996:62-65). Where should we look for ways and means that would enable us to observe this esoteric self-regulation mechanism?

One such possibility is presented by situations in which these systems fail. As the world-famous Russian physiologist Ivan Pavlov (1949:317-318) said, a pathological state opens up to us, picking apart and simplifying that which was concealed from us in a physiologically normal state. So, new and more detailed information on linguistic synergetics could be provided by the study of the pathology of individual speech. The disruptions that seem to be of the greatest interest here are those arising

- in conditions of an altered state of human consciousness (Spivak, 1997),
- as a result of local damage to the brain (Lurija, 1973),

- as a result of endogenous derangements of the brain (schizophrenia, manic-depressive psychosis, genuine epilepsy (Chaika, 1990; Wróbel, 1992).

Unfortunately, in the first two cases it is hard to obtain text and test material which is sufficiently representative and reliable from a semiotic point of view. The problem is that altered states are too brief, and local brain damage is usually connected with a severe medical condition of the patients, in which a prolonged psycholinguistic experiment is impossible. Study of the speech of patients suffering from endogenous problems proves to be more accessible and productive for information science, automatic text processing (ATP), and artificial intelligence (AI). The present article discusses some results from studies of the language of schizophrenics and of other psychiatric endogenous populations. This new field of clinical psycholinguistics is conventionally called psychiatric linguistics (PL).

1. Materials and methods

The basic material with which PL works is an oral or written text produced by the test subject spontaneously or obtained in the form of answers to the experimenter's questions or a questionnaire: a text in which one can find objective evidence of speech and thinking pathology. We will call such a text a pathological text (PT). Classification of a particular text as a PT is a far from simple problem. Let us consider the following Russian texts:

> (1) *Во-первых, гуманитарное хорошее начало. Крестьянин выдвигающий находится доли тоже Турции в восстаниях, наблюдающихся в северном Ираке защитил взгляд „виновные хотя бы немного в этом вопросе турецкое правительство".*
> 'In the first place, humanitarian good beginning. Peasant advancing is found portions of Turkey also in uprisings, observed in northern Iraq defended the view „guilty at least some in this question Turkish government';

> (2) *Автомобильный двигатель, устремленный в дачный ориентир, конечен и метафоричен.*
> 'The automobile engine directed to the dacha landmark is final and metaphorical'.

The first text, which constitutes a "verbal jumble", embodies the disjointed thinking of a patient with schizophrenia. In contrast, the second text is grammatically unexceptionable, though it does contain paradoxical metaphors, which can be evaluated, on the one hand, as a sign of a pathological text. On the other hand, this text can be perceived as a fragment of an avant-garde poem written in blank

verse, or as a fragment of a prosaic text reflecting the flow of consciousness of a hero.

In reality, the first fragment is a word-for-word machine translation of a Turkish newspaper text, while the second text was written by a psychotic patient. Thus, without information about the psychiatric condition of its author, it is practically impossible to classify a specific text fragment as a PT just on the basis of outward deviations from normal content or expression. In order to answer the question of whether or not such a text is pathological, it is necessary to reveal deep indicators, often hidden from direct observation, and therefore not controllable and not consciously imitatable by the author, which signal disruptions in the synergetics of speech and thinking activity.

If we talk about indicators in regard to expression, then they have to be revealed with the help of a statistical information model. What is primarily of interest here are such peculiarities of text formation as a consistent excess of certain styles and sublanguages, quantum distribution of information in the text, and dependence of the multiple meaning of a lexical unit (LU) on its frequency, as measured by its position in a frequency dictionary (FD).

Unfortunately, it is not possible to conduct an elaborate information experiment on guessing a text with psychiatric patients. Therefore, we will turn to statistical analysis of pathological texts, using the classic Zipf-Mandelbrot rank-frequency dependence

$$(1) \qquad f = k / (i + \rho)$$

and its parabolic variant

$$(2) \qquad f = k / i^{(\gamma + g \lg i)}$$

where i is the number (rank) of the LU in the frequency list, f is the expected relative frequency of LU's appearance in the text, k, ρ, γ and g are coefficients depending on the sample size, style, subject, and organisation of the text (Reid, 1944; Hoffmann & Piotrowski, 1979:70-74).

It will be of greatest interest, from the point of view of our task, to represent the behaviour of the parameter $\gamma = tg\varphi$ (φ is the slope of Zipf's curve to the x-axis of a bi-logarithmic graph, see Fig. 1). With an insufficient sample size (N), the recurrence of the most often used LU standing at the head of the frequency list is comparatively low, then the portion of rare LU falling into the "tail" of the list is quite significant.

Therefore, the graph of the Zipf dependence descends gradually to the x-axis, forming an angle $\varphi < 45°$ with $\gamma < 1$. As the sample size increases, the proportion of the most often used LU rises, and the relative significance of rare LU is reduced. At the same time, the average frequency (F) of the LU rises. With the same sample size, F (saturated sample) > F (unsaturated sample). In the course of the saturation of the sample, the angle φ gradually increases, reaching 45° in the so-called ideal Zipf sample, described by dependence (1). Usually it approaches 100,000 text-words. In this case, $\gamma = 1$. As the sample is progressively saturated and it approaches the general population, rare LU move into the medium-frequency zone. The middle part of the bi-logarithmic graph, "the bulge", shifts to the right and up, and the saturated sample thus acquires the form of a convex curve (Fig. 1) described by the parabolic dependence (2). In these cases, $\gamma > 1$ (Piotrowski, 1984:134). The rate of saturation of the sample is determined by the subject area and style of the general population (see below). Thus, for example, L. Brillouin (1960:74) showed that in children's speech, with its smaller lexicon, the sample rapidly reaches its saturation with $\gamma = 1.6$.

Relying on the given scheme, we will consider the compositions written by Russian test subjects independently, and some texts of Russian fiction, scientific and technical prose, military topics.

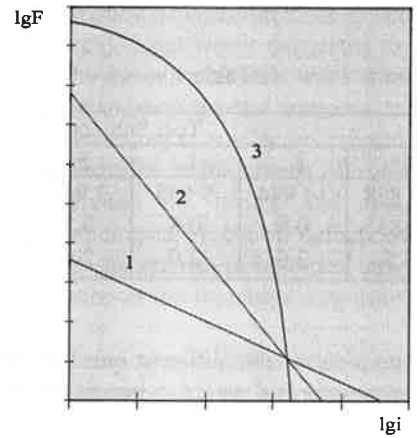


Fig. 1. Typical bi-logarithmic graphs of the rank-frequency dependence for three types of samples:
1) unsaturated sample,
2) ideal Zipf sample,
3) saturated sample (cf. texts recorded from patient Kh., military topics, definitions obtained from the healthy subjects).

2. Statistical analysis of compositions

Six texts were analysed, each of which belongs to one of six test subjects (O., K., L., F., P., and Kh.) suffering from various forms of schizophrenia. The texts contained delirious ideas of scientific and technical invention (O., K., F.), socio-economic reforms (L., P.), and political and everyday predictions (patient Kh.). For each text, we constructed a frequency list, in which the rank-frequency relationship was analysed with the help of dependencies (1-2) (Table 1).

The results that were obtained were compared with analogous data extracted from normal Russian texts (Table 2).

This comparison shows that the texts obtained from the first five test subjects form unsaturated samples with a gently sloping graph of the Zipf dependence ($\gamma < 1$; $2.1 < F < 3.5$). A different statistical structure is provided by the text written by patient Kh., who is characterised by an advanced form of paraphrenia with secondary delirium working with several recurring absurd ideas and false memories. In spite of the limited length ($N < 1000$ textwords) and small volume of the vocabulary (162 w/f), his text is already a saturated sample. This is indicated by the parabolic nature of its Zipf curve (Fig. 1), with $\gamma > 1$ and a large value of F.

Table 1

Statistical features	Test Subjects					
	O.	K.	L.	F.	P.	Kh.
N(w/fs)	30 828	14 944	5 525	2 947	3 000	993
γ	0.83	0.9	0.8	0.7	0.73	1.5
$\overline{F}$(*w / fs*)	3.5	2.9	2.0	2.9	2.2	6.2

Standard Russian texts show also different quantitative structures. The first three samples are of an unsaturated nature, whereas the military topics forms a saturated sample.

Table 2

Statistical features	Sources (FDs) and texts			
	D.N.Mamin-Sibirjak. "*Privalovskie milliony*" (Genkel, 1974)	Texts on wireless technology (Mežlumova, 1973)	Frequency Dictionary of Russian Words, 1977	Military topics (Kolguškin, 1970)
N(w/fs)	103 941	400 000	1 056 382	689 214
γ	1.0	1.05	1.0	1.4
$\overline{F}$(*w / fs*)	9.2	57.0	29.0	229.7

The differences in purpose and subject of the pathological and non-pathological texts under consideration mask deep disruptions in their synergetics. Therefore, it is necessary to continue our statistical experiment on pathological and control texts that, being generated by uniform stimuli, are characterised by narrow semantics and pragmatics. This requirement is met by texts that are definitions of the meanings of individual words suggested to the test subjects. The latter, naturally, describe the meaning of each word stimulus differently. However, the common motive stimulating the test subject's answers dictates the use of definitions constructed according to similar lexical and grammatical patterns. Significant statistical discrepancies in the choice of vocabulary and syntax can reflect the distinction between the lexical and syntactic organisation of "pathological" and "non-pathological" (control) definitions, and can be seen as a reflection of differences that characterise pathological and normal VMA from the point of view of the interaction of the thesaurus, linguistic competence, and control mechanisms.

From the above line of reasoning a further psycholinguistic experiment was carried out to reveal a difference in describing the meanings of test words by healthy subjects and psychotic patients.

Four groups of test subjects were asked to define in writing the meanings of the nouns *birch, bread, life*, and also such words as *quickly, rocky, to fly, cannot, to release, alongside*, and *dry*.

Subjects were 200 males with secondary, unfinished higher, or higher education from 18 to 40 years old, all unpaid volunteers, of whom 19 in the initial stage of paranoid schizophrenia (first hospitalisation) formed the first experimental group. The second group included 20 patients with pronounced hallucinatory paranoid symptoms. The third group was made up of 49 patients with schizophrenia The fourth group was formed of 112 psychiatrically healthy test subjects.

The set of definitions given by each group of test subjects forms an individual sample that is investigated on the Zipf rank-frequency scheme. The results of this investigation (Table 3) show that the first three sets of texts, which were obtained from the psychologically disturbed test subjects, are unsaturated samples ($\gamma < 1$), while the set of non-pathological definitions obtained from the healthy test subjects forms a saturated sample, which is indicated by high values of the parameters γ (> 1) and $\overline{F}$.

Table 3

Statistical features	Groups of subjects			
	1st	2nd	3rd	Control
N(w/fs)	490	930	1758	2594
γ	0.6	0.62	0.76	1.11
$\overline{F}$(*w / fs*)	2.05	2.1	2.28	5.04

Thus, we have a different picture for the definition test. The statistical results show that the first three groups of the subjects are unsaturated samples ($\gamma < 1$). While the set of non-pathological definitions (the 4th healthy subject group) forms a saturated sample ($\gamma > 1$).

Turning to the interpretation of the results of both experiments, we will recall that the statistical characteristics of a text are fairly mobile. They can depend on the education, social status, artistic intention, and even the mood of the communicants. In this connection, let us consider two extreme cases.

In collectives where life is governed by a rigid order (e.g. military collectives), speech activity is also strictly regulated. The texts created here (orders, dispatches, etc.) are usually characterised by a uniform thematic purpose and a unified dictionary. Therefore, comparatively small sets of such texts consist of words and *w/fs* repeated many times, thus forming saturated samples. The FD of Russian combat documents can serve as an example (Table 2).

In contrast, texts created in conditions of unregulated communication contain a diverse vocabulary. Their FD begins to show signs of saturation only in very large samples. The FD of French literary language of the nineteenth and twentieth centuries, which was constructed from a sample of 71 million running words (Dictionnaire des fréquences, 1971). Therefore, we can expect that the compositions of our test subjects on a freely chosen topic will not provide saturated FD, while the texts of definitions constructed on standard patterns of the type *birch - a tree with white bark* or *life is the biological condition of an organism opposite to death* should produce rapid saturation of the FD.

3. Statistical data

FD data of certain public, conversational and scientific texts as well as the analysis of works written by psychologically disturbed test subjects O., K., L., F. and P. confirm this hypothesis: in all the cases we deal with unsaturated samples. Quick PD saturation, made up from the definition of healthy test subjects, also conforms to this hypothesis. At the same time, the FD definitions produced by sick test subjects, on the one hand, and the small saturated text of the patient Kh. who suffers from an advanced form of schizophrenia, on the other hand, contradict the hypothesis.

To appreciate the cause of this discordance let us now look at the more specific pathogenetic traits in the informational processing of schizophrenics. Empirical studies of abnormal verbal behaviour show that schizophrenics seem to process LU meaning with less than healthy subjects' reliance on semantic, logical and conceptual cues. Thought disorder research attests to the excessive orientation of schizophrenics to the superficial and perceptually vivid, rather than semantically and taxonomically pertinent attributes such as affective tone of message (Fine et al., 1991), and the phonetic and connotative properties of a word (Kay, 1982:155; Wróbel, 1992; Piotrowski et al., 1994:32; Fine, 1995:25-41).

Weakening of emotional-volitional processes, autism and desocialization have a destructive influence on the synergetic mechanism of VMA co-ordination and regulation. Conventionally we shall call this mechanism *communicative pragmatic operator* (CPO). The CPO is a regulator or "mould" that supports the process of message generation and of its perception and analysis at a certain level and brings about convergence of various LUs at the given level (Piotrowski, 1994:19-20).

Coming back to the results of our experiment, one cannot but notice that the absence of a saturation in the frequency list of definitions produced by sick subjects is explained by the fact that the faulty CPO of a test schizophrenic cannot keep the definition generating process within the limits of a scheme: "word - its short definition", - the scheme that is required by an experimenter. The sick test subject tries to overcome this contradiction in different ways. He either repeats the stimulus word (*лететь – это лететь* 'to fly means to fly'), trying to get involved in the experimenter's scheme, or ignoring the scheme, he develops, in his definitions, additional topics, produced by his own delirious way of thinking. In the last-mentioned case non-standard lexico-grammatical constructions are used, such as: *лететь - быстро бежать, от медленного отличается то, что один синус говорит быстро идти* 'to fly - run fast, differs from a slow that one sine says go fast'. The definitions formed in this way make up a pathological text containing particular constructions with different kinds of rare vocabulary. Such a text, of course, does not give any FD saturation.

The CPO fault reveals itself differently in the texts of the subject Kh., whose PT forms a structure similar to the statistical organisations of business or combat texts and definitions produced by healthy test subjects. This patient suffers from a schizophrenic disorder with systematised paranoid delirium, in whose frame functions his VMA. That is why the faulty CPO forms here a pathological text consisting of a limited variety of words and word-combinations which can purposefully realise his delirious topic. Thus, the fact that our patient has rapidly formed a saturated sample displays his advanced form of schizophrenia with deep monothematic delirium that is a result of damage to the synergetic mechanism of VMA.

As far as the subjects O., K., L., F. and P. are concerned, who do not have systematised delirium, the vocabulary used in their works is more varied. As a result, their PTs are unsaturated and their statistical structure is similar to that of nonpathological texts. The unsaturated character of definition samples of the three sick subjects groups also reflects a synergetic defect in their verbal/mental mechanism. The reason for dispersion and diversity of words and phrases in pathological definitions is that the sick subject uses an unorganised chaotic method of making up a definition.

To understand the details of CPO functioning, and particularly the mechanism of extracting lexical items from the long-term memory in the process of generating a text, we need to conduct more subtle tests. An experiment on guessing words pursued by E. Paszkowski (1994:113-117) serves this particular purpose.

4. Word guessing

The experiment was in essence as follows: in order to guess an "unknown word" as quickly as possible the subjects asked the experimenter several questions in logical (or unorganised) succession. The words were planned by the researcher himself. The subjects were 40 males with unfinished higher education from 20 to 40 years old, all unpaid volunteers, of whom 20 were paranoid schizophrenics, and 20 healthy test subjects.

The sequences of questions addressed by each test subject to the experimenter form two sample populations; the sample of the patients' questions (*Np* = 3247 text-words) and the sample of the healthy test subjects' questions (*Nh* = 1852 text-words). Each sample is characterised by its own vocabulary *V*, i.e. a list of various word forms. In this case, *Vp* = 818 *w/fs*, and *Vh* = 463 *w/fs*. The sizes of the patients' sample and vocabulary are more than one and a half times greater than the sample and vocabulary of the healthy test subjects: *Np* : *Nh* = 1.75; *Vp* : *Vh* = 1.77.

In other words, the sick test subjects spend noticeably greater effort on guessing the control word than the healthy subjects. The reason for these discrepancies is that the sick and healthy test subjects use different guessing strategies. The former either take the path of unorganised sorting through specific traits of the supposed subject or concept (in this case, the relationship of specific and abstract questions is already significant in the first search zones: see Table 4 and Fig. 2, or they unsystematically list words that could designate the object of the search. In this case, the questions may be repeated.

Table 4

Search zone (Nos of questions)	Portion of specific questions (%)	
	Healthy test subjects	Sick test subjects
1-5	1.7	13.7
6-10	13.3	27.3
11-15	22.3	32.0
16-20	27.0	33.0
21-25	25.0	21.3
26-30	46.0	40.0

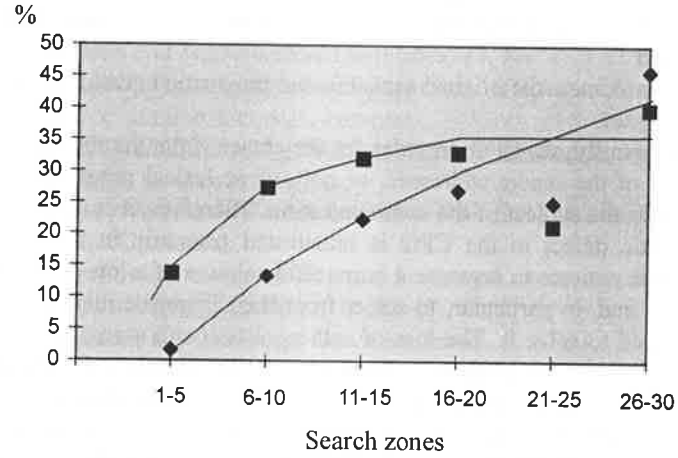


Fig. 2: Increase of specific questions (%) in the course of guessing unknown control words *ключ* 'key', *нож* 'knife', *часы* 'watch' (Paszkowski et al., 1994:113-117)

This indicates that the psychotic patient is not taking into account the information that he received in the preceding stages of guessing. The healthy subjects take a different path. Their strategy relies on a topology of questions in which they first clarify generic traits of the meaning of the word to be guessed. In this case, the number of specific questions at the beginning of guessing is comparatively small. Having discovered generic traits, the test subject switches to clarification of specific and other particular semantic traits. Thus, the results of the preceding search are not forgotten, but are used in the course of progressing toward the goal.

Conclusions

Statistical and test investigation of PT has enabled us to outline certain peculiarities of the synergetic mechanism of generating a text. The hypothesis according to which the central block of this mechanism is some communicative pragmatic operator is confirmed. It will be recalled that this CPO can be classified as a regulator or "mould" that supports the process of message generation and of its perception and analysis at a certain level, and brings about convergence of various LUs at the given level (Piotrowski, 1994:19-20). Our experiment reinforces the statement that the CPO is called upon

- to establish a purposefulness in the process of generating and recognising a message, and
- to approximate the original semantic and pragmatic intention of the sender by the addressee.

Simultaneously, the CPO provides for the choice, from the thesaurus and the competence of the sender/addressee, of only those lexical pragmatic units that correspond to the subject of the communication. Therefore, it is no accident that the synergetic defect in the CPO is manifested primarily in the inability of schizophrenic patients to organise a purposeful solution of a logical word guessing problem and, in particular, to select from their linguistic memory the information needed to solve it. The loss of self-regulation and normal control of the speech and thinking automatism is what creates the conditions for disturbed interpretation of external impressions and internal feelings (Zurabašvili, 1975:110). That is the psychiatric aspect of statistical information investigations of PT. Returning to the problems of linguistic synergetics, we can state that linguistic-psychiatric research is capable of explaining some details of the language synergetics that remain inaccessible in studying non-pathological speech.

These investigations are no less promising from the point of view of language engineering and of AI linguistic problems (Servan-Schreiber, 1986:191-201). In fact, for all of the refinement of linguistic-engineering formalism, when modern MT systems work with randomly taken text arrays not subjected to preliminary editing, they produce translations reminiscent of the "verbal jumble" that appeared in the MT of the Turkish newspaper article cited above. Frame patterns used in systems of annotation and abstracting (Apollonskaya et al., 1983:40-52; Kuhns, 1990) for organising a disintegrating output text can, of course, be seen as primary analogs of a CPO. However, even they are devoid of genuine synergetic capabilities of self-regulation and adaptive control. Construction of adequate synergetic models can begin when a sufficient store of knowledge about the self-organisation of a language system and of human VMA is accumulated.

Acknowledgements

I am indebted to W. Paszkowski, W. Piotrowska, Yu. Romanov, A. Stubbe for helpful comments, references and resources.

References

Altmann, G., & Köhler, R. (1996). "Language Forces" and Synergetic Modelling of Language Phenomena. In P. Schmitt (Ed.), *Glottometrika 15* (pp. 62-76), Trier: WVT.

Apollonskaya, T.A., Koliban, V.V., Piotrowski, R.G., & Popescul, A.N. (1983). Using Frames for Automatic Abstracting of French Patent. *Automatic Documentation and Mathematical Linguistics,* 17, No. 2.

Brillouin, L. (1960). *Наука и теория информации.* (Перевод с английского). Москва: Гос. Изд-во физ.-мат. лит-ры.

Chaika, E. (1990). *Understanding Psychotic Speech: Beyond Freud and Chomsky.* Springfield, Il.: Charles Thomas.

Dictionnaire des fréquences. Vocabulaire litteraire des XIX-e et XX-e siecles. (1971). Paris: Klincksieck.

Fine, J. (1995). Toward Understanding and Studying Cohesion in Schizophrenic Speech. *Applied Psycholinguistics*, 16, 25-41.

Fine, J., Bartolucci, G., Ginsberg, G., & Szatmari, P. (1991). The Use of Intonation to Communicate in Pervasive Developmental Disorders. *Journal of Child Psychology and Psychiatry*, 32, 771-772.

Частотный словарь русского языка (1977). Л.Н. Засорина (ред.). Москва: Русский язык. (Frequency Dictionary of Russian Words).

Генкел, М.А. (1974). *Частотный словарь романа Д. Н. Мамина-Сибиряка "Приваловские миллионы"*. Перм: Пермский Гос. университет им. А. М. Горького. (Frequency Dictionary of D.N. Mamin-Sibiryak's Novel "Privalovskie milliony". 'the Privalov millions').

Hoffmann, L., & Piotrowski, R.G. (1979). *Beiträge zur Sprachstatistik.* Leipzig: VEB Verlag Enzyklopädie.

Hřebíček, L. (1996). Word Associations and Text. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 96-101), Trier: WVT.
Kay, S.R. (1982). Conceptual Disorder in Schizophrenia as a Function of Encoding Orientation. *The Journal of Nervous and Mental Disease,* 170, no. 3.
Колгускин, А.Н. (1970). *Лингвистика в военном деле. (Разработка и использование частотных словарей военной лексики.)* Москва: Военное издательство Министерства обороны СССР.
Kuhns, R.J. (1990). New Analysis: A Natural Language Application to Text Processing. *American Association for Artificial Intelligence Spring Symposium Series, Text Based Intelligent Systems.* Palo Alto, Calif.: Stanford Univ.
Лурия, А.Р. (1973). *Основные проблемы нейролингвистики.* Москва: Изд-во MGU. (Basic Problems of Neurolinguistics).
Межлумова, А.Б. (1973). *Статистическая характеристика лексики и морфологии русских текстов по радиотехнике.* Автореф. канд дисс. Минск: Белорусский Государственный университет. (Lexico-morphological Statistical Features of the Russian Wireless Technique Texts).
Paszkowski, W. E., Piotrowska, W. R., & Piotrowski, R.G. (1994). *Психиатрическая лингвистика.* Санкт-Петербург: Наука.
Pavlov, I.P. (1949). *Лекции о работе больших полушарий головного мозга.* Лекция 18. Полное собрание сочинений. Том 4. Москва - Ленинград: Изд-во АН СССР. (Lectures on the Work of the Cerebral Hemispheres. Lecture 18. *Complete Works*, Vol. 4, AN SSSR).
Piotrowski, R. (1984). *Text - Computer - Mensch.* Bochum: Brockmeyer.
Piotrowski, R. (1994). Psycholinguistic Basis of the Linguistic Automaton. *International Journal of Psycholinguistics,* 10, 1.
Piotrowski, R.G., Pashkowski W.E., & Piotrowski, W.R. (1994). Psychiatric Linguistics and Automatic Text Processing. *Automatic Documentation and Mathematical Linguistics,* 28, 5.
Reid, I.R. (1944). French Word-frequency Distribution Curve. *Language*, 20, 4.
Servan-Schreiber, D. (1986). Artificial Intellegence and Psychiatry. *The Journal of Nervous and Mental Disease,* 17, 3.
Spivak, D. (1997). Quantitative Measurement in Linguistics of Altered States of Consciousness. *Proceedings of the XVIth International Congress of Linguists.* Paris, July 20-25 1997.
Wróbel, J. (1992). *Language and Schizophrenia.* Amsterdam: J. Benjamins.
Зурабашвили, А.Д. (1975). *Теоретические и клинические искания в психиатрии.* Тбилиси: Мецинереба. (Theoretical and Clinical Research in Psychiatry).

Sentence Length in Old Church Slavonic

Otto A. Rottmann

0. The present paper is intended as a first attempt to analyse sentence length in Old Church Slavonic (or Old Bulgarian, the predecessor of Modern Bulgarian). It will not include a general discussion of sentence length studies (this will be done in a forthcoming monograph), but just present first results.

1. Old Church Slavonic is a language whose origin dates back to the ninth century A.D. It differs in some respects from other old Indo-european languages which have already been examined for the same purpose (e.g. Latin, Greek): Old Church Slavonic was never spoken, it was a language used to write the Bible and other holy texts such as legends in a language which was understood by all Slavic people, who in their everyday life spoke their own native, Slavic languages; in other words, it was a mere *written language*. This means that all available texts had their origin in the Greek language from which the Bible and other holy texts were translated.

2. Old Church Slavonic sentences mostly have a very simple structure; word order in sentences / clauses is arbitrary. For our study the following criteria were applied: counting is based on clauses, a clause is a syntactic unit centering around a predicate. This predicate can be a finite verbal form or a non-finite forme replacing a finite verbal form (participial construction with a present participle active or past participle active if subjects in main and subordinate clauses are identical; "dativus absolutus", i.e. dative absolute, if subjects in main and subordinate clauses are different; infinitive and supine constructions).

The analysis comprised the following texts, all of which were taken from Leskien's *Handbook of Old Bulgarian* (adapted texts, with adaption meaning that words are not written in their typically abbreviated form, which, however, is irrelevant for sentence length analysis):

a) Luke VI, originally included in the Codex Zographensis,
b) Luke V, originally included in the Codex Zographensis,
c) Psalm 103, originally included in the Psalterium Sinaiticum,

d) the Isaacios legend, originally included in the Codex Suprasliensis,
e) the Chrysostomos homily, originally included in the Codex Suprasliensis,
f) the Euchologium Sinaiticum 50b, originally included in the Codex Zographensis,
g) the Basiliskos legend, originally included in the Codex Suprasliensis,
h) Luke VII, originally included in the Codex Zographensis,
i) Luke XIII, originally included in the Codex Zographensis,
j) Luke IX, originally included in the Codex Zographensis,
k) Luke X, originally included in the Codex Zographensis,
l) Luke XI, originally included in the Codex Zographensis,
m) Luke XII, originally included in the Codex Zographensis,
n) Luke VIII, originally included in the Codex Zographensis.

3. Data

In his paper on sentence length analysis, Altmann (1988) proposes the exploitation of the truncated (= positive) negative binomial distribution; as, however, this distribution converges under special conditions observed in our data ($k \to \infty$, $q \to 0$, $kq \to a$) to the positive Poisson distribution, here merely this limiting case, i.e.

$$P_x = \frac{a^x}{x!(e^a - 1)}, \quad x = 1,2,3,\ldots \text{ will be used.}$$

The following data are specified for each text:

x	sentence length (the number of clauses)
f_x	number of sentences having length x in a text
NP_x	theoretical values in compliance with the formula of the positive Poisson distribution
a	parameter
X_k^2	chi-square with k degrees of freedom
P	the probability of chi-square

	Text 1		Text 2		Text 3	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	9	9.82	16	14.83	2	10.46
2	17	14.22	16	17.45	19	11.05
3	12	13.73	15	13.69	12	7.78
4	9	9.94	7	8.05	2	4.11
5	8	5.76	3	3.79	1	2.59
6	1	2.78	0	1.49		
7	2	1.75	2	0.50		
8			0	0.15		
9			1	0.05		
	$a = 2.8960$		$a = 2.3534$		$a = 2.1123$	
	$X_5^2 = 2.97$		$X_4^2 = 0.95$		$X_3^2 = 4.36$	
	$P(X_5^2) = 0.71$		$P(X_4^2) = 0.92$		$P(X_3^2) = 0.11$	

	Text 4		Text 5		Text 6	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	3	2.24	11	8.18	3	4.00
2	4	5.00	18	14.32	10	5.17
3	9	7.43	18	16.69	1	4.46
4	8	8.29	7	14.60	1	2.89
5	7	7.40	5	10.21	3	1.49
6	4	5.50	7	5.96	1	0.99
7	3	3.51	6	2.98		
8	3	1.96	1	1.30		
9	2	1.67	0	0.51		
10			2	0.26		
	$a = 4.4614$		$a = 3.4982$		$a = 2.5869$	
	$X_7^2 = 1.92$		$X_6^2 = 12.32$		$X_1^2 = 4.31$	
	$P(X_7^2) = 0.96$		$P(X_6^2) = 0.06$		$P(X_3^2) = 0.04$	

	Text 7		Text 8		Text 9	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	30	26.30	23	23.97	12	9.67
2	21	29.94	31	25.65	7	11.88
3	26	22.72	15	18.29	13	0.73
4	14	12.93	12	9.79	8	5.98
5	4	5.89	0	4.19	0	2.94
6	2	2.23	0	1.49	2	1.80
7	1	0.73	0	0.46		
8	0	0.21	1	0.12		
9	1	0.05	1	0.03		
10	1	0.01	1	0.01		
11	0	0.00				
12-19	0	0.00				
20	1	0.00				
	$a = 2.2764$ $X_4^2 = 6.72$ $P(X_4^2) = 0.15$		$a = 2.1398$ $X_4^2 = 6.81$ $P(X_4^2) = 0.15$		$a = 2.4571$ $X_4^2 = 7.31$ $P(X_4^2) = 0.12$	

	Text 10		Text 11		Text 12	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	27	24.94	15	15.25	16	17.66
2	22	28.62	22	18.18	25	21.23
3	25	21.90	8	14.45	13	17.02
4	13	12.57	12	8.61	15	10.24
5	8	5.77	0	4.10	3	4.92
6	1	2.21	3	1.63	1	1.97
7	0	0.72	3	0.78	1	0.95
8	1	0.28				
	$a = 2.2956$ $X_4^2 = 3.47$ $P(X_4^2) = 0.48$		$a = 2.3836$ $X_3^2 = 5.06$ $P(X_4^2) = 0.17$		$a = 2.4052$ $X_4^2 = 5.04$ $P(X_4^2) = 0.28$	

	Text 13		Text 14	
x	f_x	NP_x	f_x	NP_x
1	17	14.87	22	16.94
2	24	19.87	17	22.99
3	13	17.71	22	20.81
4	6	10.24	10	14.13
5	9	4.92	12	7.67
6	4	1.97	3	3.47
7	2	0.95	1	1.35
8			1	0.64
	$a = 2.6725$ $X_5^2 = 7.03$ $P(X_5^2) = 0.22$		$a = 2.7154$ $X_5^2 = 6.86$ $P(X_5^2) = 0.23$	

These tables show that the positive Poisson distribution generally applies. Three texts turned out to be a little difficult to handle: 3, 5 and 6. As text 6 is rather short, the value $P = 0.04$ obtained is low, but acceptable. The analysis of texts 3 and 5 results in the assumption that the exploitation of a more complex model, the hyper-Pascal distribution, might be recommended, but again the model chosen can be considered acceptable.

References

Altmann, G. (1988). Verteilungen der Satzlängen. In K.-P. Schulz (Ed.), *Glottometrika* 9 (pp. 147-170), Bochum: Brockmeyer.

Leskien, A. (1968). *Handbuch der Altbulgarischen Sprache*. Wiesbaden: Winter.

Towards an Evaluation of the Conceptual Level of a Term

S. D. Šelov

Recent publications in terminology research demonstrate constantly growing interest in the problem of term definition analysis. Definitions outline the semantics of terms and set up their logical and semantic relations. According to A. Rey, term definition is probably the very centre of terminological problems (Rey, 1979:9). Besides, the ever growing interest in terminological definitions could be at least partly explained by the facilities that information and term database would offer in case proper conceptual analysis is applied to terminological definitions and provides a database with highly reliable data in a well structured and machine-readable form. The opportunity to get important information directly from definitions opens rather promising perspectives in new computer technologies (Jose & Finatto, 1995; Martin, 1992; Meyer, Bowker & Eck, 1992; Sager & L'Homme, 1994; Sager & Ndi-Kimbi, 1995; Shelov, 1996; Šelov, 1998a; Šelov, 1998b)

In terminological practice, the most common definitions are described in logical studies as explicit definitions, i.e. those which explicitly include both the defined and defining components (Definiendum and Definiens, Dfd and Dfn). These definitions can be considered from various points of view. The definiens text (henceforth Dfn text) can be analysed, let us say, by considering its entire lexical, syntactic and semantic structure. A comparison of the common and different elements in the Dfn for different terms would then reveal their significant logical and semantic characteristics. This is important for studies of terms as elements of a field of knowledge; indirect relations of terms linked through a common semantic component are included in the analysis, as well as their direct components. Methods of this research are general enough to be applied to the common word stock as well as to terminology.

However this type of linguistic analysis requires rather detailed and intricate analytical procedures. Dfn text can be long and complex, including graphics, formulas, symbolic, and other elements, i.e. Dfn expression for some important linguistic terms consists of as many as 50 to 80 words. In its non-terminological

common language part a Dfn text can include diverse but synonymous lexical-syntactic structures. The contents of a Dfn (and, therefore, of the term defined), however, will remain unchanged, making a deep semantic analysis thus necessarily completely resistant to textual synonymy and ambiguity; its results should be independent of it. Achieving this at the current level of linguistics is an independent and extremely difficult problem.

Thus, terms can be defined through other terms in a more stable fashion than when they are defined through the words of the common language and under different circumstances one can consider only direct logical and semantic relations of terms, analysing only the facts of definability and its semantic aspects. For many purposes a much more simple analysis of Dfn texts is possible by means of determining which terms can be defined through other terms and in what way. Such a study is justified both as an element of an investigation of the systemic and semantic properties of terminology and as an object of research in its own right.

In previous works we have tried to demonstrate that describing the definability of terms through other terms is important for modelling the terminological properties of words and phrases (Shelov, 1990). By evaluating definability we could also show some logical and linguistic characteristics of the subject field and fruitful applications in thesauri construction (Shelov, 1982). Since terms are informatively the most valuable lexical units of a special text and serve as special codes for theoretical concepts of science, descriptions of the definability of terms and its semantic aspects establish a logical-semantic structure of terminology cf the concepts of "logical scheme of science" and "logical form of understanding the world" (Lotte, 1961:14).

In recent publications we have endeavoured to outline a method of term definition analysis that enables us to determine 1. what part of a Dfn denotes the nearest generic concept and 2. what parts of a Dfn denote differentiating characters (Shelov, 1996; Šelov, 1998a). Basically this method was oriented to a more or less refined definition system developed for some computer applications since a "relatively free-text form of most definitions is not normally suitable for effective use in a database environment" (Sager & L'Homme, 1994:352). So we have assumed that there exists a consistent, logically and linguistically irreproachable definition system for terms of a given domain. This assumes in turn that any ambiguity or synonymy of the Dfn expressions is eliminated. It also implies that every common word of the Dfn expression has one and the same meaning, every syntactic relation is restricted to one and the same semantic relation, and not that a single meaning is expressed in different ways, etc). Then, being applied to this normalised definition system (or at least similar to it), a rule that enables us to parse the Dfn text into nearest generic term and phrase denoting differentiating characters runs:

The nearest generic concept is denoted by a minimum (if counted in autonomous words) semantically accomplished and syntactically independent part of the Dfn that includes maximum (if counted in autonomous words) term already introduced in a subject field. The remaining part of the Dfn denotes differentiating characters of this generic concept; there is only one differentiating character if the rest part of the Dfn syntactically relates to only one word, and there are 'n' (conjunct) differentiating characters if the rest part of Dfn syntactically relates to 'n' different words.

Later it turned out to be an extra argument in favour of the great importance of term definability through other terms as it should be kept in mind when classifying different definitions. To demonstrate this let us look at some definitions borrowed from the following sources:

1. Rosenberg, J.M., *Dictionary of Computers, Information Processing & Telecommunications*, 2nd ed., V, 1–5, N.Y. et al.: John Wiley & Sons, 1987 (in abbreviated form below – COMP);
2. *Glossary of Heat Treatment*. Swedish Centre for Technical Terminology, TNC 57E, Stockholm: TNC, 1974 (in abbreviated form below – HEATTR);
3. *Personal Communications Terminology*. American National Standard for Telecommunications, N. Y.: ANSI, 1996 (in abbreviated form below – PERCOM). Here are some definitions from the sources (Dfd of the definitions below is printed in italics, Dfn is printed in ordinary font):

1. *Parallel computer.* A computer having multiple arithmetic or logic units that are used to accomplish parallel operations or parallel processing (COMP).
2. *Computer micrographics.* Methods and techniques for converting data to or from micro-form with the assistance of a computer (COMP).
3. *Computer architecture.* The specification of the relationships between the parts of a computer system (COMP).
4. *Computer-assisted management.* Management performed with the aid of automatic data processing (COMP).
5. *Austenitizing.* Heat treatment for the purpose of altering a structure to a more or less pure austenitic state (HEATTR).
6. *Blue brittleness.* Condition caused by embrittlement in connection with the precipitation of foreign phases in a material of given composition and given temperature (HEATTR).
7. *Critical cooling rate.* The lowest cooling rate at which an undesired transformation will not occur (HEATTR).

8. *Equilibrium diagram.* Graphic representation of the range of occurrence for a balanced system's phases expressed as a function of temperature, pressure and composition (HEATTR).
9. *Soaking time.* Period of time during which a material subjected to heat treatment remains at the required temperature (HEATTR).
10. *Heat treatment.* Application of a combination of *heating, holding* and *quenching* (or cooling, holding and heating) to a solid *material* below its melting point in order to affect the properties of the *material* in the manner desired (HEATTR).

Definitions 1 – 10 are generic definitions, but how can we make sure that they are really generic?

It is worth mentioning that in the sources under consideration 1. the term *parallel computer* is defined through the term *computer*; 2. the term *computer micrographics* through the terms *data* and *micro-form*; 3. the term *computer architecture* through the term *system*; 4. the term *computer-assisted management* through the term *data processing*; 5. the term *austenitizing* through the term *heat treatment*; 6. the term *blue brittleness* through the term *embrittlement*, 7. the term *critical cooling rate* through the term *cooling rate;* 8. the term *equilibrium diagram* through the term *phase;* 9. the term *soaking time* through the term *material;* 10. the term *brittleness* through the term *material.*

So, according to the rule above, for definitions 1 – 10 we finally get the following results of the genus-species analysis:

1. 'computer' – THE NEAREST GENERIC CONCEPT, 'having multiple arithmetic or logic units that are used to accomplish parallel operations or parallel processing' – DIFFERENTIATING CHARACTER;

2. 'methods and techniques for converting data to or from micro-form' – THE NEAREST GENERIC CONCEPT, 'with computer assistance' – DIFFERENTIATING CHARACTER;

3. 'the specification of the relationships between the parts of a system' – THE NEAREST GENERIC CONCEPT, 'computer' – DIFFERENTIATING CHARACTER;

4. 'management performed with the aid of data processing' - THE NEAREST GENERIC CONCEPT, 'automatic' – DIFFERENTIATING CHARACTER;

5. 'heat treatment' – THE NEAREST GENERIC CONCEPT, 'for the purpose of altering a structure to a more or less pure austenitic state' – DIFFERENTIATING CHARACTER;

6. 'condition caused by embrittlement' – THE NEAREST GENERIC CONCEPT, 'in connection with the precipitation of foreign phases in a material of

given composition and given temperature' – DIFFERENTIATING CHARACTER.

7. 'cooling rate' – THE NEAREST GENERIC CONCEPT, 'the lowest' and 'at which undesired transformation will not occur' – DIFFERENTIATING CHARACTERS;

8. 'graphic representation of the range of occurrence for a system's phases' – THE NEAREST GENERIC CONCEPT, 'balanced' and 'expressed as a function of temperature, pressure and composition' – DIFFERENTIATING CHARACTERS;

9. 'period of time during which a material remains at the required temperature' – THE NEAREST GENERIC CONCEPT, 'subjected to heat treatment' – DIFFERENTIATING CHARACTER;

10. 'application of a combination of *heating, holding* and *quenching* (or cooling, holding and heating) to a *material* – THE NEAREST GENERIC CONCEPT, 'solid' and 'below its melting point in order to affect the properties of the *material* in the manner desired' – DIFFERENTIATING CHARACTERS.

This analysis proves definitions 1 – 10 to be generic.

The argument above suffices to put the property of term definability under more detailed analysis than it has usually been given. Moreover, here we will discuss the claim that this property underlies the idea of a *conceptual level of a term* which can be generalised to take into account totally different conceptual structures of terminology. To develop a formal model of the conceptual structure of terminology, we will apply to it some concepts of *graph theory* since "the graph theory as a method of linguistic description corresponds in its basic features to the principle of simplicity of the system" (Hřebíček, 1971:9)

So we will talk of the direct definability graph $< (x, y)$ meaning that $x < y$ if and only if the term x is directly defined through y.

Given the concept of direct definability graph, one may apply some standard concepts of graph theory to the direct definability graph $< (x, y)$ and the concepts of the "path" and "length of the path" in particular. According to common logical requirements a system of explicit definitions must not contain circles (circulus vitiosus) and, besides, none of the terms can be defined through itself. Using the standard vocabulary of graph theory this means that a direct definability graph has no circles and no loops. This claim is enough to propose the following definition:

Let us define a conceptual level $L(x)$ of the term x as the maximum length of all the paths from any term to the term x in $< (x, y)$.

The following formulas hold true for $L(x)$:

I. For every x if $x < y$, then $L(x) > L(y)$.

II. For every a, x if $L(x) = a$, then for every natural $b < a$, there exists such a y, that $L(y) = b$.

If $< (x, y)$ is an *oriented tree*, then formula III holds true:

III. For every x there is not more than one y, that $x < y$.

Besides if $< (x, y)$ is an *oriented tree*, then there exists only one point a, that $L(a) = 1$ (usually it is called a root), and there is only one path to a from any other point, this path being a maximum path to a from this point in $< (x, y)$.

Let us look at a part of a definitional system in the technology subject field "Pumps", borrowed from (Shelov & Miasnikov, 1987) . Here the term *dynamic pump* is defined through one and only one (generic) term *pump*, the term *friction pump* through one and only one (generic) term *dynamic pump*, the term *peripheral pump* through one and only one (generic) term *friction pump closed-peripheral pump* through one and only one (generic) term *peripheral pump*; thus we have: *pump > dynamic pump > friction pump > peripheral pump > closed-peripheral pump.*

In the direct definability graph $< (x, y)$ there is one only one first conceptual level term *pump*, and there is one and only one path from this term to the terms *dynamic pump, friction pump, peripheral pump, closed-peripheral pump.* So, evaluating conceptual levels of these terms we get the following results: L (*pump*) = 1, L (*dynamic pump*) = 2, L (*friction pump*) = 3, L (*peripheral pump*) = 4, L (*closed-peripheral pump*) = 5.

Let us now consider a small fragment of the term definition system in HEATTR:

1. *Induction hardening. Surface hardening* by induction *heating.*
2. *Surface hardening. Hardening* to a predetermined depth.
3. *Hardening. Heat treatment* designed to render a *material* significantly harder.
4. *Heat treatment.* Application of a combination of *heating, holding* and *quenching* (or cooling, holding and heating) to a solid *material* below its melting point in order to affect the properties of the *material* in the manner desired.

If we confine ourselves by *generic and only generic relations* picked up among all the relations in $< (x, y)$, we will come to the following results: *induction hardening < surface hardening < hardening < heat treatment*, getting fi-

nally L (*heat treatment*) = 1, L (*hardening*) = 2, L (*surface hardening*) = 3, L (*induction hardening*) = 4, that matches perfectly traditional and intuitive evaluation of conceptual level of a term.

But these results can not be warranted as far as the direct definability graph $< (x, y)$ is concerned on the whole, since the following assertions remain true for it: *induction hardening* < *heating, hardening* < *material, heat treatment* < *heating, heat treatment* < *holding, heat treatment* < *quenching, heat treatment* < *material*, and we do not yet know the value of L (*heating*), L (*material*), L (*holding*), L (*quenching*). It seems evident, that as soon as, for example, L (*heating*) = 1, we have for the rest of the terms: L (*heat treatment*) >= 2 and, correspondingly, L (*hardening*) >= 3, L (*surface hardening*) >= 4, L (*induction hardening*) >= 5 and the previous results become wrong.

For this reason let us now consider a non-tree direct definability graph $< (x, y)$ (under the presumption that the condition of no circles and no loops in $< (x, y)$ is still true). Generally speaking, formula III does not hold true in this case. Besides, there might exist more than one point a1, a2, ... that L (a1) = L (A2) = ... = 1, and also there might exist more than one path to any of these points from any other point of $< (x, y)$.

Thus, for the terminology of heat treatment as represented in HEATTR, the following statements hold true:

1) there are several terms of the first conceptual level, among them being *ageing, brittleness, cooling, heating, material, metal, structure.*

2) *lead patenting* < *bath patenting* < *patenting* < *pearlite* < *ferrite* < *alpha iron* < *iron* < *steel* < *metal* and, correspondingly, L (*lead patenting*) = 9, L (*bath patenting*) = 8, L (*patenting*) = 7, L (*pearlite*) = 6, L (*ferrite*) = 5, L (*alpha iron*) = 4, L (*iron*) = 3, L (*steel*) = 2, L (*metal*) = 1.

3) there are several other paths to the term *lead patenting* from the first level terms; for example, we have the following path to this term from the first level term *material*: *lead patenting* < *bath patenting* < *patenting* < *transformation range* < *transformation* < *material*, but all these paths are shorter (or, at least, not longer) than the one mentioned above.

Similarly, for the terminology of personal telecommunication as it is represented in PERCOM, the following statements hold true:

1) there are several terms of the first conceptual level, among them being *authentication, communication, encipherment, network, validation*;

2) *terminal deregistration* < *wireless terminal* < *personal station* < *fixed personal terminal* < *wireless-access mode* < *network* and, correspondingly, L (*terminal deregistration*) = 6, L (*wireless terminal*) = 5, L (*personal station*) = 4, L (*fixed personal terminal*) = 3, L (*wireless-access mode*) = 2, L (*network*) = 1;

3) there are several other paths to the term *terminal deregistration* from the term *network*: for example, we have the following path *deregistration* < *registration* < *network*, but this is shorter than the path mentioned above; there are

also several other paths to the term *terminal deregistration* from the first level terms: for example, we have the following path to this term from the first level term *telecommunication service*: *terminal deregistration* > *registration* > *Personal Communication Service* > *terminal mobility* > *telecommunication service*, but all these paths are shorter (or, at least, not longer) than the one mentioned above.

Finally, though III does not hold true for a non-tree direct definability structure (in other words, for monohierarchical structure), formulas I and II are still correct. Besides, it looks very likely that *the maximum conceptual level in the direct definability graph coincides with what is generally called a graph diameter* (using the standard terminology of the graph theory).

Thus, for the terminology of heat treatment the maximum conceptual level equals 9 since it was demonstrated that L (*lead patenting*) = 9 and L (*lead patenting*) happened to be the maximum conceptual level in the direct definability structure of this subject field; it is equal to a graph diameter in $< (x, y)$ for this subject field since the path from the term *lead patenting* to the term *metal* is the longest path between any pair of terms.

Similarly, for the terminology of personal communications the maximum conceptual level equals 11 since L (*teleservice*) = L (*anchor*) = 11 and L (*teleservice*), L (*anchor*) happened to be the maximum conceptual level in the direct definability structure of this subject field; it is a graph diameter of this subject field since the maximum length of all the paths from the terms *teleservice* and *anchor* to the term *network* equals 11 and the corresponding paths happened to be the longest paths between any pair of terms in $< (x, y)$.

However, as soon as we confine ourselves only to generic relations between terms we immediately come to the traditional generic structure that preserves the traditional intuitive concept of hierarchical level absolutely unchanged. Thus in a small fragment of a term definition system, taken from COMP, we have:

1. *Fixed-radix numeration system.* A *radix numeration system* in which all the *digit* places, except perhaps the one with the highest weight, have the same *radix.*

2. *Numeration system.* Any notation for the representation of numbers.

3. *Positional representation system.* Any *numeration system* in which a real number is represented by an ordered set of *characters* in such a way that the value contributed by a *character* depends upon its position as well as upon its value.

4. *Pure binary numeration system. The fixed-radix numeration system* that uses the *binary digits* and the *radix* 2.

5. *Radix numeration system.* A *positional representation system* in which the ratio of the weight of any one *digit* place to the weight of the *digit* place with the next lower weight is a positive integer.

Analysing this fragment in the same way we get: *pure binary numeration*

system < *fixed-radix numeration system* < *radix numeration system* < *positional representation system* < *numeration system* and, finally, *L* (*pure binary numeration system*) = 5, *L* (*fixed-radix numeration system*) = 4, *L* (*radix numeration system*) = 3, *L* (*positional representation system*) = 2, *L* (*numeration system*) = 1.

So, the idea of conceptual level introduced above seems to be a good generalisation of the concept "conceptual level of a term" for monohierarchical structures applied to the concept analysis of terminology. In a specific field of terminological semantics the following idea (developed for a considerably wider class of linguistic phenomena) turned out to be rather fruitful: "The relation between constructs and their constituents offers a criterion for distinguishing levels in languages" (Hřebíček, 1995:19). Let us hope future investigations will contribute to the assessment of this concept when applied to some wider sphere of term definitions including so-called contextual definitions.

References

Jose, M., & Finatto, B. (1995). Towards the Characterisation of Terminological Definition Paradigms. *Terminology Science & Research,* 6, 3-13.

Hřebíček, L. (1971). *Turkish Grammar as a Graph.* Prague: NČSAV.

ISO. Principles and Methods of Terminology. Geneve: ISO, (ISO/DIS 704). 1984.

Hřebíček, L. (1995). *Text Levels. Language Constructs, Constituents and the Menzerath-Altmann Law.* Trier: Wissenschaftlicher Verlag Trier.

Лотте, Д.С. (1961). *Основы построения научно-технической терминологии. Вопросы теории и методики.* Москва: изд-во АН СССР.

Martin, W. (1992). On the parsing of definitions. *Papers submitted to the 5th EURALEX International Congress. Part I* (pp. 247-256), Tampere.

Meyer, I., Bowker, L., & Eck, K. (1992). COGNITERM: An Experiment in Building a Terminological Knowledge Base. *Papers submitted to the 5th EURALEX International Congress. Part I* (159-172), Tampere.

Rey, A. (1979). *La terminologie: noms et notions.* Paris: Presses univ. de France.

Sager, J.C., & L'Homme, M.C. (1994). A Model for Definition of Concepts: Rules for analytical definitions in terminological databases. *Terminology* V, 1(2), 351-373.

Sager, J.C., & Ndi-Kimbi, A. (1995). The conceptual structure of terminological definitions and their linguistic realisations: A report on research in progress databases. *Terminology* V, 2(1), 61-81.

Shelov, S.D. (1982). One Approach to an Information Thesaurus. *Automatic documentation and mathematical linguistics*, 16, N 4, 10-21.

Shelov, S.D. (1990). Terms, termability and knowledge. *TKE^90: Terminology and Knowledge Engineering,* V.1. Frankfurt/M.

Shelov, S.D. (1996). Concept structure of terminology and knowledge representation procedure. *TKE^96: Terminology and Knowledge Engineering* (pp. 233-237), Frankfurt/M.

Шелов, С. Д. (1998a). Опыт построения терминологической базы знаний в КТН РАН. In A.S. Narin'yani (Ed.), *Computational Linguistics and its Applications: International Workshop, Proceedings,*V, 2, (pp. 726-735), Kazan'.

Шелов, С. Д. (1998b). Построение терминологической базы знаний и анализ понятийной структуры терминологии. *Научно-техническая информация,* Сер. 2, Н 5, 1-16.

Shelov, S.D., & Miasnikov, A.G. (1987). Logical-semantic Structure of a Terminology and its Formal Properties. *Automatic documentation and mathematical linguistics,* V. 21, N. 2, 7-17.

On Word Length, Clause Length and Sentence Length in Bulgarian

Ludmila Uhlířová

0. Introduction

Hřebíček's monographs, papers and lectures are devoted to a thorough investigation of Menzerath-Altmann's law as one of the universal principles of text and language structure and its self-regulatory forces. His recent book (1997), in which he presented a synthesis of his original quantitative theory of text, is a milestone in quantitative linguistics. Below we draw inspiration from this book and try to follow one of Hřebíček's ideas about the relevance of language *type* for modelling language. As Hřebíček points out, Menzerath-Altmann's law is universal and as such it holds irrespective of the typological specificity of languages. However, language type is a factor, or a *boundary* condition, which influences frequencies of language units (language *constituents* and *constructs*, to use Hřebíček's terms) in texts, and, consequently, affects their distribution models. Hřebíček has shown this on a detailed analysis of Turkish poetic texts. He pointed out some features which are easily visible under agglutinative typology (Hřebíček, 1997:80-82).

This paper presents word length, clause length and sentence length data from Bulgarian, a language which is quite complex as a language type. The aim of the paper is to show some specific features which Bulgarian displays as far as the distribution of length units is concerned.[1]

Genetically, Bulgarian belongs to the family of Slavic languages and shares most of their characteristic flexional features. Interestingly, it has preserved some features which have become extinct in some other Slavic languages, e.g. full paradigms of two simple past tenses. On the other hand, Bulgarian has adopted some agglutinative and isolating features typical of the Balkans, and now differs from other Slavic languages in many important points. We ask whether the typological specificity of Bulgarian manifests itself at the level of word, clause and sentence length distributions in texts. The paper is a modest contribution to the typological project on the lengths of language units started several years ago (Best & Altmann, 1996; Wimmer & Altmann, 1996), the results of which were published in volumes 15 and 16 of this series as well as in other places. The paper follows the theoretical and methodological points of departure of the project, as described in detail in the above mentioned volumes.

[1] Supported by a grant from GA ČR No. 102/96/K087

1. Texts

Thirty texts of contemporary intellectual epistolary style were analysed. They are informal, personal, friendly letters written by university teachers and addressed to their colleagues either at the same university or at another university in Bulgaria or abroad, concerning their jobs and common business problems as well as topics from everyday university life. Last but not least, passages about private family affairs and greetings are added in some letters.

The letters are concise, explicit, to the point, expressing clear, rational arguments and offering various alternative solutions to problems under discussion. They are very polite as far as business problems are discussed, and cordial, kind in private passages. Typically, they have rather complex syntax, but still not too elaborate a style. They have fluent text structure without any apparent sharp stylistic changes.

Some letters are hand-written, others are typed, and have the author's original signature. Pre-editing is very probable in letters consisting of several hundred words in which no corrections, deletions, or insertions are visible. Still some other letters are shorter e-mail messages, typed online.

The letters are written by eight different speakers, three men (21 letters) and five women (9 letters). Some of them are pairs of letters and answers. The shortest letter consists of 94 words, the longest of 826 words. The corpus represents a relatively homogeneous set of texts, the properties of which are quantitatively well comparable.

A technical note: In some texts, there are *years* given in numbers. For the purpose of the analysis, they were re-written according to the standard pronunciation rules, e.g., *1989 = chiljada devetstotin osemdeset i devet,* and counted as word forms with the given number of syllables, e.g. 3-4-4-1-2 in this example. (This rule is commonly applied in this sort of analysis.) In case of Bulgarian, such high numbers are relatively long words, and, consequently, they contribute to a high frequency of long words whenever they occur in a text. However, their frequency does not seem to be statistically significant in any of the letters, i.e. it cannot be claimed that the frequency of numerals is decisive for a good or poor agreement of the data with a probability model of the word length distribution in the respective letter.

2. Word length

2.1. Data

The word length is counted in syllables. The segmentation of words into syllables does not cause any problems in the standard pronunciation. Reduction of syllables occurs in colloquial language only, and concerns certain classes of word forms. Reduced forms gradually penetrate into the written standard and some have been recently codified in orthographic dictionaries, e.g. *u-čast-vu-vam* 'take part in' = 4 syllables → *u-čast-vam* = 3 syllables. The empirical data are given in Table 1 for 30 texts.

Table 1

	Word-length frequency								
Text	0	1	2	3	4	5	6	7	8
Rad1	2	30	25	19	11	3	4		
Mumi	5	49	29	22	18	6	1		
Iskra4	6	49	33	25	12	6	2	1	
Adam	4	54	31	23	18	9	5		
Genad1	7	49	35	27	16	8	2		
Iskra2	5	55	41	34	14	6	1		
Marg	2	62	38	37	12	6	1	1	
Iskra1	3	65	37	40	9	7	2		
Juri	5	79	42	26	9	4	1		
Jorn	4	68	44	31	25	6	3	0	1
Iskra5	7	72	43	36	22	5			
Dam1	0	71	52	32	17	11	3		
Kost	8	56	51	55	19	14	4	3	
Sasa1	5	94	73	52	20	9	2		
Sasa2	5	109	60	62	21	8	2		
Boris1	8	112	85	51	11	11	3		
Dam2	14	134	90	58	28	10	9		
Jorn1	13	120	80	64	48	17	7	0	1
Cen1	14	142	75	48	41	26	5	2	1
Jan3	17	154	91	87	35	13	4	1	
Jan1	13	194	122	102	46	17	5	1	
Alb	20	198	145	90	44	17	4		
Cen2	20	186	139	106	45	11	11	1	
Ziv1	22	209	129	91	54	29	9	2	
Jorn2	18	180	121	117	75	26	11	1	
Ziv2	13	204	137	124	37	24	10	4	2
Jan4	14	262	141	151	66	37	12	5	
Jan2	14	302	164	133	67	34	8	1	
Boris2	19	275	189	173	52	32	13	1	
Bacv1	26	297	181	168	90	44	17	2	1

There exist several zero syllable words (prepositions) in Bulgarian; eight-syllable words occur rarely, longer words were not found at all.

2.2. Results

We used the Altmann-Fitter (1994) to test which probability model is the best for the empirical data. For most texts - though not for all - the word length distribution can be successfully modelled by the *extended positive - negative binomial* distribution, where

$$P_x = \begin{cases} 1-\alpha, & x=0 \\ \dfrac{\alpha\binom{k+x-1}{x}p^k q^x}{1-p^k}, & x=1,2,\ldots \end{cases}$$

$k > 0,$

$0 < \alpha < 1,$

$0 < p < 1$

For each text, the theoretical values as well as the values of the χ^2 test at the significance level 0.05 and 0.01 were calculated. For the sake of economy, we do not present the results for each text separately, but give a survey of them in Table 2. The texts which show a good fit either at the significance level 0.05 or 0.01 are marked with the sign "+" in the third (or fourth) column. If there is no fit, there is "N" in the last column of the table.

Table 2

Text	*Number of words*	$\alpha = 0.05$	$\alpha = 0.01$	*No model*
Rad1	94	+		
Mumi	130	+		
Iskra4	134	+		
Adam	144	+		
Genad1	144	+		
Iskra2	156	+		
Marg	159	+		
Iskra1	163		+	
Juri	166	+		
Jorn3	182	+		
Iskra5	185		+	
Dam1	186	+		
Kost	210	+		
Sasa1	255	+		
Sasa2	267		+	
Boris1	281	+		
Dam2	343	+		
Jorn1	350		+	
Cen1	354		(+)	N
Jan3	402	(+)	[+]	
Jan1	500		+	
Alb	518	+		
Cen2	519		+	
Ziv1	545	+		
Jorn2	549		(+)	N
Ziv2	555		+	
Jan4	688			N
Jan2	723		(+)	N
Boris2	754			N
Bacv1	826			N

So far, no other probability distribution has been found for Bulgarian. There is only one exception (text Jan3), where the extended positive Poisson can be fitted to the data at the $P = 0.01$ level; this is marked with [+] in the table above.

Therefore, for the time being, the extended positive negative binomial distribution can be considered as the appropriate word length probability model for Bulgarian.

2.3. Comment

If all texts are arranged according to their increasing length, as has been done in Table 2 above, it can be seen that short texts show a good agreement of the empirical and calculated values at the significance level of 0.05 (sixteen texts: see the symbol "+" in the third column), some of the longer texts show a good

agreement only at the $\alpha = 0.01$ level (another seven texts: see the symbol "+" in the fourth column), one text conforms to the extended positive Poisson distribution (see above), and, finally, we must admit that we did not succeed in finding any model for the longest six texts (see "N" in the last column).

The longer a text, the higher is the danger that the presupposed proportionality between the word classes will abort, and, consequently, the chance decreases that a probability model of word length distribution will be found. It can be observed that in those texts which failed to be modelled, there is a "lack" of two-syllable words, and, at the same time, an "excess" of three-syllable words, sometimes also of four- and/or five-syllable words.

The question arises why this is so. What are the reasons for the superfluous frequency of long words in texts consisting of several hundred words?

Let us have a look at some typological properties of Bulgarian. Like other Slavic languages, Bulgarian is an inflectional language, with many common Slavic features in its lexicon, word-formation, in singular and plural with nouns, in verb morphology (rich in simple and compound forms expressing tenses and modality), etc. On the other hand, unlike other Slavic languages, the Bulgarian noun, adjective and pronoun have lost their case endings. The case functions are expressed by means of prepositions, and in some cases by word order. Neither the case form, nor consequently the word length, changes if a noun or an adjective is used in different clause positions (the difference in length may occur only between singular and plural, cf. *učitel* 'teacher-sg.', *učitel-i* 'teacher-pl.'). In this respect, Bulgarian differs from other Slavic languages, such as Czech or Russian, where case endings may make a word form longer or shorter, the difference between word forms of the same lexeme being one, sometimes even two syllables. Cf. Czech *žen-a* 'woman', nom. sg., 2 syllables [že-na], *žen-0* gen. pl., 1 syllable [žen], *žen-ami*, instr. pl., 3 syllables [že-na-mi]. No such changes in word length occur in Bulgarian - exceptions are rare and more or less frozen (e.g. with personal pronouns, with vocative, etc.). On the other hand, in Bulgarian there is a grammatical category of *definiteness*, which is expressed by the postpositional definite article. The article and its head word form a single word. The article is an agglutinative affix which makes the noun (the adjective, the pronoun) one syllable longer. Cf. Bulg. *učitel* 'teacher', sg., *učitel-jat* 'teacher-the', *učitel-i*, pl., *učitel-i-te* 'teacher-s-the'.

Hence: Whereas in most Slavic languages the word length is determined by the syntactic role of the respective word in a clause (i.e. by its case form), in Bulgarian it is significantly determined by the presence or absence of the definite article as an overt marker of definiteness. The definiteness marker indicates the role of the respective NP in the flow of information in text. The textual role is, in a sense, superimposed upon the syntactic role of a word in the clause. A hypothesis arises as to whether this factor, as a factor which makes definite nouns one syllable longer, may be, at least partly, responsible for the quantitative dis-

proportionality of word length classes and for the failure of the probability model with some texts.

2.4. Experiment

To test the above hypothesis, we counted all nouns, adjectives, and pronouns *with an article* in each of the seven texts that had failed to be modelled by the extended positive negative binomial distribution. The number of occurrences of these words for each text and for each length *i* is given in Table 3 below.

Table 3

Word length in syll.	*Number of words with definite article*						
	Bacv1	Jan2	Jan3	Jorn2	Jan4	Cen1	Boris2
2	2	4	2	0	1	1	6
3	29	21	20	17	19	10	20
4	20	16	7	12	17	8	10
5	15	13	5	11	13	5	10
6	15	3	3	6	7	1	7
7	0	1	0	0	4	2	1
8	0	0	0	0	0	1	0
Σ	81	58	37	46	61	28	54
Text length	826	723	402	549	688	354	754

Then the words *i* syllables long were shifted into the length class *i-1*, i. e. they were counted as if they were used *without* an article. For example, in text Jan2 there occurred 302 one syllable words and 164 two-syllable words (see Table 1 above); 4 two-syllable words are words with the article (see table 3 above), therefore we shift them into the class of one-syllable words, so that the total of one-syllable words now comes to 302 + 4 = 306 now. At the same time 21 three-syllable words with the article (see table 3) are shifted into the class of two-syllable words, so that the total of two-syllable words comes to 164 - 4 + 21 =1 81. Similarly, we get 133 -21 + 16 = 128 for three-syllable words (see again tables 1 and 3 for the data), etc. Having carried out all the necessary shifts for all length classes, we obtain a new, experimentally shifted "empirical" word length distribution. It was then tested whether this modified distribution fits the extended positive negative binomial distribution.

The results show that the "trick" helps in some cases: The texts Cen1, Jan2 and Jan3 now fit the extended positive negative binomial distribution. For these texts, the sign "plus" is given in brackets in the respective columns in Table 2 above.

Although it is well known that for modelling longer texts the contingency coefficient C may be helpful (see Best, 1997, or Niehaus, 1997 for the details), in the case of the four longest texts the values of C, computed with the Altmann-Fitter (1994), did not help. Other factors seem to be at play which make the texts insufficiently homogeneous (cf. below in section 4 and section 5.). And last but not least, there still remains a question, why the presence of the definite article should matter in long texts, whereas it does not matter in short ones. An attempt at another solution to the problem is offered in the next section.

2.5. Local modifications

Let us consider Altmann, Best & Wimmer's proposal (1997) and let us try to find a local modification of the extended positive negative binomial for those texts in which $\alpha < 0.01$. According to the authors we can speak about a local modification if there is a considerable difference between the observed and the corresponding theoretical values in *two* adjacent length classes, rarely between *two* non-adjacent length classes, but only in *two* length classes. If this is so, - and this is the case of five texts from the six which we were not able to model by the extended positive negative binomial - it is possible to shift a proportion β from a class x into the class $x + 1$, or $x - 1$, respectively, whereas all the other classes remain unchanged. Thus we obtain a *modified extended positive negative binomial:*

$$P_x^* = \begin{cases} 1-\alpha, & x=0 \\ P_x(1-\beta), & x=2 \\ P_x+\beta P_{x-1}, & x=3 \\ \dfrac{\alpha\binom{k+x-1}{x}p^k q^x}{1-p^k}, & x=1,4,5,\ldots \end{cases}$$

This method leads to good results in texts Bacv1, Jan2, Boris2, where the value of β, as derived empirically from the data, is 15%.

Still another modification has occurred in texts Cen1 and Jan4, where we try to shift β between two non-adjacent classes:

$$P_x^* = \begin{cases} 1-\alpha, & x=0 \\ P_x(1-\beta), & x=2 \\ P_x+\beta P_{x-2}, & x=4 \\ \dfrac{\alpha\binom{k+x-1}{x}p^k q^x}{1-p^k}, & x=1,3,5,\ldots \end{cases}$$

However, the results are not so persuasive as in the former case, and also the parameter β oscillates too much, in the interval <10%;20%>.

With regard to the two local modifications, we come to a more general conclusion concerning word length distribution in Bulgarian: word length in this language may be modelled by the extended positive negative binomial and/or by the modified extended positive negative binomial; two types of modifications have been found so far.

The existence of a local modification can be interpreted - according to Altmann's synergetic linguistic theory - as a manifestation of a tendency to change the pertinent attractor: The author of the text tries to find a new attractor for his text, i.e. he tends to be original, but he has not yet found it fully - what he does is apply it locally, i.e. for certain length classes only.

3. Clause length

3.1. Rearrangement of the data

Here we start from the presupposition that *clause* is a construct which consists of *words* although we are aware that this presupposition may be criticised as too simple: One may raise an objection that clause constituents are simple or complex syntactic elements which coincide with single word forms only in some cases. However, if we started from the latter standpoint, we would hardly be able to do any counting of clause length at all, because we could not segment texts into non-inclusive syntactic units, as Hřebíček has rightly pointed out (1997:85).

Apart from this, a problem of another kind has to be solved. Some letters are rather short (see Table 2 above for the length of the texts) and we are not sure whether they provide reliable results.

Here we may take advantage of the fact that some texts are written by the same person, and, in some cases, also to the same addressee. The letters are quite similar to one another in many respects, e.g. in subject matter, in degree of spontaneity (or, pre-editing), in social distance between the author and the addressee, style, etc. Some of them look like chapters following one after another in a book. For these reasons, we decided to take all letters written by the *same* author together, as one "open-ended" text (see Hřebíček, 1997 for the term), thus dividing the whole corpus of 30 letters into four parts: we took together all letters written by a person B, then all letters written by a person C, and then all letters written by a person I; the remaining five texts written by five different persons were left aside for the moment. We counted clause length in the "hypertexts" B, C and I.

3.2. Data and results

The *mixed negative binomial* distribution gives a good fit for the three "hypertexts" B, C and I, as shown in Table 4. We may say provisionally, that this is the distribution of clause length in Bulgarian:

$$P_x = \alpha\binom{k_1+x-1}{x}p_1^{k_1}q_1^x + (1-\alpha)\binom{k_2+x-1}{x}p_2^{k_2}q_2^x$$

where $x = 0,1,2,\ldots$, $k_1, k_2 > 0$, $0 < p_1,p_2 < 1$, $0 < \alpha < 1$.

Let us also note that other models of a binomial type have shown quite a good fit, but we shall not discuss these here.

Table 4

	Text B		Text I		Text C	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	14	24,0746	3	5,3035	2	3,3732
2	82	62,5701	17	11,5923	6	10,0733
3	95	98,1626	13	15,8293	15	16,7590
4	115	120,3320	16	17,2868	28	20,5194
5	127	126,9250	15	16,5153	25	20,6568
6	123	120,9202	14	14,4235	17	18,1519
7	103	107,0489	15	11,8079	10	14,4682
8	91	89,7040	12	9,2055	11	10,7601
9	72	72,0819	3	6,9069	8	7,6438
10	53	56,0902	6	5,0250	6	5,2974
11	47	42,6039	4	3,5642	3	3,6511
12	32	31,8063	2	2,4750	1	2,5432
13	22	23,4876	2	1,6880	2	1,8093
14	18	17,2615	1	1,1335	3	1,3194
15	13	12,7000	1	0,7510	0	0,9834
16	9	9,4072	0	0,4918	2	0,7441
17	7	7,0503	0	0,3188	0	0,5672
18	9	5,3672	1	0,2048	0	0,4328
19	4	4,1603	0	0,1306	0	0,3288
20	1	3,2858	0	0,0828	2	0,2480
	$\chi^2_{25} = 20,62$ $P = 0,71$		$\chi^2_{11} = 8,65$ $P = 0,65$		$\chi^2_{11} = 9,38$ $P = 0,59$	

4. Sentence length

Sentence length is measured in number of *clauses*. In some cases the sentence boundary is marked by colon, semicolon or dash. They have, at least in our texts, the same function as the full stop and we take them as sentence boundary signals. The dash is a manifestation of spontaneity in writing.

The empirical data, together with their probability models, as calculated again with the help of Altmann-Fitter (1994), are given in Tables 5 and 6 for "hypertexts" B, I and C respectively. The data are limited in size, and we can offer nothing more than tentative results.

We may approach the task from two different points of view:

1. We can start from what is already known from research (Altmann, 1988; Niehaus, 1997) and test the Bulgarian data against it. Altmann, who presented data from seven languages, as did Niehaus, who dealt with German texts, came to the same result: The best model of sentence length measured in number of clauses is the *negative binomial*. In addition, Niehaus showed that some of her texts fit more than just one model; she mentions positive Poisson, Hyperpoisson, and Conway-Maxwell Poisson among them. Now, we ask whether the negative binomial is also a good model for Bulgarian.

2. We can start from scratch and ask about *a* model of sentence length distribution in Bulgarian; Altmann-Fitter (1994) provides a rich offer of more than two hundred distributions. Then, if more than just one model has been found, ask which is the best.

The results are summarised in Table 5 for B, and in Table 6 for I and C.

The negative binomial distribution gives good results for B and I (see the respective columns in Table 5 and 6), thus supporting Altmann's as well as Niehaus's hypothesis about the negative binomial as a good model of sentence length distribution:

$$P_x = \binom{k+x-1}{x} p^k q^x$$

where $x=0,1,2,\ldots$, $k>0$, $0<p<1$, $q=1-p$.

In addition, letters of group B may be modelled by a number of other models which fit even better than the negative binomial, e.g., positive negative binomial, modified negative binomial, Hyperpoisson, positive shifted Poisson; the best fit is given by the *Hirata-Poisson* distribution, see the theoretical values in Table 5. Also text I may be better modelled than by the negative binomial, the best alternative being the *Thomas* distribution; see the theoretical values in Table 6. Thus, on the one hand, the models reflect what is common (what is universal) to the letters - i.e. the goodness of fit of the negative binomial model. On the other hand, they reflect what is specific, unique in each group. It is worth noticing that both the Hirata-Poisson and the Thomas distributions belong to the group of generalized Poisson distributions and have already been introduced as more complex word length models (see Wimmer & Altmann, 1996). As far as the Hirata-Poisson distribution is concerned, see Dieckmann & Judt (1996), Feldt et al. (1997) and Riedemann (1997) for the details of application. The Thomas distribution has been introduced for the first time here.

Let us add that letters B were written by a man, an experienced stylist; some of his letters were typed, others were written by hand. The style of all letters is very natural, fluent, "readable". Letters I are e-mail messages, written by a woman; their style is much less elaborated, the letters are more concise and less epic.

In contrast to group B and group I, group C cannot be modelled by the negative binomial, and, what is even worse, it is difficult to model at all. True, we found four models, but all with only a poor fit; one of them is the Hirata-Poisson, see the theoretical values in Table 6, which is (similar to the other three distributions) acceptable only at $\alpha = 0.01$, with a low contingency coefficient C. (Let us notice that we succeeded in modelling the word length distribution only for one of them, Cen2, but not for Cen1, see section 2.2 above). Letters C are non-spontaneous. They are two letters, written by a woman who discusses a question (maybe a misunderstanding) concerning a submission of her article for publication. She seems very interested in a positive solution, but very uncertain, even stressed, and her argumentation gives - to some extent - an impression of a blend of logic and confusion.

The difficulty in modelling letters C confirms what has been observed by other authors (Altmann, Best & Wimmer 1997; Best, 1997): It is much easier to model a length distribution if a text was created during a spontaneous, uninterrupted process of writing, and much more difficult to model it in a text, whose "definitive" shape is a result of several cycles of editing and re-editing, during which the original author's intentions may have changed.

Table 5

Text B

	negative binomial		Hirata-Poisson	
x	N_x	NP_x	N_x	NP_x
1	108	101,8063	108	108,5553
2	112	124,0886	112	114,8563
3	97	89,9414	97	89,3669
4	48	50,3790	48	51,6952
5	26	24,0705	26	25,4484
6	9	10,3114	9	10,8340
7	5	4,0776	5	4,1458
8	2	1,5165	2	1,4423
9	0	0,8086	0	0,6559
	$k = 5,2819$ $p = 0,7692$ $\chi^2_5 = 2,80$ $P = 0,73$		$a = 1,3215$ $b = 0,1993$ $\chi^2_5 = 1,49$ $P = 0,91$	

Table 6

	Text I				Text C	
	negative binomial		Thomas		Hirata-Poisson	
x	N_x	NP_x	N_x	NP_x	N_x	NP_x
1	24	21,6016	24	24,5444	24	24,0857
2	11	13,7664	11	11,1124	25	19,7303
3	9	8,1018	9	8,6807	6	12,9924
4	4	4,6364	4	4,8811	11	6,2297
5	3	2,6156	3	2,5397	1	2,6004
6	0	1,4628	0	1,2417	0	0,9341
7	3	0,8133	3	0,5724	0	0,3043
8	0	0,4503	0	0,2513	0	0,0900
9	0	0,5517	0	0,1762	0	0,0330
	$k = 1,1807$ $p = 0,4602$ $\chi^2_4 = 2,30$ $P = 0,68$		$a = 0,7885$ $b = 0,5548$ $\chi^2_3 = 0,52$ $P = 0,91$		$a = 1,023$ $b = 0,1993$ $\chi^2_3 = 11,17$ $P = 0,01$	

5. Bulgarian as a language type

The analysis of thirty texts of contemporary epistolary genre leads to one probability model of *word length* distribution in Bulgarian, namely to the *extended positive negative binomial distribution* together with two modifications. No other model has yet been found (with the exception of one text, as mentioned above). This type of distribution differs, though not substantially, from that which has been found for some other Slavic languages. For example, Czech texts from various genres fit the extended positive binomial distribution (Uhlířová, 1997), and the same distribution holds good also for Polish (Sambor, personal communication) and for Russian (Girzig, 1997). The fact that Bulgarian displays a different, though closely related type of binomial distribution may be interpreted as a manifestation of a typological difference between Bulgarian and other Slavic languages. Nevertheless, the Bulgarian data are in full accordance with the general pattern of word length distributions, as described by Wimmer & Altmann (1996). Bulgarian corresponds well to the word length theory as proposed by the authors; the results represent just one of many applications in support of their theory.

To gain a better insight into the place of Bulgarian among the other Slavic languages, we used the so-called Ord's criterion. Ord proposed characterizing each text from a set of texts by two co-ordinates, I and S, where

$$I = \frac{M_2}{M_1}, \qquad S = \frac{M_3}{M_2},$$

M_1 being the mean and M_2 and M_3 the second and third central moments of the word length distribution. The estimates of M_1, M_2 and M_3 can easily be calculated with the help of the Fitter (Altmann, 1994). The values of M_1, M_2 and M_3, I and S for word-length distribution in Bulgarian texts are given in Table 7. All of them fall within the interval $I \in < 0.6;1.0 >$ and $S \in < 0.6;1.6 >$. Graphically (see Fig. 1), it can be seen that the values make a very compact cluster. Wimmer et al. (in prep.) mention an interval $I \in <0;1 >$ and $S \in <0;2>$ for Slovak and Czech texts of different types (short stories, newspapers, etc.). Bulgarian also falls within this interval, manifesting even greater compactness, which may be due to the stylistic homogeneity of the corpus. However, it is worth noticing that whereas most Slovak and Czech texts have $I < 0.6$, no such text occurs in our Bulgarian corpus. Although this observation has been made on limited data, it may - possibly - again indicate a specific place for Bulgarian in the Slavic language family.

Table 7

Text	*M1*	*M2*	*M3*	*I*	*S*
Rad1	2,3404	1,9267	2,2269	0,8232	1,1558
Mumi	2,1615	1,7662	1,4706	0,8171	0,8326
Iskra4	2,1418	1,8978	2,466	0,8861	1,2994
Adam	2,3056	2,1705	2,454	0,9414	1,1306
Genad1	2,1944	1,8511	1,649	0,8436	0,8908
Iskra2	2,1218	1,4916	1,2288	0,7030	0,8238
Marg	2,1384	1,5532	1,9061	0,7263	1,2272
Iskra1	2,0982	1,5118	1,6397	0,7205	1,0846
Juri	1,8253	1,2526	1,5715	0,6862	1,2546
Jorn3	2,2253	1,8888	2,5637	0,8488	1,3573
Iskra5	2,0486	1,4517	0,9718	0,7086	0,6694
Dam1	2,2151	1,6634	2,0674	0,7509	1,2429
Kost	2,4476	2,1044	2,3035	0,8598	1,0946
Sasa1	2,0902	1,3997	1,3813	0,6696	0,9869
Sasa2	2,0637	1,4154	1,3211	0,6859	0,9334
Boris1	1,9644	1,3724	1,7551	0,6986	1,2789
Dam2	2,0836	1,4532	1,4614	0,6974	1,0056
Jorn1	2,2829	2,0028	2,1264	0,8773	1,0617
Cen1	2,209	2,2444	3,1314	1,0160	1,3952
Jan3	2,0721	1,6241	1,6606	0,7838	1,0225
Jan1	2,1	1,57	1,68	0,7476	1,0701
Alb	2,0135	1,4689	1,4829	0,7295	1,0095
Cen2	2,1002	1,6316	1,9093	0,7769	1,1702
Ziv1	2,145	1,9221	2,3666	0,8961	1,2313
Jorn2	2,3242	1,9058	1,6167	0,8200	0,8483
Ziv2	2,2018	1,898	3,0445	0,8620	1,6041
Jan4	2,2573	1,944	2,4027	0,8612	1,2360
Jan2	2,1051	1,6736	1,9745	0,7950	1,1798
Boris2	2,0836	1,4532	1,4614	0,6974	1,0056
Bacv1	2,2603	1,9698	2,2275	0,8715	1,1308

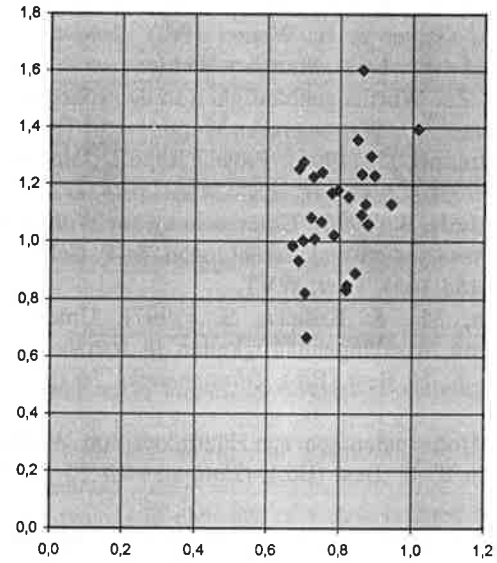


Fig. 1: Ord's criterion

As far as *clause length* (measured in number of words) is concerned, we saw above that the *mixed negative* binomial distribution is a suitable model for Bulgarian. And, finally, *sentence length* (measured in number of clauses) can be modelled by the *negative binomial* distribution, even though with some reservations.

Thus, several distributions of the common, *binomial* type seem to be, for the time being, good probability models of length unit distribution in Bulgarian. Generally, the results have confirmed that Hřebíček is right when stressing the impact of language *type* on language and text modelling.

References

Altmann, G. (1993). Phoneme counts. Marginal remarks to Pääkönen's article. In G. Altmann (Ed.), *Glottometrika 14* (pp. 54-68), Trier: WVT.

Altmann, G. (1988). Verteilungen der Satzlängen. In K.-P Schulz, (Ed.), *Glottometrika 9*, (pp.147-169), Bochum: Brockmeyer.

Altmann, G., Best, K.-H., & Wimmer, G. (1997). Wortlänge in Romanischen Sprachen. In A. Gather & H. Werner (Ed.), *Semiotische Prozesse und natürliche Sprache* (pp. 1-13), Stuttgart: Steiner.

Best, K.-H. (1997). Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten. In K.-H. Best (Ed.), *Glottometrika 16* (pp. 1-15), Trier: WVT.

Best, K.-H., & Altmann, G. (1996). Project Report. *Journal of Quantitative Linguistics,* 3, 85-88.

Dieckmann, S., & Judt, B. (1996). Untersuchung zur Wortlängenverteilung in französischen Pressetexten und Erzählungen. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 158-165), Trier: WVT.

Feldt, S., Janssen, M., & Kuleisa, S. (1997). Untersuchung zur Gesetzmäßigkeit von Wortlängenhäufigkeiten in französischen Briefen und Pressetexten. In K.-H. Best (Ed.), *Glottometrika 16* (pp. 145-151), Trier: WVT.

Girzig, P. (1997). Untersuchungen zur Häufigkeit von Wortlängen in Russischen Texten. In K.-H. Best (Ed.), *Glottometrika 16* (S. 152-162), Trier: WVT.

Hřebíček, L. (1997). *Lectures on Text Theory*. Prague: Oriental Institute.

Riedemann, G. (1997). Wortlängenhäufigkeiten in japanischen Pressetexten. In K.-H. Best (Ed.), *Glottometrika 16* (pp. 180-1184), Trier: WVT.

Uhlířová, L. (1997). Word Length Distribution in Czech: On the Generality of Linguistic Laws and Individuality of Texts. In K.-H. Best (Ed.), *Glottometrika 16* (pp. 163-173), Trier: WVT.

Wimmer, G., & Altmann, G. (1996). The Theory of Word Length: Some Results and Generalizations. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 112-133), Trier: WVT.

Wimmer, G., Wimmerová, S., Hřebíček, L., & Altmann, G., (in prep.). *Úvod do analýzy textů.*

Software

Altmann-FITTER (1994). Lüdenscheid: RAM-Verlag.

Models of Rank-Frequency Distributions in Language and Music

Gejza Wimmer, Gabriel Altmann

In musicology as well as in linguistics discrete probability distributions (p. d.) arise in two ways:

(i) Via a generating mechanism that has developed historically and brings about the genesis of data;

(ii) through the ranking of elements of a class ordered according to their frequency of occurrence. This has a long tradition in linguistics.

The variable in (i) is "natural" (measurable, countable, e.g. probability distribution of the word length, of the semantic productivity of words, etc.), that of (ii) is "artificial" (mathematical) if it is at all necessary to use these terms to differentiate them since both have been constructed conceptually by us. It is evident that in every class of musical or linguistic entities constructed by us, an "artificial" order can be established if one is able to define some pertinent criteria. If, in addition, some lawful interrelations of ranking can be discovered, then they can serve as a criterion of "naturalness" or "correctness" or "closeness to reality". The difference between (i) and (ii) can, however, be founded also genetically and the modeling techniques can be appropriately adapted.

With "natural" variables such as length, strength of voice, complexity, number of repetitions, etc., one can assume that originally only the simplest stage existed (e.g. in the language – monosyllabic words, sentences consisting of one clause, etc.). More complex forms developed because of reduction of redundancy, on the basis of the coding requirement, expression requirement, etc. (cf. Köhler, 1986), but depending on the properties of the simplest classes. Since it is impossible to reconstruct the elementary state of language or music, we restrict ourselves here to dependence in the domain of frequency, which is discrete or may be made discrete, and displays conspicuous regularities. It has already been shown in many problems of quantitative linguistics that frequency in class x is proportional to that in class $x - 1$, or even to those in all lower classes (c.f. Altmann, 1991; & Altmann, 1996; Wimmer & Altmann, 1996; Wimmer, Köhler, Grotjahn, & Altmann, 1994). The complex approach is

(1) $P_x = g(x)\sum_{j=1}^{x} h(j)P_{x-j}$,

where g (.) is a proportionality function, h (.) is a weighting function. Thus the summation concerns the classes 0 to $x - 1$, i. e. the frequency in class x turns out to be proportional to the weighted sum of the frequencies in all lower classes. The simplest form of (1) is

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

and this approach has been successfully used in hundreds of cases (for word length see e.g. Best & Altmann, 1996). This approach already yields an elementary justification or explanation, but if necessary, one can go a step deeper and consider equations (1) or (2) as steady-state solutions of stochastic processes, e.g. of the birth-and-death process. This fact enables us to embed this approach into a more general theory and, at the same time, to use it as an instrument for capturing the variability of languages, texts, compositions, etc.

With ranking, i.e. with variables of type (ii) the argumentation can be precisely the reverse. We have here three points of departure:

(a) the lowest rank is 1 but conventionally it can be set to 0,
(b) the probability distribution is monotone decreasing,
(c) we observe the actual state, not the genesis of the considered set.

From (c) it follows that the frequency at rank 1 (the most frequent element of the set under consideration) depends on the number of the other elements and their frequencies. This fact can easily be illustrated, for example, in the case of the rank-frequency distribution of phonemes of a language. The more phonemes there are in the inventory (investigated set), the more even is the curve, and the smaller is the relative frequency of the phoneme with rank 1. Thus we can assume that P_1 can be considered as a function of the sum of frequencies of other phonemes with rank greater than (or equal to) 1, P_2 as a function of the sum of frequencies with rank greater than (or equal to) 2, etc. However, it is evident that the summed variable, called *parent* and marked as P_j^*, is different. Below we shall show four simple possibilities (schemes).

Scheme I.

$$P_1 = C_1\{P_1^* + P_2^* + P_3^* + P_4^* + \ldots\}$$
$$P_2 = C_1\{P_2^* + P_3^* + P_4^* + \ldots\}$$
$$P_3 = C_1\{P_3^* + P_4^* + \ldots\}$$
$$\vdots$$
$$\cdot$$

As $\sum_{x\geq 1} P_x = 1$, we obtain

$$P_x = \frac{1}{\mu_1^*}\sum_{j\geq x} P_j^*, \quad x = 1,2,\ldots,$$

The recurrence formula for probabilities is ($\mu'_1{}^*$ being the mean of the parent p.d.)

$$P_x = P_{x-1} - \frac{P_{x-1}^*}{\mu_1'^*}, \quad x = 2,3,\ldots.$$

Scheme II.

$$P_1 = C_2\{P_1^* + \frac{P_2^*}{2} + \frac{P_3^*}{3} + \frac{P_4^*}{4} + \ldots\}$$
$$P_2 = C_2\{\frac{P_2^*}{2} + \frac{P_3^*}{3} + \frac{P_4^*}{4} + \ldots\}$$
$$P_3 = C_2\{\frac{P_3^*}{3} + \frac{P_4^*}{4} + \ldots\}$$
$$\vdots$$
$$\cdot$$

Now we obtain

$$P_x = \sum_{j\geq x}\frac{P_j^*}{j}, \qquad x = 1, 2, \ldots,$$

and the recurrence formula is

$$P_x = P_{x-1} - \frac{P_{x-1}^*}{x-1}, \qquad x = 2, 3, \ldots.$$

Scheme III.

$$P_1 = C_3\{P_2^* + P_3^* + P_4^* + P_5^* + \ldots\}$$
$$P_2 = C_3\{P_3^* + P_4^* + P_5^* + \ldots\}$$
$$P_3 = C_3\{P_4^* + P_5^* + \ldots\}$$
$$\vdots$$
$$\cdot$$

For the probabilities we obtain

$$P_x = \frac{1}{\mu_1'^* - 1} \sum_{j \geq x+1} P_j^*, \qquad x = 1, 2, \ldots,$$

and for the recurrence formula

$$P_x = P_{x-1} - \frac{P_x^*}{\mu_1'^* - 1}, \qquad x = 2, 3, \ldots .$$

And finally *Scheme IV.*

$$P_1 = C_4\{P_2^* + \frac{P_3^*}{2} + \frac{P_4^*}{3} + \frac{P_5^*}{4} + \ldots\}$$

$$P_2 = C_4\{\frac{P_3^*}{2} + \frac{P_4^*}{3} + \frac{P_5^*}{4} + \ldots\}$$

$$P_3 = C_4\{\frac{P_4^*}{3} + \frac{P_5^*}{4} + \ldots\}$$

$$\vdots$$

where

$$P_x = \frac{1}{1 - P_1^*} \sum_{j \geq x+1} \frac{P_j^*}{j-1}, \quad x = 1, 2, \ldots,$$

and the recurrence formula is

$$P_x = P_{x-1} - \frac{P_x^*}{(x-1)(1 - P_1^*)}, \qquad x = 2, 3, \ldots .$$

Remarks

(a) If the parent and the resulting partial sums probability distributions begin with 0, the formulas must be slightly modified.

(b) From a particular distribution one can construct several others by partial summation.

We believe that many ranking problems in musicology and linguistics could be captured in this way.

Example (1)

Let the parent probability distribution be a zero-truncated (= positive) Poisson probability distribution with probability mass function (p. m. f.)

$$P_j^* = \frac{e^{-a} a^j}{j!(1 - e^{-a})}, \qquad j = 1, 2, \ldots, a > 0.$$

According to Schemes I – IV we obtain four partial sums probability distributions as follows

(1.I) $$P_x = e^{-a} \sum_{j \geq x} \frac{a^{j-1}}{j!}, \qquad x = 1, 2, \ldots,$$

(1.II) $$P_x = \frac{e^{-a}}{1 - e^{-a}} \sum_{j \geq x} \frac{a^j}{j!\, j}, \qquad x = 1, 2, \ldots,$$

(1.III) $$P_x = \frac{e^{-a}}{e^{-a} + a - 1} \sum_{j \geq x+1} \frac{a^j}{j!}, \qquad x = 1, 2, \ldots,$$

(1.IV) $$P_x = \frac{e^{-a}}{1 - e^{-a} + a e^{-a}} \sum_{j \geq x+1} \frac{a^j}{j!(1-1)}, \qquad x = 1, 2, \ldots .$$

(2) Using the 1-displaced Poisson probability distribution with p. m. f.

$$P_j^* = \frac{e^{-a} a^{j-1}}{(j-1)!}, \qquad j = 1, 2, \ldots, a > 0$$

we obtain only one new probability distribution, namely

(2.I) $$P_x = \frac{e^{-a}}{a+1} \sum_{j \geq x} \frac{a^{j-1}}{(j-1)!}, \qquad x = 1, 2, \ldots .$$

(Type (2.II) and (2.III) are identical with type (1.I); Type (2.IV) is identical with Type (1.II).)

In order to make fitting easier, we introduce some other partial sums probability distribution

(3) From the 1-displaced geometric probability distribution with p. m. f.

$$P_j^* = pq^{j-1}, \quad j=1,2,\dots, \quad 0<p<1, q=1-p$$

we obtain only one new probability distribution

$$\text{(3.II)} \qquad P_x = \frac{p}{q}\sum_{j\geq x}\frac{q^j}{j}, \qquad x=1,2,\dots$$

identical with (3.IV).

(4) From the zero-truncated binomial probability distribution with p. m. f.

$$P_j^* = \binom{n}{j}\frac{p^j q^{n-j}}{1-q^n}, \qquad j=1,2,\dots,n, \quad 0<p<1, q=1-p$$

we obtain

$$\text{(4.I)} \qquad P_x = \frac{1}{np}\sum_{j=x}^{n}\binom{n}{j}p^j q^{n-j}, \qquad x=1,2,\dots,n,$$

$$\text{(4.II)} \qquad P_x = \frac{1}{1-q^n}\sum_{j=x}^{n}\binom{n}{j}\frac{p^j q^{n-j}}{j}, \qquad x=1,2,\dots,n,$$

$$\text{(4.III)} \qquad P_x = \frac{1}{np+q^n-1}\sum_{j=x+1}^{n}\binom{n}{j}p^j q^{n-j}, \qquad x=1,2,\dots,n,$$

$$\text{(4.IV)} \qquad P_x = \frac{1}{1-q^n-npq^{n-1}}\sum_{j=x+1}^{n}\binom{n}{j}\frac{p^j q^{n-j}}{j-1}, \qquad x=1,2,\dots,n.$$

(5) From the zero-truncated negative binomial probability distribution with p. m. f.

$$P_j^* = \binom{k+j-1}{j}\frac{p^k q^j}{1-p^k}, \quad j=1,2,\dots, \quad 0<k, 0<p<1, q=1-p$$

we obtain

$$\text{(5.I)} \qquad P_x = \frac{p^{k+1}}{kq}\sum_{j\geq x}\binom{k+j-1}{j}q^j, \quad x=1,2,\dots,$$

$$\text{(5.II)} \qquad P_x = \frac{p^k}{1-p^k}\sum_{j\geq x}\binom{k+j-1}{j}\frac{q^j}{j}, \quad x=1,2,\dots,$$

$$\text{(5.III)} \qquad P_x = \frac{p^{k+1}}{qk-p+p^{k+1}}\sum_{j\geq x+1}\binom{k+j-1}{j}q^j, \quad x=1,2,\dots,$$

$$\text{(5.IV)} \qquad P_x = \frac{p^k}{1-p^k-kp^k q}\sum_{j\geq x+1}\binom{k+j-1}{j}\frac{q^j}{j-1}, \quad x=1,2,\dots.$$

(6) Finally from the logarithmic probability distribution with p. m. f.

$$P_j^* = \frac{q^j}{-j\log_e(1-q)}, \quad j=1,2,\dots, \quad 0<q<1$$

we obtain

$$\text{(6.I)} \qquad P_x = (1-q)\sum_{j\geq x}\frac{q^{j-1}}{j}, \quad x=1,2,\dots,$$

$$\text{(6.II)} \qquad P_x = \frac{1}{-\log_e(1-q)}\sum_{j\geq x}\frac{q^j}{j^2}, \quad x=1,2,\dots,$$

$$\text{(6.III)} \qquad P_x = \frac{-(1-q)}{q+(1-q)\log_e(1-q)}\sum_{j\geq x+1}\frac{q^j}{j}, \quad x=1,2,\dots,$$

$$\text{(6.IV)} \qquad P_x = \frac{-1}{q+\log_e(1-q)}\sum_{j\geq x+1}\frac{q^j}{(j-1)j}, \quad x=1,2,\dots.$$

Here are some examples from linguistics:

1. The rank-order distribution of the suffix *–e* in modern German according to its meaning (Rothe, 1990:112):

probability distribution Type (1.III)

rank	*frequency of occurrence*	*fitted value*
1	13	11.65
2	11	8.66
3	6	6.12
4	5	4.40
5	5	3.43
6	3	2.97
7	2	2.77
8	2	2.69
9	2	2.67
10	2	2.66
11	2	2.66
12	1	2.66
13	1	2.66
14	1	0.00

$\hat{a} = 3.38840$ $P = 0.98$

2. The rank-order distribution of word classes in the FAZ corpus (FAZ = Frankfurter Allgemeine Zeitung) (Becker, 1995:228):

probability distribution Type (1.IV)

rank	frequency of occurrence	fitted value
1	104	80.91
2	56	69.24
3	53	56.54
4	41	44.08
5	34	33.22
6	24	24.77
7	15	18.85
8	14	15.09
9	1	0.72

$\hat{a} = 6.5372$ $P = 0.11$

3. The rank-order distribution of reflexives in German according to their meaning (Rothe, 1990:113):

probability distribution Type (3.II)

rank	*frequency of occurrence*	*fitted value*
1	51	44.45
2	19	24.39
3	11	15.89
4	10	11.10
5	9	8.05
6	8	5.99
7	7	4.53
8	5	3.48
9	3	2.69
10	3	2.10
11	2	1.65
12	2	1.31
13	1	0.37

$\hat{q} = 0.8469$ $P = 0.44$

A very good model for rank-order distribution in music seems to be the 1-displaced negative hypergeometric probability distribution with p. m. f.

$$P_x = \frac{\binom{M+x}{x-1}\binom{K-M+n-x}{n-x+1}}{\binom{K+n-1}{n}}, \qquad x = 1, 2, \ldots,$$

$K > M > 0, n \in \{0, 1, 2, \ldots\}$.

We can illustrate this model on Ludwig van Beethoven's Sonata Op. 27, No.2 (under investigation is the ranked frequency of occurrence of tones):

rank	*frequency of occurrence*	*fitted value*	*rank*	*frequency of occurrence*	*fitted value*
1	106	113.17	29	11	11.07
2	89	85.34	30	11	10.22
3	84	73.19	31	10	9.42
4	79	65.26	32	10	8.66
5	68	59.28	33	5	7.95
6	66	54.43	34	5	7.27
7	58	50.31	35	5	6.64
8	50	46.72	36	5	6.04
9	44	43.52	37	4	5.47
10	42	40.63	38	4	4.94
11	42	37.99	39	3	4.45
12	34	35.56	40	3	3.98
13	33	33.32	41	3	3.55
14	28	31.23	42	3	3.15
15	25	29.27	43	2	2.77
16	21	27.44	44	2	2.43
17	20	25.72	45	2	2.11
18	19	24.09	46	2	1.82
19	18	22.56	47	1	1.55
20	17	21.11	48	1	1.31
21	17	19.74	49	1	1.09
22	16	18.44	50	1	0.90
23	14	17.21	51	1	0.73
24	14	16.05	52	1	0.57
25	14	14.94	53	1	0.44
26	14	13.90	54	1	0.33
27	13	12.90	55	1	0.86
28	11	11.96			

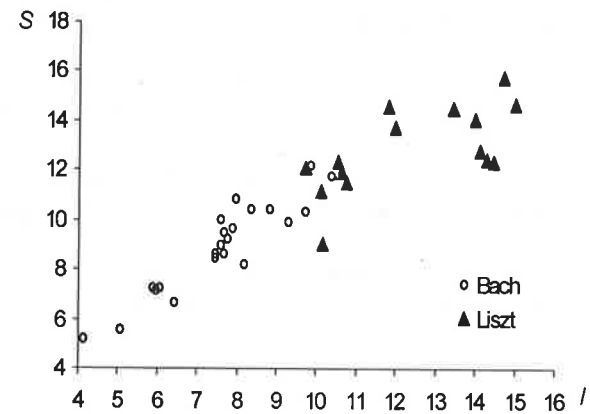
where $\hat{K} = 4.0964$ $\hat{M} = 0.7836$ $\hat{n} = 59$ $P = 0.994$.

Using the $< I,S >$ characterisation (suggested by Ord) where

$$I = \frac{\mu_2}{\mu_1'},$$

$$S = \frac{\mu_3}{\mu_2},$$

of the best fitted 1-displaced negative hypergeometric probability distribution to the measured data (ranked frequency of occurrence of tones) we obtain the next figure for some compositions of Johann Sebastian Bach and Franz Liszt.



Evidently Bach and Liszt can be fairly well discriminated.

We hope that this method of investigation will offer the possibility of discovering law-like hypotheses in linguistics as well as in musicology or other sciences.

Note

Supported by VEGA, the Grant Agency of the Slovak Republic, grant No. 1/4196/97, grant No. 2/5126/98 and grant No. 1/3171/96.

References

Altmann, G. (1991). Modelling diversification phenomena in language. In U. Rothe (Ed.), *Diversification Processes in Language: Grammar*, Hagen: Rottmann.

Best, K.-H., & Altmann, G. (1996). Project report. *Journal of Quantitative Linguistics*, 3, 85-88.

Becker, H. (1995). *Die Wirtschaft in der deutschsprachingen Presse.* Frankfurt: Lang.

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

Köhler, R., & Altmann, G. (1996). "Language Forces" and synergetic modelling of language phenomena. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 62-67), Trier: WVT.

Rothe, U. (1990). Verteilung der Suffixe denominaler Verben nach ihren semantischen Wortbildungsmustern. In P. Schmidt (Ed.), *Glottometrika 15* (pp. 107-114), Trier: WVT.

Wimmer, G., & Altmann, G. (1996*). The theory of word length: Some results and generalizations.* In P. Schmidt (Ed.), *Glottometrika 15* (pp. 112-133), Trier: WVT.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, 1, 98-106.

Word Class Frequencies in Portuguese Press Texts

Arne Ziegler

1. Introduction

This paper presents an approach to analysing the empirical frequency of word classes in Portuguese press texts and asks whether their distributions follow one or even more probabilistic models. The hypothesis that word class distributions in texts in Portuguese are not chaotic, but abide by specific laws is already corroborated in a preceding examination of Brazilian-Portuguese press texts (Ziegler, 1998b). Thus, this analysis is the second one that deals with a Portuguese variety and enables us to get a more exact view of the phenomenon of word class frequency distributions in Romance languages and in particular in the Portuguese language.

Starting from the background hypothesis that word classes are the result of a (historical) diversification process of coding means (Köhler, 1991), it might be supposed that the frequencies of individual classes are not uniformly distributed, but abide by a special 'ranking' law whose different forms have been shown in Altmann (1991) and tested on different language phenomena (Hammerl, 1990; Schweers & Zhu, 1991; Rothe, 1991; Zhu & Best, 1992; cf. also Arapov & Šrejder, 1977). Since we use here a rather exploratory approach, we restrict our attention merely to two probability distributions and leave it to further research to decide which of them can find a better theoretical foundation.

2. Word classes in Portuguese and data aquisition

For the Portuguese language nine different word classes are differentiated, namely: noun, adjective, verb, adverb, article, pronoun, preposition, conjunction and numeral (Ministèrio da Educação e Cultura 1979; Cunha & Cintra, 1984; Caetano, Mayr, Plachy & Ptacek, 1992). However, the attribution of words to different classes is not always absolutely certain, because Portuguese grammar defines its word classes inconsistently, taking into account both mor-

phological (Portuguese grammar distinguishes between inflected and uninflected word classes) as well as semantic and syntactic criteria.

As in preceding examinations, in this analysis a "word" is defined orthographically, i.e. a sequence of letters that is not interrupted by blanks or punctuation marks (Bünting & Bergenholtz, 1989:36; Best, 1994:145; Ziegler, 1998a). Nevertheless, with regard to Portuguese, this definition has to be modified. Word groups and compounds, as for instance, adverbial constructions (e.g. *as vezes* = 'sometimes'), reflexive verbs (e.g. *lembremo-nos* = 'we remember') or periphrastic conjugations (e.g. *O Paulo está a dormir/O Paulo está dormindo* = 'Paul is sleeping') are counted by classifying each single word. Contractions like *comigo* ('with me') or *dele* ('of him') are always counted as one word, whereby the problem of deciding on just one single word class appears. Since it is mostly prepositions functioning as affixes which form the prefixes of these words thus giving them a prepositional character, they are counted as prepositions. A further difficulty in determining word classes is the frequently occurring conversions (e.g. numerals very often change the word class). Although an attempt has been made to tackle this problem, classification - according to the nature of each kind of categorisation - is not absolutely certain in all cases.

Abbreviations and numbers are expanded and counted separately and assigned to the appropriate word class (e.g. *ONU/Organização das Nações Unidas* = 'UNO/United Nations Organisation' is counted as 4 words; *21/vinte e um* = 'twenty-one' is counted as 3 words).

Basically all titles, subtitles, recommendations etc. are not counted, but only the body text.

3. The study

For this analysis and the subsequent presentation of the results:

X^2	= value of the empirical chi-square
P	= probability of the given or greater X^2 value
df	= degrees of freedom
a, b, a, K, M, n	= parameters
X	= rank of the word class
$E[X]$	= empirical frequency
$NP[X]$	= theoretical frequency

The first step in this analysis is to examine whether the data from the 20 Portuguese press texts are homogeneous or not. Therefore, homogeneity is checked regarding its qualitative aspect, i.e. homogeneity of the texts is examined without considering the ranks of the observed frequencies, but the identical word classes, and regarding its quantitative aspect, i.e. homogeneity of the texts is examined

considering the ranks of word classes, that is by comparing rank-identical classes. Since the qualitative homogeneity with $X^2 = 427.62$ with 152 *df* and the quantitative homogeneity with $X^2 = 270.24$ with also 152 *df* is not given, a great variety of suitable distribution models might be expected to apply. However, merely two models seemed to be adequate to fit the empirical data of word classes. Thus, a preceding examination of Brazilian-Portuguese texts can be preliminarily corroborated, and the data are examined to see whether they are consistent with the model of the 1-displaced mixed Poisson distribution that is given by the formula

$$P_x = \frac{\alpha a^{x-1} e^{-a}}{(x-1)!} + \frac{(1-\alpha) b^{x-1} e^{-b}}{(x-1)!}, \quad x = 1,2,....$$

$a, b > 0$, $0 \leq \alpha \leq 1$, and whether they are consistent with the 1-displaced negative hypergeometric distribution defined as

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

$K > M$, $n \in N$. Because the sample size (N) is in all cases large, the coefficient $C = X^2/N$ as a discrepancy coefficient is also applied (Grotjahn & Altmann, 1993: 143).

The results are basically considered as satisfactory if $P \geq 0.05$, and are still acceptable if $0.01 \leq P < 0.05$. If C is used, the results are considered as satisfactory if $C \leq 0.01$, and are still acceptable if $0.01 < C \leq 0.02$.

4. The results

The examination of the data from the 20 Portuguese press texts according to (the model of) the 1-displaced mixed Poisson distribution shows the following results:

Text 1 / Text 2

X	*word class*	*E[X]*	*NP[X]*	*word class*	*E[X]*	*NP[X]*
1	noun	96	92.17	noun	86	83.80
2	preposition	67	69.91	preposition	57	57.72
3	verb	47	43.83	article	37	34.47
4	adjective	45	40.62	numeral	31	32.74
5	pronoun	34	41.27	adjective	30	34.04
6	adverb	30	36.30	verb	28	30.34
7	article	28	26.93	conjunction	26	22.76
8	numeral	20	17.16	adverb	18	14.65
9	conjunction	19	17.81	pronoun	13	15.48
∑		386	386.00		326	326.00
	$a = 4.4615$	$P = 0.5560$		$a = 4.5069$	$P = 0.7584$	
	$b = 0.6571$	$C = 0.0102$		$b = 0.5977$	$C = 0.0080$	
	$\alpha = 0.5516$			$\alpha = 0.5436$		
	$X^2 = 3.955$			$X^2 = 2.620$		
	$df = 5$			$df = 5$		

Text 3 / Text 4

X	*word class*	*E[X]*	*NP[X]*	*word class*	*E[X]*	*NP[X]*
1	noun	100	97.43	noun	119	117.63
2	preposition	67	69.99	preposition	73	71.42
3	verb	45	40.87	verb	41	41.68
4	pronoun	34	35.29	article	37	37.75
5	adjective	30	33.85	pronoun	30	34.69
6	article	27	28.24	adjective	28	26.81
7	adverb	21	19.89	adverb	24	17.38
8	numeral	14	12.03	conjunction	7	9.67
9	conjunction	11	11.41	numeral	6	7.97
∑		349	349.00		365	365.00
	$a = 4.2356$	$P = 0.9073$		$a = 3.8943$	$P = 0.4805$	
	$b = 0.6266$	$C = 0.0044$		$b = 0.5037$	$C = 0.0123$	
	$\alpha = 0.4909$			$\alpha = 0.4829$		
	$X^2 = 1.549$			$X^2 = 4.495$		
	$df = 5$			$df = 5$		

Text 5 / Text 6

X	*word class*	*E[X]*	*NP[X]*	*word class*	*E[X]*	*NP[X]*
1	noun	115	109.65	noun	126	124.97
2	preposition	83	85.42	preposition	101	96.94
3	verb	54	50.29	verb	54	56.30
4	numeral	48	44.60	adjective	47	46.68
5	adjective	39	47.69	article	45	45.62
6	conjunction	38	45.43	pronoun	36	39.80
7	pronoun	36	36.68	adverb	26	29.48
8	article	33	25.44	conjunction	24	18.77
9	adverb	30	30.80	numeral	19	19.44
∑		476	476.00		478	478.00
	$a = 4.8581$	$P = 0.3123$		$a = 4.4598$	$P = 0.7733$	
	$b = 0.7029$	$C = 0.0125$		$b = 0.6945$	$C = 0.0053$	
	$\alpha = 0.5432$			$\alpha = 0.4877$		
	$X^2 = 5.983$			$X^2 = 2.521$		
	$df = 5$			$df = 5$		

Text 7 / Text 8

X	*word class*	*E[X]*	*NP[X]*	*word class*	*E[X]*	*NP[X]*
1	noun	92	88.98	noun	135	131.94
2	preposition	74	73.69	preposition	93	94.29
3	adjective	52	50.48	verb	56	53.82
4	verb	48	46.39	article	45	45.31
5	pronoun	37	44.35	adjective	44	42.97
6	article	33	36.51	pronoun	30	35.61
7	adverb	28	25.38	adverb	21	24.95
8	conjunction	23	15.16	conjunction	18	15.01
9	numeral	8	14.06	numeral	16	14.10
∑		395	395.00		458	458.00
	$a = 4.1821$	$P = 0.1227$		$a = 4.2133$	$P = 0.7653$	
	$b = 0.6943$	$C = 0.0220$		$b = 0.6274$	$C = 0.0056$	
	$\alpha = 0.5662$			$\alpha = 0.4736$		
	$X^2 = 8.675$			$X^2 = 2.574$		
	$df = 5$			$df = 5$		

Text 9

X	word class	E[X]	NP[X]
1	noun	131	123.45
2	preposition	99	103.79
3	verb	70	61.12
4	numeral	49	51.37
5	article	47	55.33
6	pronoun	46	54.86
7	conjunction	46	46.42
8	adjective	45	33.79
9	adverb	42	44.87
∑		575	575.00

$a = 5.0996$ $P = 0.1233$
$b = 0.7742$ $C = 0.0151$
$a = 0.5415$
$X^2 = 8.662$
$df = 5$

Text 10

word class	E[X]	NP[X]
noun	134	131.79
preposition	79	81.78
verb	63	57.40
article	53	54.89
adjective	41	47.51
pronoun	38	33.82
adverb	18	20.13
conjunction	10	10.28
numeral	9	7.40
	445	445.00

$a = 3.5741$ $P = 0.7387$
$b = 0.4560$ $C = 0.0062$
$a = 0.5574$
$X^2 = 2.748$
$df = 5$

Text 11

X	word class	E[X]	NP[X]
1	noun	164	161.33
2	preposition	107	107.02
3	adjective	61	58.71
4	verb	45	48.78
5	article	43	44.95
6	pronoun	36	35.87
7	adverb	23	24.14
8	conjunction	20	13.95
9	numeral	8	12.25
∑		507	507.00

$a = 4.0467$ $P = 0.4614$
$b = 0.5783$ $C = 0.0092$
$a = 0.4465$
$X^2 = 4.640$
$df = 5$

Text 12

word class	E[X]	NP[X]
noun	178	174.23
preposition	101	103.31
adjective	68	62.85
verb	61	63.69
article	61	64.94
conjunction	52	55.27
pronoun	39	39.40
numeral	26	24.08
adverb	25	23.23
	611	611.00

$a = 4.2798$ $P = 0.9245$
$b = 0.4926$ $C = 0.0023$
$a = 0.5457$
$X^2 = 1.398$
$df = 5$

Text 13

X	word class	E[X]	NP[X]
1	noun	140	138.28
2	preposition	91	88.49
3	adjective	46	48.01
4	verb	45	42.31
5	article	37	41.92
6	pronoun	34	35.80
7	adverb	24	25.74
8	conjunction	22	15.88
9	numeral	13	15.57
∑		452	452.00

a = $P = 0.5629$
b = $C = 0.0086$
α =
$X^2 = 3.907$
$df = 5$

Text 14

word class	E[X]	NP[X]
noun	116	114.00
preposition	70	71.63
verb	38	35.61
adjective	27	28.04
article	22	26.08
pronoun	22	21.29
adverb	16	14.68
conjunction	11	8.69
numeral	6	7.98
	328	328.00

$a = 4.1463$ $P = 0.8306$
$b = 0.5623$ $C = 0.0065$
$\alpha = 0.4008$
$X^2 = 2.132$
$df = 5$

Text 15

X	word class	E[X]	NP[X]
1	noun	163	162.08
2	preposition	100	95.27
3	adjective	39	42.81
4	verb	35	33.64
5	article	35	34.17
6	pronoun	25	30.78
7	conjunction	22	23.41
8	numeral	18	15.28
9	adverb	17	16.56
∑		454	454.00

$a = 4.5719$ $P = 0.8034$
$b = 0.5420$ $C = 0.0051$
$\alpha = 0.3932$
$X^2 = 2.320$
$df = 5$

Text 16

word class	E[X]	NP[X]
noun	111	107.61
verb	66	68.62
preposition	65	56.22
pronoun	52	59.40
adjective	44	54.06
article	43	40.07
adverb	32	24.80
conjunction	19	13.16
numeral	2	10.06
	434	434.00

$a = 3.7146$ $P = 0.1423$
$b = 0.4302$ $C = 0.0159$
$\alpha = 0.6428$
$X^2 = 6.882$
$df = 4$

Text 17

X	word class	E[X]	NP[X]
1	noun	108	106.01
2	preposition	65	65.26
3	pronoun	40	38.96
4	verb	39	37.77
5	article	34	37.61
6	adjective	27	31.48
7	adverb	24	22.09
8	conjunction	17	13.30
9	numeral	11	12.52
∑		365	365.00
	$a = 4.2152$	$P = 0.7831$	
	$b = 0.5167$	$C = 0.0067$	
	$a = 0.5261$		
	$X^2 = 2.456$		
	$df = 5$		

Text 18

X	word class	E[X]	NP[X]
1	noun	145	142.03
2	preposition	53	55.93
3	article	33	27.92
4	adjective	29	28.77
5	verb	24	28.88
6	pronoun	21	23.77
7	conjunction	15	16.34
8	adverb	12	9.63
9	numeral	10	8.73
∑		342	342.00
	$a = 4.1249$	$P = 0.6730$	
	$b = 0.3300$	$C = 0.0093$	
	$a = 0.4321$		
	$X^2 = 3.175$		
	$df = 5$		

Text 19

X	word class	E[X]	NP[X]
1	noun	146	144.53
2	preposition	84	84.89
3	verb	53	50.72
4	pronoun	45	47.57
5	adjective	43	44.18
6	article	34	34.25
7	adverb	22	22.24
8	conjunction	14	12.39
9	numeral	10	10.23
∑		451	451.00
	$a = 3.8997$	$P = 0.9915$	
	$b = 0.4796$	$C = 0.0011$	
	$a = 0.4986$		
	$X^2 = 0.516$		
	$df = 5$		

Text 20

X	word class	E[X]	NP[X]
1	noun	181	174.63
2	preposition	79	85.24
3	article	49	40.30
4	verb	37	38.70
5	adjective	34	41.28
6	pronoun	32	37.05
7	numeral	28	27.85
8	conjunction	26	17.95
9	adverb	16	19.00
∑		482	482.00
	$a = 4.5124$	$P = 0.1223$	
	$b = 0.4327$	$C = 0.0180$	
	$a = 0.4492$		
	$X^2 = 8.686$		
	$df = 5$		

The examination of the 20 Portuguese press texts by the model of the 1-displaced negative hypergeometric distribution shows the following results:

Text 1

X	word class	E[X]	NP[X]
1	noun	96	98.71
2	preposition	67	62.24
3	verb	47	49.44
4	adjective	45	41.76
5	pronoun	34	36.10
6	adverb	30	31.38
7	article	28	27.03
8	numeral	20	22.51
9	conjunction	19	16.83
∑		386	386.00
	$K = 1.9344$	$P = 0.9017$	
	$M = 0.6527$	$C = 0.0041$	
	$n = 8.0000$		
	$X^2 = 1.596$		
	$df = 5$		

Text 2

X	word class	E[X]	NP[X]
1	noun	86	85.90
2	preposition	57	52.52
3	article	37	41.30
4	numeral	31	34.71
5	adjective	30	29.95
6	verb	28	26.04
7	conjunction	26	22.48
8	adverb	18	18.83
9	pronoun	13	14.27
∑		326	326.00
	$K = 1.8927$	$P = 0.8392$	
	$M = 0.6314$	$C = 0.0064$	
	$n = 8.0000$		
	$X^2 = 2.071$		
	$df = 5$		

Text 3

X	word class	E[X]	NP[X]
1	noun	100	102.51
2	preposition	67	60.76
3	verb	45	46.22
4	pronoun	34	37.40
5	adjective	30	30.86
6	article	27	25.41
7	adverb	21	20.44
8	numeral	14	15.49
9	conjunction	11	9.91
∑		349	349.00
	$K = 2.1286$	$P = 0.9178$	
	$M = 0.6297$	$C = 0.0042$	
	$n = 8.0000$		
	$X^2 = 1.459$		
	$df = 5$		

Text 4

X	word class	E[X]	NP[X]
1	noun	119	118.71
2	preposition	73	67.61
3	verb	41	49.56
4	article	37	38.52
5	pronoun	30	30.35
6	adjective	28	23.64
7	adverb	24	17.73
8	conjunction	7	12.20
9	numeral	6	6.68
∑		365	365.00
	$K = 2.3711$	$P = 0.2006$	
	$M = 0.6228$	$C = 0.0199$	
	$n = 8.0000$		
	$X^2 = 7.281$		
	$df = 5$		

Text 5

X	word class	E[X]	NP[X]
1	noun	115	119.96
2	preposition	83	71.69
3	verb	54	56.86
4	numeral	48	48.72
5	adjective	39	43.23
6	conjunction	38	39.06
7	pronoun	36	35.57
8	article	33	32.32
9	adverb	30	28.59
∑		476	476.00

$K = 1.6795$ $P = 0.7496$
$M = 0.6034$ $C = 0.0056$
$n = 8.0000$
$X^2 = 2.677$
$df = 5$

Text 6

word class	E[X]	NP[X]
noun	126	133.88
preposition	101	81.13
verb	54	62.68
adjective	47	51.52
article	45	43.23
pronoun	36	36.30
adverb	26	29.94
conjunction	24	23.47
numeral	19	15.85
	478	478.00

$K = 2.0561$ $P = 0.1465$
$M = 0.6377$ $C = 0.0171$
$n = 8.0000$
$X^2 = 8.182$
$df = 5$

Text 7

X	word class	E[X]	NP[X]
1	noun	92	92.90
2	preposition	74	68.78
3	adjective	52	56.60
4	verb	48	47.60
5	pronoun	37	39.94
6	article	33	32.91
7	adverb	28	26.06
8	conjunction	23	19.02
9	numeral	8	11.19
∑		395	395.00

$K = 2.4642$ $P = 0.7203$
$M = 0.8017$ $C = 0.0073$
$n = 8.0000$
$X^2 = 2.868$
$df = 5$

Text 8

word class	E[X]	NP[X]
noun	135	141.42
preposition	93	80.35
verb	56	60.06
article	45	48.04
adjective	44	39.27
pronoun	30	32.04
adverb	21	25.55
conjunction	18	19.19
numeral	16	12.09
	458	458.00

$K = 2.1180$ $P = 0.3439$
$M = 0.6046$ $C = 0.0123$
$n = 8.0000$
$X^2 = 5.630$
$df = 5$

Text 9

X	word class	E[X]	NP[X]
1	noun	131	138.47
2	preposition	99	83.43
3	verb	70	66.86
4	numeral	49	58.01
5	article	47	52.26
6	pronoun	46	48.13
7	conjunction	46	44.96
8	adjective	45	42.44
9	adverb	42	40.44
∑		575	575.00

$K = 1.6009$ $P = 0.3342$
$M = 0.6024$ $C = 0.0100$
$n = 8.0000$
$X^2 = 5.722$
$df = 5$

Text 10

word class	E[X]	NP[X]
noun	134	131.62
preposition	79	84.68
verb	63	64.26
article	53	50.46
adjective	41	39.60
pronoun	38	30.37
adverb	18	22.14
conjunction	10	14.54
numeral	9	7.33
	445	445.00

$K = 2.6385$ $P = 0.4000$
$M = 0.7174$ $C = 0.0115$
$n = 8.0000$
$X^2 = 5.132$
$df = 5$

Text 11

X	word class	E[X]	NP[X]
1	noun	164	167.16
2	preposition	107	93.65
3	adjective	61	68.31
4	verb	45	52.99
5	article	43	41.72
6	pronoun	36	32.52
7	adverb	23	24.44
8	conjunction	20	16.89
9	numeral	8	9.32
∑		507	507.00

$K = 2.3445$ $P = 0.3931$
$M = 0.6116$ $C = 0.0102$
$n = 8.0000$
$X^2 = 5.191$
$df = 5$

Text 12

word class	E[X]	NP[X]
noun	178	175.68
preposition	101	100.34
adjective	68	76.71
verb	61	63.25
article	61	53.72
conjunction	52	46.04
pronoun	39	39.17
numeral	26	32.26
adverb	25	23.83
	611	611.00

$K = 1.8777$ $P = 0.5293$
$M = 0.5916$ $C = 0.0068$
$n = 8.0000$
$X^2 = 4.141$
$df = 5$

Text 13

X	word class	E[X]	NP[X]
1	noun	140	142.85
2	preposition	91	77.55
3	adjective	46	57.49
4	verb	45	46.06
5	article	37	37.94
6	pronoun	34	31.40
7	adverb	24	25.58
8	conjunction	22	19.85
9	numeral	13	13.28
∑		452	452.00

$K = 1.9903$ $P = 0.3825$
$M = 0.5713$ $C = 0.0117$
$n = 8.0000$
$X^2 = 5.281$
$df = 5$

Text 14

word class	E[X]	NP[X]
noun	116	119.09
preposition	70	59.41
verb	38	41.83
adjective	27	31.94
article	22	25.00
pronoun	22	19.52
adverb	16	14.80
conjunction	11	10.43
numeral	6	5.98
	328	328.00

$K = 2.1914$ $P = 0.5656$
$M = 0.5395$ $C = 0.0119$
$n = 8.0000$
$X^2 = 3.888$
$df = 5$

Text 15

X	word class	E[X]	NP[X]
1	noun	163	171.05
2	preposition	100	74.95
3	adjective	39	52.28
4	verb	35	40.75
5	article	35	33.22
6	pronoun	25	27.54
7	conjunction	22	22.77
8	numeral	18	18.27
9	adverb	17	13.17
∑		454	454.00

$K = 1.7525$ $P = 0.0130$
$M = 0.4545$ $C = 0.0318$
$n = 8.0000$
$X^2 = 14.443$
$df = 5$

Text 16

word class	E[X]	NP[X]
noun	111	103.82
verb	66	78.67
preposition	65	64.50
pronoun	52	53.48
adjective	44	43.89
article	43	35.05
adverb	32	26.60
conjunction	19	18.26
numeral	2	9.73
	434	434.00

$K = 2.6839$ $P = 0.0405$
$M = 0.8379$ $C = 0.0268$
$n = 8.0000$
$X^2 = 11.615$
$df = 5$

Text 17

X	word class	E[X]	NP[X]
1	noun	108	107.24
2	preposition	65	61.92
3	pronoun	40	47.06
4	verb	39	38.35
5	article	34	32.03
6	adjective	27	26.85
7	adverb	24	22.16
8	conjunction	17	17.46
9	numeral	11	11.93
∑		365	365.00

$K = 2.0016$ $P = 0.9035$
$M = 0.6060$ $C = 0.0043$
$n = 8.0000$
$X^2 = 1.581$
$df = 5$

Text 18

word class	E[X]	NP[X]
noun	145	143.89
preposition	53	53.78
article	33	36.14
adjective	29	27.77
verb	24	22.57
pronoun	21	18.83
conjunction	15	15.82
adverb	12	13.10
numeral	10	10.10
	342	342.00

$K = 1.5863$ $P = 0.9756$
$M = 0.3832$ $C = 0.0024$
$n = 8.0000$
$X^2 = 0.823$
$df = 5$

Text 19

X	word class	E[X]	NP[X]
1	noun	146	144.97
2	preposition	84	81.52
3	verb	53	60.06
4	pronoun	45	47.20
5	adjective	43	37.77
6	article	34	30.03
7	adverb	22	23.16
8	conjunction	14	16.60
9	numeral	10	9.69
∑		451	451.00

$K = 2.2414$ $P = 0.7399$
$M = 0.6069$ $C = 0.0061$
$n = 8.0000$
$X^2 = 2.741$
$df = 5$

Text 20

word class	E[X]	NP[X]
noun	181	180.49
preposition	79	74.87
article	49	52.38
verb	37	41.53
adjective	34	34.77
pronoun	32	29.95
numeral	28	26.13
conjunction	26	22.76
adverb	16	19.12
	482	482.00

$K = 1.5263$ $P = 0.8219$
$M = 0.4204$ $C = 0.0045$
$n = 8.0000$
$X^2 = 2.193$
$df = 5$

5. Conclusions

In order to provide a visual representation of heterogeneity we use Ord's criterion (Ord, 1972) and define

$$I = \frac{m_2}{m_1'}, \quad S = \frac{m_3}{m_2} \text{ where } m_1' = \frac{1}{N}\Sigma x f_x, \quad m_r = \frac{1}{N}\sum_x (x - m_1')^r f_x, \quad r \geq 2$$

i.e. m_1' is the mean and m_r are the rth central moments of the distribution. The results are shown in the following table.

m_1'	m_2	m_3	I	S
3.6995	6.0755	9.4164	1.6423	1.5499
3.6687	6.1111	8.9259	1.6657	1.4606
3.3668	5.4414	10.0266	1.6162	1.8427
3.1014	4.7377	8.7363	1.5276	1.8440
3.8739	6.7362	9.5955	1.7389	1.4245
3.4812	5.7434	10.5059	1.6498	1.8292
3.6025	5.3230	7.4252	1.4776	1.3949
3.2838	5.3473	10.8909	1.6284	2.0367
4.0104	7.0608	9.1494	1.7606	1.2958
3.1753	4.5446	7.9465	1.4312	1.7486
3.0868	4.8445	9.8663	1.5694	2.0366
3.5205	5.8830	9.6162	1.6711	1.6346
3.2965	5.5581	10.8315	1.6861	1.9488
2.9695	4.8588	11.0969	1.6362	2.2839
3.0749	5.6464	13.9645	1.8363	2.4732
3.4977	4.7846	5.3274	1.3679	1.1135
3.4219	5.6302	9.6423	1.6453	1.7126
2.9327	5.3961	13.3551	1.8400	2.4750
3.1663	4.9679	9.4305	1.5690	1.8983
3.2033	6.0541	13.1951	1.8900	2.1795

The points <I,S> can be displayed graphically in order to give an optical impression (cf. Fig. 1).

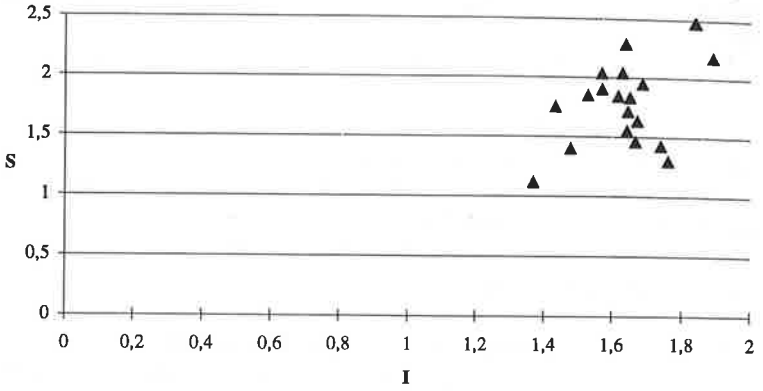


Fig. 1: Ord's criterion for Portuguese press texts

As we can see, the empirical frequency distribution of the word classes in all 20 texts are clearly following both the model of the 1-displaced mixed Poisson and the model of the negative hypergeometric distribution. Regarding the examination of Brazilian-Portuguese texts (Ziegler, 1998b) we ascertained a few difficulties in justifying and applying the mixed Poisson distribution. It is assumed that these differences of fitting refer to an areal problem. Particularily in Brazilian-Portuguese there arises the problem of interlingual influences as a result of intensive multiple language contacts, that is various interferences from different languages are manifested in Brazilian-Portuguese. Therefore, differences within one individual language are probably caused by different linguistic influences and social environments. Comparing Portuguese and Brazilian-Portuguese a better founded answer could be given by a historical analysis or by exploring other text types. This is the second analysis that can corroborate in a preliminary way the assumption of a non-accidental distribution of word classes in Portuguese press texts, but the question of whether other types of Portuguese texts are in line with the Poisson or the negative hypergeometric distribution still remains and has to be answered by further studies.
Nevertheless, concerning the text type *press texts*, both varieties of the Portuguese language can be fitted to the two models mentioned above.

Sources

text 1: Dias, João Miguel (1996): Em busca da deontologia. In: Público, segunda-feira, 20 Maio. Lisboa/Porto, 3.
text 2: Dias, João Miguel (1996): Uma classe protegida. In: Público, segunda-feira, 20 Maio. Lisboa/Porto, 3.

text 3: Amado, Joana (1996): O mundo das 260 licenciaturas. In: Público, segunda-feira, 20 Maio. Lisboa/Porto, 3.
text 4: Ralha, Leonardo (1996): À espera da ordem. In: Público, segunda-feira, 20 Maio. Lisboa/Porto, 3.
text 5: Viana, Luís Miguel (1996): Faltam qualificações, há trabalho. In: Público, segunda-feira, 20 Maio. Lisboa/Porto, 4.
text 6: Viana, Luís Miguel (1996): Desemprego ou mais mercado. In: Público, segunda-feira, 20 Maio. Lisboa/Porto, 3.
text 7: Pessoa, Carlos (1996): Com a saúde pública nas mãos. In: Público, segunda-feira, 20 Maio. Lisboa/Porto, 4.
text 8: Talixa, Jorge (1996): Centro Apoio ao Empresário nas instalações da Chemina. In: Público, segunda-feira, 20 Maio. Lisboa/Porto, 47.
text 9: Graça, Franco (1996): As "boas"e "más" razões da desinflação. In: Publico Economia, segunda-feira, 20 Maio. Lisboa/Porto, 15.
text 10: Catalão, Rui (1996): Uma avaria nas hélices. In: Público Cultura, segunda-feira, 20 Maio. Lisboa/Porto, 31.
text 11: Lapa, Fernando C. (1996): Singular e plurais. In: Público Cultura, segunda-feira, 20 Maio. Lisboa/Porto, 31.
text 12: Delgado, Alexandre (1996): Da indigestão á obra-prima. In: Público Cultura, segunda-feira, 20 Maio. Lisboa/Porto, 31.
text 13: Agualusa, José Eduardo (1996): Uma noite com África. In: Público Cultura, segunda-feira, 20 Maio. Lisboa/Porto, 30.
text 14: Gomes, Manuel João (1996): Uma beleza de planeta. In: Publico Local, segunda-feira, 20 Maio. Lisboa/Porto, 64.
text 15: Pedrosa, Maria Ermelinda (1996): Porcelanas de qualidade na Europa a preços competitivos. In: Correio de Portugal, Ano III Nº 21, Maio. Dortmund, 23.
text 16: Calvário, Renato (1996): Centro cultural organizou uma festa agradável. In: Correio de Portugal, Ano III Nº 21, Maio. Dortmund, 16.
text 17: Rita, Jorge Martins (1996): Centro juvenil organizou "Baile da Pinha". In: Correio de Portugal, Ano III Nº 21, Maio. Dortmund, 16.
text 18: Correio de Portugal, Redacção (1996): "O Etnográfico" de Solingen venceu (mais uma vez) festival de folclore. In: Correio de Portugal, Ano III Nº 21, Maio. Dortmund, 16.
text 19: Correio de Portugal, Redacção (1996): Muito futebol e muita garra. In: Correio de Portugal, Ano III Nº 21, Maio. Dortmund, 15.
text 20: Correio de Portugal, Redacção (1996): "Os Águias" venceram torneio de futebol. In: Correio de Portugal, Ano III Nº 21, Maio. Dortmund, 14.

References

Altmann, G. (1981). The Homogeneity of Metric Patterns in Hexameter. In R. Grotjahn (Ed.), *Hexameter Studies* (pp. 137-150), Bochum: Brockmeyer.

Altmann, G. (1991). Modelling Diversification Phenomena in Language. In U. Rothe (Ed.), *Diversification Processes in Language: Grammar* (pp. 33-46), Hagen: Rottmann.

Altmann, G. (1991). Word Class Diversification of Arabic Verbal Roots. In U. Rothe (Ed.), *Diversification Processes in Language: Grammar* (pp. 57-59), Hagen: Rottmann.

Arapov, M.V., & Šrejder, J.A. (1977). Klassifikacija i rangovye raspredelenija. *Naučno-techničeskaja informacija, Ser.*, 2, 1-12, 15-21.

Best, K.-H. (1994). Word Class Frequencies in Contemporary German Short Prose. *Journal of Quantitative Linguistics*, 1, 144-147.

Büntig, K.-D., & Bergenholz, H. (1989). *Einführung in die Syntax*. Frankfurt a. M.: Athenäum.

Bußmann, H. (1983). *Lexikon der Sprachwissenschaften*. Stuttgart: Kröner.

Caetano, J.A.P., Mayr, J.J., Plachy, R., & Ptacek, F. (1992). *Grammatik Portugiesisch*. München: Hueber.

Cunha, C., & Cintra, L. (1984). *Nova Gramática do Portugûes Contemporâneo*. Lisbóa: Fundação Nacional de Material Escolar. Ed. Ministério da Educação e Cultúra.

Fucks, W. (1955). *Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen*. Köln, Opladen: Westdeutscher Verlag.

Grotjahn, R., & Altmann, G. (1993). Modelling the Distribution of Word Length: Some Methodological Problems. In R. Köhler & B. Rieger (Eds.), *Contributions to Quantitative Linguistics* (pp. 141-135), Dordrecht: Kluver.

Hammerl, R. (1990). Untersuchungen zur Verteilung der Wortarten im Text. *Glottometrika 11* (pp. 142-156), Bochum: Brockmeyer.

Köhler, R. (1991). Synergetic Modelling of Grammatical Diversification Phenomena. In U. Rothe (Ed.), *Diversification Processes in Language: Grammar:* (pp. 47-56), Hagen: Rottmann.

Ministério da Educação e Cultúra (Ed.) (1979[5]). *Gramática da Língua Portuguesa*. Rio de Janeiro: Fundação Nacional de Material Escolar.

Ord, J.K. (1972). *Families of Distributions*. London: Griffin.

Rothe, U. (Ed.) (1991). *Diversification Processes in Language: Grammar*. Hagen: Rottmann.

Schweers, A., & Zhu, J. (1991). Wortartenklassifikation im Lateinischen, Deutschen und Chinesischen. In U. Rothe (Ed.), *Diversification Processes in Language: Grammar* (pp. 157-167), Hagen: Rottmann.

Wimmer, G., & Altmann, G. (1994). A Model of Morphological Productivity. *Journal of Quantitative Linguistics*, 2, 212-216.

Zhu, J., & Best, K.-H. (1992). Zum Wort im modernen Chinesisch. *Oriens Extremus,* 35, 45-60.

Ziegler, A. (1996). Word Length Distribution in Brazilian-Portuguese Texts. *Journal of Quantitative Linguistics*, 3, 73-79.

Ziegler, A. (1998a). Word Length in Portuguese Texts. In G. Altmann, J. Mikk, P. Saukkonen & G. Wimmer (Eds.). Linguistic structures. To honor J. Tuldava. *Journal of Quantitative Linguistics,* 5, 115-120. (Special Issue).

Ziegler, A. (1998b). Word Class Frequencies in Brazilian-Portuguese Texts. *Journal of Quantitative Linguistics*, 5,3,269-280.