QUANTITATIVE LINGUISTICS

Volume 54

Juhan Tuldava

METHODS

in

Quantitative Linguistics

preface by G. Altmann

WVT Wissenschaftlicher Verlag Trier

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Tuldava, Juhan :
Methods in quantitative linguistics /
Juhan Tuldava. Pref. by G. Altmann. -
Trier : WVT Wissenschaftlicher Verlag Trier, 1995
(Quantitative linguistics ; Vol. 54)
ISBN 3-88476-126-9
NE: GT

Umschlag: Brigitta Disseldorf
(Marco Nottar, Agentur für Werbung und Design, Konz)

ISBN 3-88476-126-9
ISSN 0179-3616

WVT Wissenschaftlicher Verlag Trier
Bergstraße 27, 54295 Trier
Postfach 4005, 54230 Trier
Tel. 0651-41503, Fax 41504

Contents

Preface

Science should be free, democratic and international. The scientist should have freedom of speech, should (have the possibility to) take into account the research of other scientists and should publish his/her own results in a generally understandable language. Though in history there were (and still are) religious and political institutions which have applied bans on publication, putting books on the black list, burning of books and men, exile, excommunication, prisons, concentration camps and other effective means in their struggle against science, such measures could not hinder forever its course, merely slow it down. Even in very unfavourable circumstances, including language barriers, works could always come into existence which meant an enrichment for science. Later on one could pick them up like scattered pearls merely to discover that, as a matter of fact, they filled existing gaps.

Half talking J. Tuldava round, half begging him, the editorial board of this series succeeded in persuading him that he should voluntarily publish at least a part of his Estonian and Russian papers in the form of an English book. He himself made the choice of the papers, the translations were prepared by Estonian specialists (see below) whose excellent English was cautiously and with much feeling "canadianized" by Sheila Embleton. The editorial board cordially thanks all of them. J.T. authorized the translations, completed the literature and *mutabat mutanda*.

The decision to publish a selection of his works in this series (besides a monograph which will appear in another place in 1994) has been made on several grounds:

(i) J.T's works contain perhaps the most valuable and durable ideas that can be recovered from the rubble of "Soviet" linguistics. For modern (qualitative) linguistics the ephemerality of its particular approaches is characteristic: paradigms come and go, the fall of the next one is programmed in advance because they all are oriented towards an antiquated philosophy of science and thus cannot be incorporated in the bulk of science. Tuldava's work, however, displays a seamless amalgamation with the general tendencies in all sciences heading to mathematization, systems theory (synergetics) and methodology of natural sciences. This ideal - not even present as such in qualitative linguistics

- is a substantial part of the "Zipfian linguistics" which has been in the process of incessant expansion for seventy years. The introductory article, *Some notes on the methodology of quantitative linguistics*, shows the perspective which is gaining acceptance in quantitative linguistics.

(ii) Juhan Tuldava (*1922), Professor emeritus at the University of Tartu (Estonia), is the founder of the Tartu school of quantitative linguistics and was for many years the head of the text-analytical group of quantitative linguists in the former SU, for whom it was not easy even to maintain their right to exist if they were not willing to serve an ideology. I still remember the time when "mathematization" was officially identified there with "dehumanization" and cybernetics was called "bourgeois pseudo-science". J.T. is not only a "grand old man" - he is perhaps the most active grand old man among linguists - but his works were unfortunately published mostly in "scientifically dead" languages. Since the errors of history are corrigeable - at least in linguistics - his works and the best ones of the group mentioned will be published in this series in English.

(iii) The present book yields not only a wealth of theoretical and empirical insights in texts and languages, but is also an excellent text-book of methods of quantitative linguistics. The lucidity and the simplicity with which quite complicated affairs are mediated here have a scarcity value in this domain of research. It offers us instruments which we must laboriously seek in the labyrinthine world of mathematics. At the same time it opens a new world of problems which can grow from a pilot flame to a fire in Western quantitative linguistics.

The fundamental aim of any scientific research is the discovering of the mechanism of becoming which is equivalent to explanation. In the article *On causal relations in language* J.T. takes a modern and reasonable standpoint: causal relations are connected with some conceptual system. The problem of causality that has not released us from its fangs since Aristotle is nowadays out of fashion. On the one hand, it is treated merely as a special case of determination (cf. Bunge 1959); on the other hand, even the Aristotelian forms of causality are called into query (cf. e.g. Riedl 1982). Where is its place in the world dominated by stochastic processes and in which there is more chaos than order? In linguistics one rather speaks about *influence*[1], and the present research on causality in language is to be understood in this way. In this sense it acquires relevance because it is in agreement with the modern efforts to introduce explanations in linguistics in the form of self-regulation and self-organization schemes. The classical explanation schemes (e.g. Hempel-Oppenheim) are problematic even in purely deterministic cases; modern science explains rather by means of imbedding a phenomenon in a nomological net (cf. Salmon 1984),

[1] "Whatever causes probabilifies but not conversely. Anything that probabilifies (increases chance propensities) may be called *influence*" (Bunge 1977, Vol. 3: 211).

which in the domain of synergetic linguistics based on self-regulation and adhering to functional analytic procedures happens without serious problems. Using the concept of *influence* one can successfully operate with conditional probabilities and obtain valuable hypotheses. Changes in conditional probabilities (against non-conditional ones) signal the existence of *bonds* between two entities which come into existence "if at least one of them acts upon the other" (Bunge 1977, Vol. 3: 261). Their discovery is a way to establish systems. Correlation and regression are also explorative means of this kind but they cannot replace creative deductive work. In self-regulation cycles both the direction and the extent of influence are relevant. Sometimes both can be discovered by means of a conditional-probability-technique. Here J.T. shows among other things that against the intuitive assumption that frequency influences word length more than conversely, the empirical evidence corroborates the latter direction. If the probability of B under the condition of the existence of A changes more than the other way round, then it is an indication that in a self-regulating cycle the way from A to B is shorter than from B to A.

The problem of influence is the contents of the chapter *On the measurement of correlation between qualitative features in linguistics: contingency of alternative features*, in which a number of methods are presented with the aid of which we can compute the probability of association, contingency, dependence, correlation etc. (e.g. chi-square test, Fisher's exact test, analysis of residuals, different measures of association, entropy and informations, etc.). All methods are illustrated on linguistic data and will surely serve as a stimulus for further research. No text-book of statistics explains so much and in such a lucid way.

Seven articles of this volume are dedicated to text analysis. In *An attempt at quantitative analysis of the style of fiction* and in *A comparison of subjective and objective characteristics of style*, the exploratory way of inquiry mentioned above is applied. In texts of seven Estonian authors some properties are analyzed (in the first article: parts of speech, entropy, concentration, hapax legomena, etc.) and submitted to factor analysis. The result is the discovery of four factors: lexical concentration, activity, richness and analyticality, and a classification of the authors. In the second article, in which besides objective properties subjective judgements of experts are also used, the factors emotionality, readability, intellectuality and popularity are found. Methods of this kind not only enrich our taxonomic knowledge (see Bunge 1983: Vol 6:16ff.) but are even able to enlarge our theoretical knowledge if, for example, we join the properties loading on one factor in a self-regulating cycle and set up hypotheses about the kind of these bonds. No property (feature) in text or language is isolated - it unfolds or develops in cooperation and competition with some others. A linguist can tentatively examine the *existence* of a relation between objective and subjective properties in an exploratory way, but in order to grasp its form and its mechanism he/she possibly needs the help of a psychologist or that of a non

conventional literary theoretician, and, in any case, the brain of a mathematician. Thus progress in this domain depends on cooperation between scientists. Only with joint efforts shall we be able to penetrate into the mysteries of texts and languages.

Two articles concern the comparison of the vocabulary of two texts. In *On the lexical connection of texts* vocabularies of texts are compared without reference to frequencies. The author shows the computation of some indices, confidence intervals and statistical tests for the lexical homogeneity of texts. In *A statistical method of comparison of lexical composition of two texts*, the frequency distributions are compared. The problems as such are not new - they have been discussed several times by G. Herdan and by French researchers - nevertheless, they give rise to several questions:

(i) How can we test the homogeneity of *one* text, i.e. how can we find breaks in the text, how can we ascertain that the text has not been written under a unique regime of creation? The only possible answer is the agreement of the given text properties with a theoretical model - if there is one at hand.

(ii) From this follows immediately the question of what the models are according to which the individual properties are distributed sequentially or globally. One of the possible solutions follows from so-called synergetic modelling, which enables us to derive both control cycles and probability distributions using the same approach. The Waring-distribution presented in the second article can also be derived from such a model (see below).

(iii) Can we conclude from a disturbance (break) of a special kind that at the given place in the text (a) there is an intentional code-switching, (b) a pause in writing took place or the theme changed, (c) the text was corrected by others?

(iv) Which text properties tell us something about the kind and place of break? For example, the transition from indirect to direct speech is a manifest break of type (a); a break in a type-token curve is a break of type (b), etc.

(v) A quite different problem is the homogeneity of *two* texts. This problem is so multidimensional that it can be approximated merely stepwise, in small one-dimensional steps. The time at which all properties of texts will be connected in a control cycle and the differences will follow from the values of the parameters still lies in the dark and uncertain future.

Two articles are concerned with the relation between text length and the increase of new words in text, known as the TTR-problem (type-token-ratio). In *On the relation between text length and vocabulary size* the increase of the vocabulary is divided into phases and the author introduces two Tornquist functions whose parameters are interpreted linguistically and he ascertains that the relation L/N is a special case of the Zipf-Mandelbrot function. In the second article, *The ratio of word forms and lexemes in texts*, the allometric function is used and the ratio L/V is interpreted as the coefficient of analyticism of the language.

This problem is one of the most discussed ones in text analysis. Up to now about fifty papers have been written, even by pure mathematicians (e.g. B. Brainerd, J. Gani, D. McNeil); however, the balance-scale seems to fall in favor of Tuldava and his allies. Both approaches in this volume are in congruence with the synergetic modelling of language phenomena, which has recently gained acceptance. For example the simpler Tornquist function can be derived from the linguistic synergetic approach as $y_x = g(x)y_{x-1}$ with $g(x) = x(b+x-1)/(b+x)(x-1)$ and $y_1 = a/(b+1)$, or the Zipf-Mandelbrot function can be presented with $g(x) = [(b+x-1)/(b+x)]^c$ and its adequacy in modelling some phenomena can hardly be proven wrong. If there is a collision of theoretical approaches, then the one that can be derived from a more general theory and has a better linguistic foundation will be the winner. Those which start from ad hoc assumptions and miss a linguistic foundation will be refuted even if they are mathematically more sophisticated.

The "TTR-problem" is a problem of information unfolding in text, the examination of which is still in its infancy. However, if we succeed in penetrating into the mysteries of the text on this path, then not only linguistics but also other sciences can only stand to gain from this research. Some possibilities and vistas are presented e.g. by Hřebíček (1994) and Wildgen (1994).

The generalization of the TTR-problem brings an overwhelming number of problems. Let us mention at least some of them:

(i) Which are the properties and dimensions of texts and how do they unfold in the linear course of text?

(ii) Is this unfolding deterministic, stochastic or chaotic? Nobody believes in the first case - only applied linguistics struggles with it - the latter two are our merciless gnoseological and methodological enemies.

(iii) Which methods should be applied in order to grasp processes which present themselves as time series, stochastic and chaotic sequences? Does the text have its own mathematics that has not been discovered as yet?

(iv) How is it possible that the linear unfolding of texts gives rise to nonlinear structures which produce a text-form out of a selected knowledge and in turn an interpreted text out of a text-form (cf. Hřebíček 1994, Herrmann & Hoppe-Graf 1988)?

(v) How can we integrate knowledge about the linear unfolding of the text into the theory of language self-regulation? Etc.

The last article, *Quantitative analysis of the phonemic system of the Estonian language,* analyzes the frequency structures of the Estonian phonemic system and compares it with several other languages. Special attention is given to the ranking of phonemes according to their frequency. The author shows different functions: the Zipf function, two exponential functions, the Zipf-Mandelbrot and the Zipf-Alekseev functions, a logarithmic function and a new one - which can be baptized the Zipf-Tuldava function - yielding quite good residuals. Some of

the curves bring up the problem of parameter estimation. It can be shown in any case that point estimators should be used merely as starting values and the curve fitting should be done by means of an optimization procedure.

Here, too, problems arise, the solution of which could be of much help for us:

(i) How should we estimate the parameters of the curves if the classical estimators fail? Is the use of p_n (the frequency of the highest-ranking (last) class) relevant for the estimation at all, if it is that small and loaded with fluctuation?

(ii) Even the first class (most frequent) is problematic because in many cases it yields great residuals. Since it is the source of the greatest contribution to the generation of redundancy, it should perhaps be distinguished and the rank curve modified, for example: $p_1 = \alpha$, $p_x = (1-\alpha)f_x$ (x = 2,3,...,n), a technique which can be linguistically well founded.

(iii) Can this problem be generalized, i.e. do all language entities display ranking regularities? If so, are there general ranking laws, and how good are the curves presented here as candidates for laws?

(iv) Are there ranking differences between languages, texts, entities etc., or are there none?

(v) What is the relation between the number of phonemes in the inventory or the parameters of the ranking curve and other properties of language, e.g. word length, phoneme distribution, tone, accent, etc.? In other words, how can we integrate the ranking problem into the system of synergetic linguistics?

As is evident, this book can serve both as a problem generator and a method donator, not only for quantitative but also for qualitative linguistics. It opens many doors and gives us vehicles that can be used for passing their thresholds. We hope to present herewith the best part of Estonian linguistics.

G. Altmann

References

Bunge, M. (1977). *Treatise on Basic Philosophy. Vol 3. Ontology I: The Furniture of the World.* Dordrecht, Reidel.

Bunge, M. (1983). *Treatise on Basic Philosophy. Vol 6: Epistemology & Methodology II: Understanding the World.* Dordrecht, Reidel.

Bunge, M. (1959). *Causality.* Cambridge, Mass., Harvard University Press.

Herrmann, T., Hoppe-Graff, S. (1988). Textproduktion. In Mandl, H. & Spada, H. (eds.), *Wissenspsychologie: 283-298.* München/Weinheim, Psychologie Verlags-Union.

Hřebíček, L. (1994). *Text Levels.* Trier, WVT (to appear).

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

Riedl, R. (1982). *Evolution und Erkenntnis.* München, Piper.

Salmon, W.C. (1984). *Scientific Explanation and the Causal Structure of the World.* Princeton, N.J., Princeton University Press.

Wildgen, W. (1994). The distribution of imaginistic information in oral narratives: a model and its application to thematic continuity. In: Hřebíček, L., Altmann, G. (eds.), *Quantitative Text Analysis: 175-199.* Trier, WVT .

Sources of the articles in this volume

Nr. 1. Metodologičeskie problemy sovremennoj kvantitativnoj lingvistiki. In: Perebejnos, V.I. (ed.), *Statističeskaja leksikografija i učebnyj proces.* Kiev, KGPIIJa 1990, 1-18. Translated by Ilmar Anvelt.

Nr. 2. O verojatnostno-statističeskom modelirovanii pričinno-sledstvennych zavisimostej v jazyke. In: *Evrističeskie vozmožnosti kvantitativnych metodov issledovanija jazyka.* Smolensk 1991: 9-11. Translated by Mall Tamm.

Nr. 3. Ob izmerenii svjazi kačestvennych priznakov v lingvistike (1): soprjažennost' alternativnych priznakov. *Acta et Commentationes Universitatis Tartuensis 827, 1988, 146-162* and Statistiline sõltuvus keeleteaduses (1). In: Tuldava, J. (ed.), *Linguistica (Tartu) 6, 1975, 169-188.* Translated by Malle Laar.

Nr. 4. Opyt kvantitativnogo analiza chudožestvennogo teksta. In: Põldmäe, J. (ed.), *Studia Metrica et Poetica (Tartu) 1, 1976, 122-141.* Translated by Ilmar Anvelt.

Nr. 5. K probleme sopostavlenija subjektivnych i objektivnych charakteristik stilja. *Studia Metrica et Poetica (Tartu) 2, 1977, 82-93.* Translated by Ilmar Anvelt.

Nr. 6. Ob izmerenii leksičeskoj svjazi tekstov na urovne slovarja. In: Golovin, B.N. (ed.), *Voprosy statističeskoj stilistiki.* Kiev, Naukova dumka 1974, 35-44. Translated by Mall Tamm.

Nr. 7. Statističeskij metod sravnenija leksičeskogo sostava dvuch tekstov. *Linguistica (Tartu) 4, 1971, 199-220.* Translated by Mall Tamm.

Nr. 8. O statističeskoj strukture teksta. In: Liimets, H. et al. (eds.), *Sovetskaja pedagogika i škola (Tartu) 9, 1974, 5-33.* Translated by Ilmar Anvelt.

Nr. 9. O sootnošenii slovoform i leksem v tekste. In: Bartkov, B. (ed.), *Sistemnyj analiz lingvističeskich javlenij v tekste.* Vladivostok, ANDO 1988, 94-102. Translated by Laine Hone.

Nr. 10. Opyt kvantitativnogo analiza sistemy fonem estonskogo jazyka. *Acta et Commentationes Universitatis Tartuensis 838, 1988, 120-133.* Translated by Malle Laar.

1

Some Notes on the Methodology of Quantitative Linguistics

In recent years the methodology of quantitative linguistics has received a great deal of attention in the pertinent literature. The development of linguistic theory and of appropriate mathematical models is the common theme of books such as those by Altmann (1980, 1988), Grotjahn (1979), Piotrowski (1979, 1990), Těšitelová (1985), Embleton (1986), Köhler (1986), Perebejnos (1990), Altmann & Schwibbe (1992), Hřebíček (1992), Saukkonen (1992), Köhler & Rieger (1993) and many others. The present article deals with only a few of the problems concerning the general principles, the quantitative-systemic and interdisciplinary approaches and the ways of evolution of quantitative linguistics in our days.

1. The status of methodology

By methodology of investigation we understand an ordered set or system of separate methods as well as the use of this system. From the viewpoint of scientific knowledge and generalization, concrete practical methods of investigation are on the level of "technique and procedures". This is the lowest level in the hierarchy including the higher levels of philosophical, general-scientific and concrete-scientific methodology (cf. Tuldava 1980-1982). By attributing the practical ways of problem solving to the lowest level, we do not mean, however, that it loses its ties with the higher levels mentioned. The aim of science is not merely to discover simple facts and correlations but to generalize and to seek for higher and higher generalizations, calling for more and more abstract and complex concepts, setting up hypotheses and theories.

Quantitative-linguistic analysis, like any other systemic analysis, is characterized not by a specific body of concrete methods but by an ordered, logically grounded approach to the application of the existing methods already elaborated in other sciences and disciplines (mathematics, linguistics, etc.). The choice and combination of these methods are of greatest importance both on the level of observation and experiment and on the level of analysis and theoretical generalization of the results. The elaboration of a system of methods of research is

especially important on the boundary between sciences, including quantitative linguistics, where the choice of adequate methods can promote the discovery of new linguistic phenomena and dependences. In addition to the usual methods of "quantitative" mathematics (mathematical statistics, theory of probabilities, information theory, some sections of mathematical analysis) there are methods available which help us establish the relative importance of different sources of causality or contingency (see, e.g., the corresponding articles in this issue) and testing procedures, including the analysis of residuals, which may give us confidence in the scientific correctness of our investigation and its results (or lead us to their "falsification").

On the level of concrete research a general approach has to be determined which specifies the central notions and principles of the given investigation (or a cycle of investigations). Present-day quantitative linguistics can be characterized by a great number of approaches to the modelling of linguistic material. We could mention the informational approach and linguistic engineering (Piotrowski 1979; Piotrowski et al. 1990); modelling the "Zipf size" (Orlov 1982); the "variational" approach (Arapov 1988); the stylometric approach (Martynenko 1988); the quantitative typology of text (Alekseev 1988); micro- and macro-world mechanisms (Králík 1990); modelling the word's life cycle (Polikarpov 1993); text as a construct of aggregations (Hřebíček 1992); dialectometry (Embleton 1993; Goebl 1993); and many others. Nowadays, the trend known as "language synergetics" has gained special popularity (Köhler 1986); a great number of investigations have been accomplished as part of the "Language Synergetics Project" directed by G.Altmann and R.Köhler (e.g. Saukkonen 1982; Hammerl & Sambor 1993; analogous works by Silnitsky 1993, and others). The fundamental axiom of synergetic linguistics is that language systems possess self-regulating and self-organizing control mechanisms which "change the language towards an optimal steady state and an optimal adaptation to its environment - in analogy to biological evolution" (Köhler 1993: 41). Discovering dependences between different levels of language, revealing their mechanism and explaining their causes is significant for the theory of language synergetics. The final aim is building an adequate linguistic theory within the framework of the theory of self-organizing systems.

On the general scientific level linguistic objects and research procedures are viewed in the context of such notions as *interdisciplinarity, isomorphism, bipolarities (quantity - quality, concentration - dispersion, unification - diversification, etc.), theory and regularities (text laws)* and others. Present-day works on quantitative linguistics deal with these general methodological notions in one or another way and they are included in the general procedure of describing and explaining the concrete material under study.

Finally, we have to admit that "the content of any scientific theory will in fact overlap with philosophy" (Lewis 1966: 771). On this highest level of generalization the key-words are *ontology, synergy, integrity, causality, cognition* and others which belong to the fundamental philosophical problems and concepts related to scientific investigation (see e.g. Seppänen 1992).

2. Quantitative-systemic aspect

In present-day systemic investigation, along with the so-called dynamic systems (with a strong regularity of relations and connections between the elements), a significant role also belongs to *probabilistic systems* whose integrity and stability are combined with sufficiently great autonomy of their parts (cf. Tuldava 1987). The parameters of such systems belong to different levels, they can be divided into two classes: incidental or occasional events (on a lower level) and regularities in a mass of occasional events (on a higher level). At that the higher-level characteristics which determine the general structure of the system do not determine every concrete occasional event, to be more exact - they determine these occasional events only in a generalized, integral way. Such an approach makes it clear that any kind of seemingly "asystemic" or "antisystemic" phenomena may be organic parts of probabilistic systems which are characterized by both regular and irregular (occasional, transitional, variable, etc.) properties.

Treating objects of investigation as probabilistic systems is the theoretical basis for modelling the objects by means of *statistical distributions*.

Distribution as a generalizing, integral notion is the most important structural characteristic of probabilistic systems; it expresses not only the existence of internal regularities in the system (interactions between its elements), but also the integrity and stability of the system as a whole.

The notion of distribution can be treated in the broad sense as an ordered aggregate of quantitatively expressed values, i.e. results of the object's (objects') measurement, usually with the indication of their significance (frequency, probability, rank) within the given set. For a deeper understanding of the structure of distribution it is necessary to familiarize oneself with the notion of measurement as the source of formation of elements included in the distribution.

Measurement can be defined as a procedure of ascribing numerical values to the observed objects according to definite predetermined rules. It can be easily shown that the logical foundation of measurement is the categories of *property* and *relation*. When measuring, we find out the relation of an object (X) to a

quantitatively expressed value (*Y*) on the basis of a property (*P*). This abstract structure can be presented as follows:

$$X \xrightarrow{P} Y$$

where the component "→" expresses the "operator" under which the object (*X*) is brought into relation with the value (*Y*) on the basis of a quantitative property (*P*), i.e. such a property (feature, characteristic) which admits of quantitative estimation (e.g. size, number, frequency, extent). The following statement can serve as an illustration of the formula:

"The text (X) is connected with the quantitatively expressed value (Y) in the sense that the size of the text (P) is equal to this value."

In a more common form this statement can be expressed as follows:

"The size (P) of the text (X) equals Y".

This example shows that the quantitative presentation of the object on the basis of some property is formally ascribing a quantitative value not to the object, but to the property of the object. As far as the notion "the value of the property" (*Y*) is concerned it should be specified that as a rule it consists of two elements: of a quantity (number) and units of measurement accepted for the given property (e.g. "the length of the word is equal to *6 letters*"). It can also be added that besides purely numerical values of the properties there also exist evaluative ones: evaluative-quantitative meanings (very little, moderately, much, very much, etc.) and evaluative-binary ones (present - absent, much - little, etc.). Evaluative meanings can be used in definite conditions or at certain stages of analysis. When necessary these evaluative meanings can be replaced by numerical signs, for example, when evaluative-quantitative values are expressed by "points" (forming an ordered scale) or when *1* and *0* are ascribed to binary relations (forming the so-called dichotomous scale).

Besides the differentiation of theoretical and empirical distributions we can also distinguish between static and dynamic distributions. Static distributions, also called synchronic distributions, express chiefly synchronic and paradigmatic aspects of the analysis of linguistic phenomena. Dynamic distributions in linguistics differ from static ones by the fact that they express either the process connected with speech generation phenomena (in synchrony) or the change, the development, of language (in diachrony).

Technically, any distribution can be presented in the form of a table, graph or mathematical formula (function). In all these cases the distribution can be either differential (non-cumulative) or integral (cumulative).

Without going into technicalities of preliminary arrangement and statistical estimation of material (grouping, calculation of dispersion characteristics, etc.), we shall dwell on those variants of data presentation in the form of distributions which correspond to the basic rules of measurement expounded above. We proceed from the measurement formula in which three major components correlate with one another: object (*X*), property or feature (*P*), and value, i.e. result of measurement (*Y*). These three components of measurement presented in a table can serve as a basis for revealing the corresponding distributions. We suggest three main schemes for making tables of distribution.

Scheme 1. *Uni-object* distribution
(one object, several features)

	P_1	P_2	...	P_n
X	Y_1	Y_2	...	Y_n

According to this initial scheme one object correlates with the measurement results under different conditions. For example, if *X* is a concrete linguistic unit (or class of units), P_i means "frequency in text *i*", and Y_i the values of corresponding frequencies, it is possible to trace the behaviour of this linguistic unit in a series of tests. If we range the data in the order of increase or decrease of the quantitative values of *Y* it becomes possible to make up a variational series or to form a rank distribution (ascribing ranks to Y_i values).

Scheme 2. *Multi-object* distribution
(several objects - one feature)

	P
X_1	Y_1
X_2	Y_2
...	...
X_m	Y_m

According to this scheme different objects are measured by the same common feature. For example, if X_i are different linguistic units (or classes of units) and

P means "frequency in a given text", then Y_i are concrete frequency values of the linguistic units in the given text. An ordinary frequency dictionary as well as any frequency list of some other kind of linguistic units is compiled according to this scheme. Two forms of presentation are distinguished here: a) "spectral" distribution when identical results of measurement are combined into groups and the number of items with this result of measurement given; b) rank distribution, under which ranks (i) are ascribed to ranged frequency values (Y_i).

Scheme 3. *Complex (multidimensional)* distribution
(several objects - several features)

	P_1	P_2	...	P_n
X_1	Y_{11}	Y_{12}	...	Y_{1n}
X_2	Y_{21}	Y_{22}	...	Y_{2n}
...				
X_m	Y_{m1}	Y_{m2}	...	Y_{mn}

This scheme is a combination of the first two. In its simplest form the number of objects (X) and the number of features (P) equal two, for example, when the frequencies of two different words in two different texts are compared or the frequencies of different classes of words are compared in two aspects - in the vocabulary and the text. The scheme of complex distribution helps to solve multivariate problems: interconnection and joint variation of a number of objects or features. The correlation between distributions of quantitative values (horizontal and vertical; see Scheme 3) can be expressed by a functional dependence (by an equation of regression) and the strength of this correlation can be measured by the correlation coefficient (linear or non-linear). As far as inner relations within the whole aggregate of data are concerned, it is possible to reveal them by means of factor analysis, cluster analysis and other types of classificatory analysis.

Thus, all the variants of distributions described above can contribute to the quantitative-systemic analysis and modelling of linguistic objects. Preference of modelling by means of distributions under the quantitative-systemic approach arises from the specific character of systemic analysis itself. It consists in the fact that the object is studied in the aspect in which it represents a system under the given approach. In this sense linguistic distribution is a model-description of those language objects which can be regarded as probabilistic systems. The aim of the analysis of linguistic objects as probabilistic systems is the revelation

of deep quantitative and qualitative regularities of structure and functioning of language and speech along with the obligatory interpretation of the results of the analysis.

3. Interdisciplinary approach

The inclination to semiotic and systemic generalizations characteristic of present-day investigations has led the researchers in the field of quantitative linguistics into a close contact with representatives of some other areas of semiotics - theories of music, painting, motion pictures, and theatre as well as with psychologists, educationalists, sociologists, and physicians, not to mention mathematicians and physicists. On this ground the interdisciplinary sphere of language and text research is formed. The foundation of the international research team "Text as an object of interdisciplinary studies" in 1985 (see, e.g., Boroda & Dolinskij 1988) was a concrete embodiment of this trend. The team united within the framework of quantitative text analysis researchers of different specialities from several universities and research institutions in Estonia, Latvia, Russia, Ukraine, Georgia and Armenia. Annually (1985-1991) the team arranged seminars and conferences. The collection "Quantitative Linguistics and Automatic Text Analysis" was published yearly in the series of Transactions of the University of Tartu (*"Acta et Commentationes Universitatis Tartuensis"*).

The most interesting studies in this field are the investigations of regularities of statistical organisation of musical texts (see, e.g., Boroda 1982), in particular the discovery of parallels in the composition of musical and literary texts (cf. Zörnig & Boroda 1992). Co-operation with psychologists has also proved fruitful. Representatives of psychophysiology have connected some types of linguistic distributions as models of speech activities with the structural peculiarities of the brain, in particular with the spatial and temporal organization of cyclic processes in the brain. Proceeding of the idea about codification of word images as "packets of waves of neuronal activity", A.N. Lebedev (1986) deduced a formula which coincides with Zipf's formula for the description of word frequency distribution in speech, and another formula which describes the connection between vocabulary growth and text size.

Interdisciplinary studies allow us not only to extend the study of problems in each of the specific fields involved, but also to give a broader theoretical understanding of the results of linguistic studies, to bring the problems of language and text into a theoretical and systemic context of problems of optimization of human communicative behaviour.

Naturally, methodological problems of interdisciplinary research are in close contact with the problems of constructing synergetic models and with general problems of quantitative and systemic models of the functioning of language which were touched upon earlier.

4. Ways of development of quantitative linguistics

The general course of development of science can be imagined as an *alternation of paradigms* (according to Thomas S.Kuhn). By alternation of paradigms we mean that a new theory (concept, model) entirely ousts the old one(s). In present-day quantitative linguistics the situation of paradigm alternation can probably occur only in a narrower sense and in a few individual cases, e.g. if Zipf's law is entirely denied and an entirely different approach to the modelling of the statistical structure of the text is proposed (e.g., the attempt made by Martynenko 1988). Somehow, the concept of paradigm alternation is related to the viewpoint of *decision theory* in which an investigation is seen as a mechanism for making a decision between competing scientific hypotheses. It is noteworthy that like most problems, linguistic ones when solved by mathematical modelling, give a set of alternatives. Very often different functions can serve as mathematical models for the phenomena under investigation, and the preference for this or that model (if they correspond equally to the empirical data from the formal point of view) depends on the content analysis of the concrete problem. The main thing is the legitimacy of the system of underlying postulates applied to the phenomenon in question. What is required is the analysis of the process or phenomenon in their logic or physical essence with the view of forming an adequate idea of this phenomenon and its use in the construction of the desired model or its interpretation.

In a number of cases the development of concepts in quantitative linguistics can be characterized as "the modification of prior belief" (cf. McPherson 1990: 155). This corresponds to the Bayesian viewpoint of the progress of science towards perfect understanding by a sequence of steps, each of which is aimed at reducing uncertainty. Mostly it happens according to the *correspondence principle* to the effect that old models or theories will not be discarded, but will be viewed as limiting or particular cases of the new theory.

As a classical example illustrating the development according to the correspondence principle, we could present the fate of Zipf's law (in the form of a function analytically describing the rank and spectral distributions of words). Originally, the function for rank distribution had the following form:

(1) $$F_i = Ci^{-1}$$

where F_i denotes the frequency of the word with rank i, and C is a constant. Thereafter, the function has been repeatedly modified. First of all, it was specified by G.K.Zipf himself:

(2) $$F_i = Ci^{-\gamma}$$

and later by B.Mandelbrot:

(3) $$F_i = C(i + B)^{-\gamma}$$

where C, B and γ are constants. It is obvious that if $\gamma = 1$ and $B = 0$, then formulas (2) and (3) will turn into the original formula (1), i.e. formula (1) can be considered a particular case of the new formulas (2) and (3). In reality this means that formula (1) indicates a strictly inverse proportionality between the variables, formula (2) allows a certain variability in the form of the dependence, and formula (3) takes into account the possible deviation at the beginning of the distribution. It should be noted that formula (3) was originally derived by B. Mandelbrot from some theoretical considerations about the rank-size relation (the basic idea in Mandelbrot's formulation was "minimization under constraint"; cf. Rapoport 1982: 6).

As is known, Zipf's law was modified later in order to be useful in those cases when there was a deviation from linearity (in bilogarithmic coordinates) both at the beginning and at the end of the rank distribution of words. The authors of the most popular modifications were J.Woronczak (1967), P.M. Alekseev (1978) and Ju.K.Krylov (1982). So, for instance, Alekseev's formula (which he called "the fourth approximation" to Zipf's law) is the following:

(4) $$F_i = Ci^{-(\gamma + \phi \ln i)}$$

where C, γ and ϕ are constants. This formula includes the main forms of Zipf's function as particular cases: if $\phi = 0$, we get formula (2), and if $\phi = 0$ and $\gamma = 1$, we get formula (1).

Several more examples could be presented concerning the development of quantitative-linguistic models according to the principle of correspondence. We can mention here the model of Waring-Herdan which describes the frequency distribution of words, being a generalization of an earlier model suggested by

G.U. Yule; some quantitative models of polysemy (Polikarpov 1987); models of the growth and evolution of vocabulary (Arapov & Cherc 1983; Tuldava 1987, chap. 4.2), etc.

It also happens that several competing models express different aspects of the phenomenon under study and, although contradictory and mutually exclusive, they are nevertheless indispensable to give a complete description of experience. The appearance of such a situation can be explained by the well-known *concept of complementarity*, introduced by Niels Bohr.

As an example we shall discuss the models of Zipf, Waring-Herdan and Carroll used for the description and explanation of the frequency distribution of words.

Zipf's model (the power function) points at a simple allometric dependence between the frequency (F) and the number of words with the given frequency (m_F):

$$\frac{dm_F/m_F}{dF/F} = b \ (const.). \tag{5}$$

Allometric dependence means "harmony" in the relative growth (or decrease) of the variables, and structural affinity with some other important laws (or models) known in quantitative linguistics, such as Menzerath-Altmann's law.

Waring-Herdan's model expresses the dependence between F and m_F from the viewpoint of the law of "uneven transition" (Muller 1976) meaning constant decrease of the ratio m_i/m_{i+1} as the result of some kind of equilibrium (balance) observed in the process of speech generation. The model of lognormal distribution, proposed by J.B. Carroll (1967), makes it possible to regard speech generation as a probability process, which can serve as a basis for certain conclusions about the nature of language.

The investigator can view these solutions as a complex (in the sense of complementarity) or prefer one or several of them depending on his theoretical or practical considerations.

Thus, the basic principles according to which quantitative linguistics will develop in the foreseeable future are, in our opinion, the principles of decision theory (competition between hypotheses), correspondence (development on the basis of previously conceived ideas) and complementarity (integration of different hypotheses). In fact, to many investigators these principles and approaches by themselves are complementary rather than competitive.

New ideas and approaches are necessary which take into account the amount and depth of knowledge already at our disposal. Extensive application of

computers in the practice of linguistic studies by quantitative methods is also an important instrument and stimulus for the development of quantitative linguistics.

References

Alekseev, P.M. (1978). O nelinejnych formulirovkach zakona Zipfa (On nonlinear formulations of Zipf's law). *Voprosy kibernetiki 41, 53-65.*

Alekseev, P.M. (1988). *Kvantitativnaja tipologija teksta (Quantitative typology of text).* Leningrad, LGPI.

Altmann, G. (1980). *Statistik für Linguisten.* Bochum, Brockmeyer.

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

Altmann, G. & Schwibbe, M.H. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen. Mit Beiträgen von W.Kaumanns, R.Köhler und J.Wilde.* Hildesheim etc., Olms.

Arapov, M.V. (1988). *Kvantitativnaja lingvistika (Quantitative linguistics).* Moscow, Nauka.

Arapov, M.V. & Chere, M.M. (1983). *Mathematische Methoden in der historischen Linguistik.* Bochum, Brockmeyer.

Boroda, M.G. (1982). Häufigkeitsstrukturen musikalischer Texte. In: Orlov Ju. K., Boroda, M.G., Nadarejšvili, I.Š. (eds.), *Sprache, Text, Kunst. (Quantitative Analysen).* Bochum, Brockmeyer, 231-262.

Boroda, M.G. & Dolinskij, V.A. (1988). Problems of quantitative text analysis. *Glottometrika 9, 135-145.*

Carroll, J.B. (1967). On sampling from a lognormal model of word-frequency distribution. In: Kučera, H. & Francis, W.N. *Computational Analysis of Present-day American English.* Providence, R.I., Brown University Press, 406-424.

Embleton, S.M. (1986). *Statistics in historical linguistics.* Bochum, Brockmeyer.

Embleton, S. (1993). Multidimensional scaling as a dialectometrical technique: outline of a research project. In: Köhler, R. & Rieger, B.B. (eds.). *Contributions to Quantitative Linguistics.* Dordrecht etc., Kluwer, 267-276.

Goebl, H. (1993). Dialectometry: a short overview of the principles and practice of quantitative classification of linguistic atlas data. In: Köhler, R. & Rieger, B.B. (eds.). *Contributions to Quantitative Linguistics.* Dordrecht etc., Kluwer, 277-315.

Grotjahn, P. (1979). *Linguistische und statistische Methoden in Metrik und Textwissenschaft.* Bochum, Brockmeyer.

Hammerl, R. & Sambor, J. (1993). Synergetic studies in Polish, In: Köhler, R. & Rieger, B.B. (eds.). *Contributions to Quantitative Linguistics.* Dordrecht etc., Kluwer, 331-359.

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

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

Köhler, R. (1993). Synergetic linguistics. In: Köhler, R & Rieger, B.B. (eds.). *Contributions to Quantitative Linguistics.* Dordrecht etc., Kluwer, 41-51.

Köhler, R. & Rieger, B.B. (eds.) (1993). *Contributions to Quantitative Linguistics. Proceedings of the First International Conference on Quantitative Linguistics, QUALICO, Trier 1991.* Dordrecht-Boston-London, Kluwer.

Králík, J. (1990). On some special models in quantitative linguistics. *Prague Studies in Mathematical Linguistics 10, 85-105.*

Krylov, Ju. K. (1982). Ob odnoj paradigme lingvostatističeskich raspredelenij (A paradigm of linguostatistical distributions). *Acta et Commentationes Universitatis Tartuensis 628, 80-102.*

Lebedev, A.N. (1986). Nejrofiziologičeskie predely pamjati čeloveka i bogatstva jego leksiki (Neurophysiological limitations of man's memory and of his vocabulary). *Acta et Commentationes Universitatis Tartuensis 745, 95-108.*

Lewis, C.I. (1966). Philosophy. In: *Encyclopedia Americana. International Edition. Vol. 21, 769-771.*

Martynenko G.Ja. (1988). *Osnovy stilemetrii (The foundations of stylometrics).* Leningrad University Press.

McPherson, G. (1990). *Statistics in scientific investigation.* New York, Springer.

Muller, C. (1976). Some recent contributions to statistical linguistics. *Statistical Methods in Linguistics 1976, 136-147.*

Orlov, Ju. K. (1982). Ein Modell der Häufigkeitsstruktur des Vokabulars. In: Orlov, Ju. K., Boroda, M.G., Nadarejšvili, I.Š. (eds.). *Sprache, Text, Kunst. (Quantitative Analysen).* Bochum, Brockmeyer, 118-192.

Perebejnos, V.I. (ed.) (1990). *Statističeskaja leksikografija i učebnyj process (Statistical lexicography and the process of learning).* Kiev, Pedagogical Institute Press.

Piotrowski, R.G. (1979). *Inženernaja lingvistika i teorija jazyka (Computational linguistics and the theory of language).* Leningrad, Nauka.

Piotrowski, R., Lesohin, M., Lukjanenkov, K. (1990). *Introduction of elements of mathematics to linguistics.* Bochum, Brockmeyer.

Polikarpov, A.A. (1987). Polisemija: sistemno-kvantitativnye aspekty (Polysemy: systemic-quantitative aspects). *Acta et Commentationes Universitatis Tartuensis 774, 135-154.*

Polikarpov, A.A. (1993). A model of the word life cycle. In: Köhler, R. & Rieger, B.B. (eds.). *Contributions to Quantitative Linguistics.* Dordrecht etc., Kluwer, 53-63.

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

Saukkonen, P. (ed.) (1992). *What is Language Synergetics? Seminar on the International Language Synergetics Project. Oulu, 5-6 October 1990.* Oulu University Press (Acta Universitatis Ouluensis, series B, Humaniora 16).

Seppänen, J. (1992). Synergy, emergence and complexity in mind, language and culture. A systems view of cognitive, linguistic and cultural evolution. In: Saukkonen, P. (ed.). *What is Language Synergetics?* Oulu University Press, 118-159.

Silnitsky, G. (1993). Correlational system of verbal features in English and German. In: Köhler, R. & Rieger, B.B. (eds.). *Contributions to Quantitative Linguistics.* Dordrecht etc., Kluwer, 409-420.

Těšitelová, M. et al. (1985). *Kvantitativní charakteristiky současné češtiny (Quantitative characteristics of the present-day Czech language).* Praha, Academia.

Tuldava.J. (1980-1982). O teoretiko-metodologičeskich osnovach kvantitativno-sistemnogo analiza leksiki 1-3 (On the theoretical-methodological premises of the quantitative-systemic analysis of lexis). *Acta et Commentationes Universitatis Tartuensis 544, 143-158; 585, 114-133; 619, 123-143.*

Tuldava, J. (1987). *Problemy i metody kvantitativno-sistemnogo issledovanija leksiki (Problems and methods of the quantitative-systemic investigation of vocabulary).* Tallinn, Valgus.

Woronczak, J. (1967). On an attempt to generalize Mandelbrot's distribution. In: *To Honor Roman Jakobson, vol. 3.* The Hague, Mouton, 2254-2268.

Zörnig, P. & Boroda, M. (1992). The Zipf-Mandelbrot law and the interdependencies between frequency structure and frequency distribution in coherent texts. *Glottometrika 13, 205-218.*

2

On Causal Relations in Language

The present article deals with the problem of probabilistic-statistical modelling of causal relations in language, using concrete illustrative material as a basis for the analysis. The possibilities of the establishment and measurement of causal dependence and interdependence are examined by applying the methods of correlation and regression analysis.

1. The probabilistic conception of causality

The definition of causal relations between phenomena is always connected with some conceptual system. The author of the present paper proceeds from the probabilistic conception of causality, according to which "causality is something that may be found to a greater or smaller degree and not only exist or not exist" (Wiener 1956). Using this approach it is assumed that some phenomenon may be the cause of some other if the appearance of the first (X) with a high degree of probability is followed by the appearance of the other (Y), while it is stated that, in symbolic form, $P(X) > 0$ and

$$P(Y|X) > P(Y),$$

i.e. "the appearance of Y, on condition that X appeared, is more probable than the appearance of Y without X". In this way causality is expressed as *conditional probability*, as the dependence between phenomena which are characterized by the expression "on condition that".

In the case of multiple causes it turns out that

$$P(Y|X_1, X_2, \ldots) > P(Y|X_1),$$

i.e. the addition of new factors increases the probability of the appearance of Y. At the same time X_1 may happen to be a "false" cause if $P(Y_1|X_1,X_2) = P(Y|X_2)$. (For more details on causal interpretation of statistical relationships see, e.g. Nowak 1965; Suppes 1970).

The dependence of Y on the single X may be established and measured with the help of the methods of simple correlation and regression. In the case of

multiple causes the total dependence is measured by means of the methods of multiple correlation and regression. A false, and also an indirect, cause may be revealed with the help of the method of partial correlation.

According to the probabilistic approach, deterministic causality is naturally included in the probabilistic scheme of causality as a particular (extreme) case which has probability equal to 1 (or 0), symbolically $P(Y|X) = 1$ (or 0). Because of the fact that chance and secondary factors cannot be excluded from the interrelations between phenomena, the dependence necessarily acquires a probabilistic (stochastic) character.

Causal analysis in its probabilistic treatment can be considered one of the most important subsidiary methods for the description and explanation of complex systems. It calls for the use of statistical methods which, when applied in linguistics, may help discover new heuristic possibilities for the investigation of interrelations and dependences in language.

2. Aspects of causal relations

Causal relation can be said to be present when one phenomenon in some way influences the other, viz.:

(i) it gives rise or birth to the other (usually, it is a relation between external cause and internal effect, or vice versa);

(ii) determines the existence of the other (for instance, the phonemic structure of the language determines its morphologic structure; of course, the dependence may be mutual);

(iii) causes changes in the other (e.g., the strengthening of the dynamic stress causes the reduction of the unstressed part of the word).

These main aspects of causal relations may often appear in combination. In all cases causality may have different forms. For example, X may be a sufficient and necessary condition for Y, symbolically $P(Y|X) = 1$ and $P(Y|\bar{X})$ (Y always takes place if X takes place, and Y never takes place if X does not take place). However, in practice, also in language, we usually see a less strong relationship, for example: X is a sufficient but not necessary condition, i.e. $P(Y|X) = 1$ and $0 < P(X|Y) < 1$, if it happens that in addition to X there are also other causes which may influence Y. Consequently, the appearance of X, in case Y takes place, has a probabilistic character. The influence of a strong dynamic stress on the reduction of the unstressed parts of a word can be given as an example. Reduction may be caused by other reasons (the speed of speech, the frequency of use, etc.) besides the strong dynamic stress, but the latter seems to be a sufficient condition for reduction.

With regard to the character of the interrelations between quantity and quality, the following dependences may be of interest:

"quantity - quality" (for example, the frequency of the word and its word-formational activity),

"quality - quantity" (the build-up of the system and its statistical characteristics),

"quantity - quantity" (the number of phonemes in a language and the length of the word).

From another point of view causes (conditions) may be exterior or interior (extra- and intralinguistic), direct or indirect, principal or additional (subsidiary), diachronic or synchronic. All these types of causality appear in language, as well as in other fields of reality.

In a diachronic analysis of causality, the fact that causal relations have a tendency not to appear instantly, but gradually, is taken into consideration. But it is not always possible to identify causal relations with the succession of events in time. The character of these relations is neither simple nor clear. The effect (consequence) is usually influenced by a whole complex of multiple interwoven causes. The effect itself may simultaneously influence its cause as a result of which the cause and the effect may be replaced by each other in the evolution of things. A complex "cause - effect" may become the cause of the successive process, etc. For example, the frequency of the use of the word may influence the semantic scope (polysemy) of the word which in its turn influences the frequency of the given word and further both factors together may cause changes in the morphemic or phonemic structure of the word.

With these considerations in mind we shall discuss the interrelations between some linguistic phenomena and give some examples of the application of statistical methods in the establishment and measurement of causal relations in language.

3. Simple linear correlation

Classical methods of correlation and regression analysis are the principal statistical tools for the analysis of causal relations. However, correlation as a formal statistical concept does not in itself reveal the causal character of relations. Correlation in itself gives only an assessment of the intensity of association (or joint occurrence) but the formulation of the question about causality, its existence and direction (cause → effect) must be non-statistical, i.e. professional-

logical. Depending upon the aims and tasks of the investigation, linear or non-linear, simple, partial, and multiple correlation analyses are used. We begin our treatment of causal relations between linguistic phenomena from the least complicated case - the method of simple linear correlation.

As an illustration we shall analyze the multivariate distribution of some quantitative characteristics of words on the basis of a frequency dictionary of lexemes of authors' monologues (i.e. non-conversational material) taken from contemporary Estonian prose fiction (Kaasik et al. 1977; Tuldava 1977). The material is represented by a corpus of 20 samples - 5,000 words each - from 20 authors, in total 100,000 word occurrences in the whole corpus. A fragment of the frequency dictionary - the 1,200 most frequent lexemes (the "basic vocabulary") covering 75% of the text - has been examined with regard to the following quantitative features: frequency of occurrence, length, semantic scope (polysemy, or the number of meanings of a word, determined on the basis of a one-language dictionary)[1], and the age of the word. Because of the large number of individual observations (1,200 words) and in the interest of better presentation of the material, it is convenient to group the words into frequency zones of equal size, 100 words in each group, and to calculate the mean values of the features under examination. As the words in the frequency dictionary are arranged according to decreasing frequencies and increasing ranks, so also the frequency zones are arranged as follows: zone No. 1 consists of words with individual ranks 1 to 100, zone No. 2 of words with individual ranks 101 to 200, etc. In this way we get a sort of gradation of lexical groups according to significance. The distribution of other quantitative features will be examined in relation to the frequency zones.

In terms of the above grouping of words, we can choose between three variants of statistical characteristics of the feature "frequency": mean frequency (F) of words in the given frequency zone; mean individual rank (M) and, if we mark each zone with a number, the rank of the zone (Z). We shall begin with the analysis of interrelations between mean frequency (F) and the following features:

- mean length of a word in syllables (L);
- mean number of meanings (polysemy) (P);
- mean age of words in the given frequency zone (A), expressed by a coefficient denoting the ratio of ancient words (from the period before A.D. 1200)

[1] Semantic scope (meaning complexity) in this sense is sometimes called "polylexy" (Köhler 1986).

in the given frequency zone (by analogy with the method proposed by Arapov and Cherc 1983).

Table 1 represents the initial data of the investigation.[2]

Table 1
Distribution of quantitative-linguistic characteristics on the basis of frequency zones

Freq. zone Z	Individual ranks i	Mean rank M	Mean freq. F	Word length L	Age A	Polysemy P
1	1-100	50	413.58	1.77	0.91	9.46
2	101-200	150	89.26	1.85	0.77	6.03
3	201-300	250	55.01	1.93	0.70	4.55
4	301-400	350	37.98	2.32	0.66	4.50
5	401-500	450	29.34	2.30	0.60	3.66
6	501-600	550	24.31	2.16	0.54	3.32
7	601-700	650	20.38	2.48	0.46	3.43
8	701-800	750	17.47	2.40	0.44	2.75
9	801-900	850	15.49	2.37	0.43	2.80
10	901-1000	950	13.40	2.41	0.37	2.58
11	1001-1100	1050	11.96	2.66	0.33	2.50
12	1101-1200	1150	10.69	2.62	0.35	2.32
Total (Σx)		7200	738.87	27.27	6.56	47.90
Mean ($\bar{x}$)		600	61.57	2.2725	0.5467	3.9917
Standard deviation (s_x)		360.55	113.15	0.2896	0.1835	2.0333

The correlation between the features will be calculated by means of the method of linear correlation using Pearson's product-moment formula:

[2] For some recent studies on the problem of interrelations between quantitative-linguistic features the reader is referred to Altmann et al. 1982; Rothe 1983; Köhler 1986; Zörnig et al. 1990; Hammerl 1992; Hammerl & Rogaliñska 1992; Miyajima 1992.

(1) $$r_{yx} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y}$$

where r_{yx} is the coefficient of correlation between the comparable features (X and Y); $\bar{x}$ and $\bar{y}$ are mean values, s_x and s_y the corresponding standard deviations, n is the number of observations (in the given case 12 pairs are being compared, consequently, n = 12).

The calculation of linear correlation between F (mean frequency) and the features L (word length), A (age) and P (polysemy) yields the result:

$$r_{LF} = -0.6834; \qquad r_{AF} = 0.7523; \qquad r_{PF} = 0.9330.$$

The interrelations between the remaining features are expressed by the following coefficients:

$$r_{PL} = -0.8419; \qquad r_{PA} = 0.9242; \qquad r_{LA} = -0.9260.$$

All the obtained correlation coefficients (except r_{LF}) are statistically significant at the 0.01 level of significance; for the number of degrees of freedom $df = n - 2 = 12 - 2 = 10$, the critical value of the coefficient is 0.708 in a two-tailed (non-directional) test.

However, the problem of correctness of the calculations arises. From Table 1 it is clear that the values of F (mean frequency) for the first 100 most frequent words are disproportionately large and can therefore distort the results of the analysis of linear correlation. We could exclude the values of F for the first zone from our experiment but then there would be considerable loss of information as the words of the first zone cover about 40% of the given text. Therefore, it is more suitable to use the mean rank (M) or, even simpler, the rank of the frequency zone (Z) as the representative of the feature "frequency of occurrence" (exactly the same value of correlation coefficients will be obtained for M and Z, since the intervals on each variable are uniform). The new calculation gives the following results:

$$r_{LM} = r_{LZ} = 0.9085; \quad r_{AM} = r_{AZ} = -0.9714; \qquad r_{PM} = r_{PZ} = -0.8384.$$

The interrelations between P, L and A remain unchanged. The results of the correlation analysis for the whole system of features is depicted in Fig. 1.

In the framework of the present investigation the question about the direction of correlation arises, i.e. the question of which characteristic in every pair should be considered the cause and which the effect. Here, it should be taken into consideration that although the causal influence, as a rule, is manifested in the form of a correlation (linear or non-linear), it does not follow that the obverse statement will hold - that every correlation conceals a causal relation.

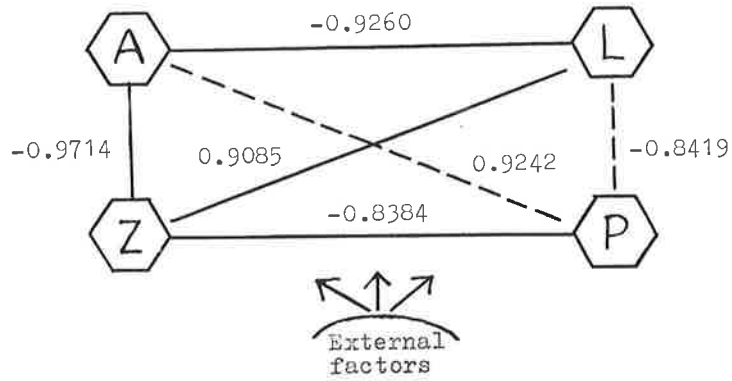

Figure 1. Correlations between Z - frequency zone, P - polysemy, A - age, L - length of the word

As already mentioned, the direction of causal relation is clarified by an investigator's application of professional knowledge and logic. It may be supposed that in this study, in all the cases, the dependence is mutual, but in some cases, coming from synchronic or diachronic viewpoint, it is possible to distinguish the dominant direction of causality. It seems likely that frequency has a greater influence on the length of the word than vice versa, at least in diachrony (frequent use of the word very often results in the shortening of the word). Logical reasoning also suggests that in many cases the age of the word to a consider- able degree depends on the frequency of usage, namely because the words of high frequency, denoting the main fields of people's activity, are more stable and "alive" and so "live longer" (verbs denoting going and coming, taking and giving, seeing and hearing; nouns denoting kinship and nature; adjectives denoting colour and size; pronouns, etc.). On the whole, frequent use of a word gives support to ist "viability". Frequent use of a word in various contexts may lead to the accumulation of new meanings (sometimes to homonymy) and back to an increase in frequency; etc.

In such discussions one should not forget that the observed characteristics, as well as all other phenomena, may represent in themselves results of multiple causes all operating together and simultaneously, including the influence of "concealed" factors, occasional circumstances, all kinds of internal (intralinguistic or psychological) and external forces, which are either unknown or difficult to detect. Such factors, which are not implicitly considered in an investigation of dependences, are usually called "disturbances", and they are reflected in deviations from the expected values.

At the disposal of the investigator are some statistical methods which formally help find the dominating direction of influence, for example, the methods of implication and regression (cf. Silnitsky et al. 1990: 33-34). These methods can be illustrated by means of a 2 x 2 contingency table. In a very simple test the bivariate distribution of frequency and length of nouns in the selected Estonian texts was examined (cf. Kaasik et al. 1977: 128-130); see Table 2.

Table 2

	Y_1	Y_2	Total
X_1	50 (a)	12 (b)	62 ($a+b$)
X_2	142 (c)	199 (d)	341 ($c+d$)
Total	192 ($a+c$)	211 ($b+d$)	403 ($a+b+c+d$)

Here, X denotes frequency and Y word length; X_1 nouns with high frequency ($F \geq 50$) and X_2 nouns with lower frequency ($10 \leq F < 50$; text length ~ 30,000 nouns); Y_1 short nouns (less than 5 phonemes), Y_2 long nouns (5 or more phonemes).

First, in order to test the significance of the interdependence between X and Y, we shall apply the chi-square test using the formula

$$\chi^2 = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}. \tag{2}$$

The calculation yields the result $X_2 = 31.99$ which is statistically significant at least at the 0.001 level (the critical value being $\chi^2 = 10.83$ for one degree of freedom). Consequently, there is significant interdependence between the frequency of occurrence and the length of the word (noun).

Then the test for implication is calculated using the following formulas:

$$I_{y/x} = \frac{a}{a + b} \quad \textit{and} \tag{3a}$$

$$I_{x/y} = \frac{a}{a + c}. \tag{3b}$$

Formula (3a) expresses the "implication" of Y under the influence of X, and formula (3b) shows the opposite direction of influence. In our example:

$$I_{y/x} = \frac{50}{50 + 12} = 0.8065 \quad \textit{and} \quad I_{x/y} = \frac{50}{50 + 142} = 0.2604.$$

i.e. the share of short and frequent words (cell a in the above contingency table) among the frequent words (row total $a+b$) is 0.8065, or 80.65 per cent, but the share of short and frequent words (cell a) among the short words (column total $a+c$) is only 0.2604, or 26.04 per cent. One may conclude that the dependence of word length (Y) on frequency (X) is stronger than vice versa. Short nouns are not especially frequent, but frequent nouns are almost always short.

Similarly, the comparison of the regression coefficients $b_{y/x}$ and $b_{x/y}$ (in the case of the same dimension of units) might shed light on the tendency or dominating direction of causal relation. The formulas for 2x2 tables are

$$b_{y/x} = \frac{ad - bc}{(a + +b)(c + d)} \quad \textit{and} \tag{4a}$$

$$b_{x/y} = \frac{ad - bc}{(a + c)(b + d)}. \tag{4b}$$

In the first case (4a) the influence of X on Y will be calculated, in the second case (4b) the influence of Y on X. In our example $b_{y/x} = 0.3900$ and $b_{x/y} = 0.2035$.

This kind of analysis, undoubtedly, allows us to reveal some essential sides of the inner structure of the bivariate distribution of characteristics, but the disclosure of the complicated causal mechanism should be made on the basis of qualitative professional interpretation. As mentioned above, we are inclined to consider the feature "frequency of occurrence" the dominating factor in its relation to the feature "length of the word", especially from the diachronic point of view. The frequency of occurrence as a characteristic feature of lexical items certainly expresses the degree of importance and significance of a given word

for the native speaker of the language. So the frequency of usage may embody some concealed factor, the factor of social and historical experience, which was not considered in our study of correlation analysis. We would subscribe to the opinion that "frequency value of the word is as a rule a most reliable and objective factor indicating the relative value of the word in language in general and conditioning the grammatical and lexical valency of the word. The frequency value of the word alone is in many cases sufficient to judge of its structural, stylistic, semantic and etymological peculiarity. If the word has a high frequency, it is in all probability monomorphic, simple, polysemantic and stylistically neutral. Etymologically it is likely to be native or to belong to early borrowings" (Ginzburg et al. 1966: 238).

4. Partial correlation

A further task in the analysis of causal relations is the specification of correlative ties in the system of quantitative characteristics. In the previous section the assessment of simple linear correlation between several features of the word was made. It must be taken into consideration that in case two features, e.g. Y and X, correlate with each other, the value of the coefficient of simple correlation may be influenced by other features which are connected with Y and X. For the objective assessment of correlative dependence it is necessary to eliminate in some way the outside influences, otherwise the analysis of causal relations between the two features will remain simplistic. For the measurement of the "pure" correlation between Y and X_1, when the influence of another factor X_2 is eliminated (more precisely: remains constant), the partial correlation can be calculated using the formula:

$$(5) \qquad r_{y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1 - r_{y2}^2)(1 - r_{12}^2)}},$$

where $r_{y1.2}$ is the coefficient of partial correlation between X and X_1 after eliminating the effect of X_2. As with the simple linear correlation coefficient, the partial correlation coefficient may vary between -1 and +1. A value of zero implies no correlation and increasing magnitude implies an increasing level of correlation. The theoretical criterion of statistical significance to be used is the t- test with the formula (Storm 1972: 243):

$$(6) \qquad t = \frac{r_{y1.2}\sqrt{n - k}}{\sqrt{1 - r_{y1.2}^2}}$$

where n denotes the number of observations and k the number of characteristics under examination.

For instance, we might be interested in establishing the real dependence of polysemy on other features in the system. We ascertained that with regard to simple linear correlation the feature "polysemy" (P) was closely connected with the feature "frequency zone" (Z) and simultaneously with the feature "length of the word" (L) - with coefficient values r_{PZ} = -0.8384 and r_{PL} = -0.8419, correspondingly (see Fig. 1). The length of the word in turn depends on the frequency of the word (the frequency zone): r_{LZ} = 0.9085. The question arises as to how strong the connection between polysemy (P) and the length of the word (L) is if we eliminate the influence of frequency (Z), the latter being an important factor influencing both polysemy and word length. It is possible that polysemy (P) and word length (L) in reality are connected indirectly as a result of the so-called effect of cohesion (graphically, see Fig. 2). In Fig. 2 the correlation between P and L has been caused by the medium of Z.

We shall try to calculate the correlation between P and L by eliminating the influence of Z with the help of the method of partial correlation. The initial data needed for the calculation are (Fig. 1):

$$r_{y1} = r_{PL} = -0.8419; \qquad r_{y2} = r_{PZ} = -0.8384; \qquad r_{12} = r_{LZ} = 0.9085.$$

Using formula (5) we get

$$r_{y1.2} = r_{PL.Z} = \frac{-0.8419 - (-0.8384 \cdot 0.9085)}{\sqrt{(1 - 0.8384^2)(1 - 0.9085^2)}} = -0.3522.$$

The calculation of the t-test yields the result t = 1.1289 which does not reach the critical value at the 0.05 level for 9 degrees of freedom ($t_{0.05;9}$ = 2.26). Thus, in our system of features the partial linear correlation between polysemy and word length (by eliminating frequency) tends to be statistically not significant. Of course, this may be so due to the small sample size (n = 12).

However, as G.McPherson (1990: 506) has pointed out, it tends not only to be the size of the partial correlation coefficient which is of interest. The extent to which the partial correlation coefficient differs from the coefficient of simple

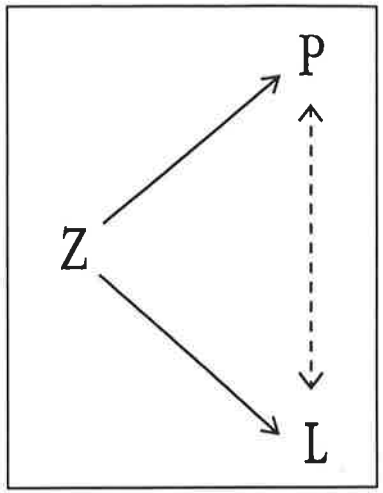

Figure 2. The effect of cohesion.

correlation is equally important. Of primary interest is whether the partial correlation coefficient is substantially different from the correlation coefficient. In comparison with simple linear correlation between polysemy and word length (r_{PL} = -0.8419) the coefficient of partial correlation, after eliminating the influence of frequency, is indeed considerably smaller ($r_{PL.Z}$ = -0.3522). We might draw the conclusion that real correlation between polysemy and word length is spurious. Unfortunately, there is no formal test available to establish when a partial correlation is significantly different from the simple correlation. Particularly when sample sizes are small, findings should be treated with caution (cf. McPherson 1990: 506). In any case, the difference revealed between the correlation coefficients cautions against overrating the strength of the simple linear correlation between polysemy and word length.

Using the procedure outlined above, that is, calculating the partial correlation between features by eliminating one outside factor, we shall gauge the dependence of polysemy (P) on the age of the word (A) after eliminating the influence of frequency (Z). The coefficient of simple linear correlation between features P and A was equal to 0.9242 (Fig. 1). Eliminating Z, we get $r_{PA.Z}$ = 0.8454. The conclusion is that on the basis of the data of our study, the correlation between polysemy and age (eliminating frequency) is significant (t = 4.7482) even when the influence of frequency is eliminated. Apparently, the development of polysemy in a word, as a rule, cannot be momentary, but needs time. If we eliminate the influence of age (A), the partial correlation between polysemy (P) and frequency (Z) is r_{PZA} = -0.6547, that is, still statistically significant (t = 2.60).

Generalizing the partial correlation to any number of factors, the formula can be expressed in the following compact form (Förster & Rönz 1979: Chap. 4.5):

$$(7) \qquad r_{y1.2\ldots m} = \frac{r_{y1.3\ldots m} - r_{y2.3\ldots m} \cdot r_{12.3\ldots m}}{\sqrt{(1 - r^2_{y2.3\ldots m})(1 - r^2_{12.3\ldots m})}} .$$

Only as an example. we shall calculate the partial correlation between polysemy (P) and word length (L), eliminating the combined influence of frequency (Z) and age (A). From formula (7) it can be seen that we have to calculate the auxiliary coefficients (using formula (6)):

$$r_{y1.3} = r_{PL.A} = -0.0965;$$
$$r_{y2.3} = r_{PZ.A} = -0.6547;$$
$$r_{12.3} = r_{LZ.A} = 0.1002.$$

Then we calculate the coefficient of linear partial correlation by eliminating two factors, using formula (7):

$$r_{y1.23} = r_{PL.ZA} = \frac{-0.0965 - (-0.6547)0.1002}{\sqrt{(1 - 0.6547^2)(1 - 0.1002^2)}} = -0.0411.$$

As can be seen, the elimination of A, in addition to Z, diminishes the role of L (word length) in its relation to P (polysemy) even more.

5. Multiple correlation

When investigating causal relations between linguistic phenomena, we may state that in the majority of cases any linguistic phenomenon can be considered as an effect (consequence) of the total influence of different causes (not to forget the external and concealed forces that lie behind the formal quantitative factors). From this, the task of determining the causal relations of one phenomenon with a set of many other quantitative characteristics arises. Simultaneous investigation of the correlation of one feature with a number of other features can be carried out by means of the method of multiple correlation.

For the case of the dependence of the resultant variable (Y) on two independent factors (X_1 and X_2) the formula for linear multiple correlation has the form:

$$(8) \qquad R_{y.12} = \sqrt{\frac{r^2_{y1} + r^2_{y2} - 2r_{y1}r_{y2}r_{12}}{1 - r^2_{12}}}$$

where $R_{y.12}$ is the coefficient of linear multiple correlation.

In order to calculate $R_{y.12}$ it is necessary to know the coefficients of simple linear correlation between all three variables. The value of the coefficient of

multiple correlation ranges from 0 to 1 and it may not be numerically smaller than any of the coefficients of simple linear correlation which produce it. With the help of the coefficient of multiple correlation we cannot draw any conclusion concerning the character of interrelation, i.e. either of positive or negative correlation. Only if all the coefficients of simple correlation have the same sign (plus or minus) is it possible to attribute this sign also to the coefficient of multiple correlation.

As an example let us look at the dependence of polysemy (P) on the total influence of the frequency (Z) and the length of the word (L). The coefficients of simple linear correlation are (Fig. 1):

$$r_{y1} = r_{PZ} = -0.8384; \qquad r_{y2} = r_{PL} = -0.8419; \qquad r_{12} = r_{ZL} = 0.9085.$$

In application to our experiment, the coefficient of linear multiple correlation equals

$$R_{P.ZL} = \sqrt{\frac{0.8384^2 + 0.8419^2 - 2(-0.8384)(-08419)(0.9085)}{1 - 0.9085^2}} = 0.8601.$$

If some coefficients of partial correlation are known, the formula has the form:

$$R_{y.12} = \sqrt{1 - (1 - r_{y1}^2)(1 - r_{y2.1}^2)}. \tag{9}$$

For example, if $r_{y1} = r_{PZ} = -0.8384$ and $r_{y2.1} = r_{PL.Z} = -0.3527$, the calculation of the coefficient of linear multiple correlation gives the result

$$R_{P.ZL} = \sqrt{1 - (1 - 0.8384^2)(1 - 0.3527^2)} = 0.8601,$$

i.e. the same result which was obtained using formula (8).

The appropriate test to use is the F-test, which enables us to state the statistical significance of the F-value using the formula (Storm 1972: 245)

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - k}{k - 1} \tag{10}$$

where n is the number of observations and k the number of characteristics involved. The F-statistic tests the null hypothesis that the multiple correlation between the dependent variable and the total of factors (independent variables) is zero in the population.

In our case $R_{y.12} = r_{P.ZL} = 0.8601$, $n = 12$ and $k = 3$; consequently,

$$F = \frac{0.8601^2}{1 - 0.8601^2} \cdot \frac{12 - 3}{3 - 1} = 12.7926.$$

For df_1, $df_2 = (k - 1)$, $(n - k)$, i.e. 2 and 9 degrees of freedom, at the 0.01 level the critical F-value is 8.02 (for F-distribution tables see e.g. Storm 1972; Förster & Rönz 1983). As the calculated F-value (12.7926) exceeds the critical value, we can conclude that the correlation is statistically significant (at the 0.01 level).

Generalizing, the formula of the coefficient of linear multiple correlation may be presented in the following way:

$$R_{y.12...m} = \sqrt{1 - (1 - r_{y1}^2)(1 - r_{y2.1}^2)...(1 - r_{ym.12...(m-1)}^2)}. \tag{11}$$

Let us calculate the coefficient of multiple correlation between the feature P (polysemy), on the one side, and the features Z (frequency), L (word length) and A (age of the word) on the other. The coefficients calculated earlier are:

$$r_{y1} = r_{PZ} = -0.8384; \qquad r_{y2.1} = r_{PL.Z} = -0.3527; \qquad r_{y3.12} = r_{PA.ZL} = 0.8249.$$

As a result we get

$$R_{P.ZLA} = \sqrt{1 - (1 - 0.8384^2)(1 - 0.3527^2)(1 - 0.8249^2)} = 0.9575.$$

We can see that the addition of the feature A (age) increases the coefficient of multiple correlation in comparison with the coefficient calculated on the basis of the dependence between P and the total influence of F and L ($R_{P.ZL}$ = 0.8601).

As was already mentioned, the coefficient of linear multiple correlation R varies within the limits of 0 to 1. The nearer to 1, the closer the correlation. If the value of R is not statistically significant, we may state that between the de-

pendent and the predictor variables (factors) there is no correlational dependence (but there may be non-linear dependence which has to be specially checked; see the next section).

Thus, the coefficent of linear multiple correlation, which reflects the total influence of three quantitative factors on the resultant feature *P* (polysemy), was 0.9575 on the basis of the data of our experiment. This is greater than any of the coefficients of simple linear correlation between *P* and other features (r_{PZ} = -0.8384; r_{PL} = -0.8419; r_{PA} = 0.9242).

We can also calculate the so-called coefficient of determination, which is equal to the square of the coefficient of correlation. The coefficient of multiple determination R^2 = 0.95752 = 0.9168 shows, according to the traditional interpretation, the proportion of variance accounted for by the factors involved. In the given case about 92% of the variation of *P* can be explained by the total influence of the above-mentioned three factors (*Z*, *L* and *A*), but 1 - 0.9168 = 0.0832, i.e. about 8% would constitute the proportion of the variance of *P* which can be explained by the influence of other factors which were not considered.

However, this interpretation should be taken with caution. The calculation of the coefficients of linear correlation and determination was carried out assuming that there was a linear relationship between the variables and that they were jointly normally distributed. In case of non-linear correlation the application of the coefficients of linear correlation may to a certain degree distort the result of the analysis. In the following section we can see that the relations between features in our experiment deviate insignificantly from linearity and the application of linear coefficients of correlation is justified, especially if we are interested foremost in statics and not in prediction. In the theory of statistics it is accepted that even if the process under investigation has a non-linear tendency, it very often is reduced to the linear one, which in the investigation of social processes does not result in grave errors. The assumption of normality of distribution is also justified if we have a large number of observations. In our case we assume that there is an approximately normal joint distribution of the variables, although the grouping of individual observations (on the basis of frequency zones with 100 words in each zone) may play a disturbing role. The results of correlation analysis will be specified by the following regression analysis.

6. Regression

Regression is a one-direction stochastic dependence of one phenomenon on another or several other phenomena. Dealing with causal relations, we establish the dependence of the resultative variable on the predictor variables (factors) with the help of regression models. The mathematical solution is reduced to the obtainment of the equation (function) of regression.

The basic model of linear regression, i.e. linear dependence of the dependent (resultant, response) variable *Y* on the independent (predictor, explanatory) variables, or factors X_1, X_2, ..., X_m, can be expressed by the following equation:

$$X = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \epsilon \qquad (12)$$

where α is the free member (intercept) of the regression equation; β_1, β_2, etc. are the coefficients of regression; ϵ is a random disturbance ("error term"). Due to ϵ the relation expressed by equation (12) is stochastic. For the construction of tests and confidence limits it is required that ϵ is normally distributed and no exact linear relation exists between any of the independent variables (factors X_i). The correlation between the factors is called multicollinearity. For more about the assumptions and potential pitfalls in regression analysis, see Grotjahn (1992).

Bearing these reservations in mind we can interpret the regression model in the following way. The coefficients of regression β_1, β_2, etc. are directly connected with the corresponding factors X_1, X_2, etc. For example, β_1 denotes the mean value of *Y* while X_1 changes by one unit on condition that other factors remain unchanged. If all the coefficients of regression of the given regression are known, we can obtain, giving definite values to factors, the theoretical value of *Y* on the average. The number of factors included in the equation of regression should not be too large, to avoid distraction from the principal factors: it is recommended that the number of factors should be 5-6 times smaller than the number of observations. It is necessary to deal with only the most important factors which have a significant influence on the dependent variable.

In the preceding sections we ascertained that the feature *P* (polysemy) was influenced most of all by *Z* (frequency) and *A* (age of the word). Let us first look at the dependence of *P* on *Z* on the basis of simple linear regression with one predictor variable. On the basis of the information in Table 1 we shall calculate the estimates *a* and b_1 for the parameters α and β:

$$\hat{Y} = a + b_1 X_1.$$

(The disturbance ε actually appears in the form of deviations (d_i) from the expected values $\hat{Y}_i$.)

Taking $Y = P$ and $X_1 = Z$, we receive the following regression equation when applying the method of least squares:

$$\hat{P} = 7.0648 - 0.4728\ Z.$$

If we know the coefficients of linear correlation, the mean values and the standard deviations of the variables (see Table 1), we can easily calculate the parameters of the regression equation using formulas

(13*a*) $$b_{yx} = r_{yx}\frac{s_y}{s_x} \qquad \textit{and}$$

(13*b*) $$a = \bar{y} - \beta\bar{x}.$$

In our case $r_{yx} = r_{PZ} = -0.8384$, $s_y = s_P = 2.0333$, $s_x = s_Z = 3.6056$, $\bar{y} = \bar{P} = 3.9917$ and $\bar{x} = \bar{Z} = 6.5$. We get

$$b_{PZ} = -0.8384\frac{2.0333}{3.6056} = -0.4728 \qquad \textit{and}$$

$$a = 3.9917 - (-0.4728)6.5 = 7.0649.$$

(The small difference in *a* values is due to rounding error).

The conformity of the empirical data to the theoretical ones is relatively good, but under the assumption that it is possible to get better conformity if we also take the second factor *A* (age) into consideration, we shall try to construct the corresponding regression equation with two factors.[3] One caution about the use of this statistical device should be noted here. In our example the factors (predictor variables) are interrelated and therefore we have *multicollinearity*. In this case the application of the method of least squares tends to be inaccurate and the estimate of R^2 and testing results may be highly misleading. However, as R. Grotjahn (1992: 160) has noted, the suppression variables (i.e. the factors

[3] There are more complicated methods available which help scientists establish the relative importance of different sources of causality by "partitioning" causality between the predictive variables - factors (e.g. path analysis; cf. Kang & Seneta 1980; McPherson 1990, chap. 20).

which are correlated with other factors) could be used to increase the predictive power of the regression model, whenever optimization of prediction is the main objective. We take the risk and calculate the first order linear regression with two factors:

$$\hat{P} = -11.6669 + 0.5944\ Z + 21.5745\ A.$$

The values of the parameters have been calculated using the following formulas (cf. Förster & Rönz 1979, chap. 2.7; Ehrenberg 1975, chap. 15):

(14*a*) $$b_1 = \frac{r_{yx_1} - r_{yx_2}r_{x_1x_2}}{1 - r^2_{x_1x_2}} \cdot \frac{s_y}{s_{x_1}}$$

(14*b*) $$b_2 = \frac{r_{yx_2} - r_{yx_1}r_{x_1x_2}}{1 - r^2_{x_1x_2}} \cdot \frac{s_y}{s_{x_2}}$$

(14*c*) $$a = \bar{y} - b_1\bar{x_1} - b_2\bar{x_2}.$$

Table 3 shows that the regression model with two factors gives a little better result than the model with one factor (the residual sums of squares $\Sigma d_i^2 = 13.5356$ and $\Sigma d_i^2 = 3.7995$, correspondingly). For instance, if we take $Z = 1$ and $A = 0.91$ (see Table 1), the expected value is $P = -11.6669 + 0.5944 + 21.5745(0.91) = 8.56$ (observed $P = 9.46$); if $Z = 2$ and $A = 0.77$, then $\hat{P} = 6.13$ ($P = 6.03$), etc. But extrapolation (prediction) beyond the limits of observed data does not work well. For instance, if we take $Z = 20$, i.e. the frequency zone consisting of words ranking from 1901 to 2000 in the frequency dictionary, and $A = 0.14$ (experimentally checked), we get the expected value $\hat{P} = 3.45$ which is unrealistic as we really expect a value under 2 (cf. Table 1). Further, if both factors *Z* and *A* are included in the one equation, and they are highly correlated with one another and with *Y*, it is obvious that both factors will be judged to be making little contribution to the prediction of the resultant variable *Y* in the presence of the other factor.

It has been stated that in quantitative linguistics the primary and ultimate aim is to develop *explanatory models* (cf. Altmann 1980; Köhler 1986; Grotjahn 1992; and others). The question arises as to what the theoretical assumptions about the nature of the interrelation between the features under examination,

specifically between the frequency of occurrence of a word and the number of its meanings (polysemy), are. Under the assumption that their values grow in solidarity with one another (on the average) so that the ratio of relative growths remains constant, like the universal "allometric" growth model (cf. Land 1973), the above assumption can be expressed in the form of the differential equation

$$\frac{dY/Y}{dX/X} = b. \tag{15}$$

Table 3
Regression analysis:
the relation between frequency zone (Z), age of the word (A) and polysemy (P)

Z	A	P	$\hat{P}$	$d = P-\hat{P}$	d/s_d	$\hat{P}$	$d = P-\hat{P}$	d/s_d
1	0.91	9.46	6.59	2.87	2.47	8.56	0.90	1.46
2	0.77	6.03	6.12	-0.09	-0.08	6.13	-0.10	-0.16
3	0.70	4.55	5.65	-1.10	-0.95	5.22	-0.67	-1.09
4	0.66	4.50	5.17	-0.67	-0.58	4.95	-0.45	-0.73
5	0.60	3.66	4.70	-1.04	-0.89	4.25	-0.59	-0.96
6	0.54	3.32	4.23	-0.91	-0.78	3.55	-0.23	-0.37
7	0.46	3.43	3.76	-0.33	-0.28	2.42	1.01	1.64
8	0.44	2.75	3.28	-0.53	-0.46	2.58	0.17	0.28
9	0.43	2.80	2.81	-0.01	-0.01	2.96	-0.16	-0.26
10	0.37	2.58	2.34	0.24	0.21	2.26	0.32	0.52
11	0.33	2.50	1.86	0.64	0.55	1.99	0.51	0.83
12	0.35	2.32	1.39	0.93	0.80	3.02	-0.70	-1.14
	Σd_i			0.0			0.01	
	Σd_i^2			13.5356			3.7995	
	$s_d = \sqrt{\Sigma d_i/(n-2)}$			1.1634			0.6164	
Formulas			$\hat{P}$ = 7.0648 - 0.4728 Z			$\hat{P}$ = -11.6669 +0.5944Z + 21.5745A		

$\hat{P}$ - expected values; d - residuals; d/s_d - standardized residuals

This equation can be rewritten as $dY/Y = b(dX/X)$ and after intergation we get

$\log Y = A + b \log X$, where A is a constant. Taking $A = \log a$, we obtain the power function

$$Y = aX^b \tag{16}$$

where a and b are parameters.

This non-linear function has been widely used in social sciences and biology and, in fact, it can be brought into parallelism with some well-known linguistic laws such as Zipf's law of least effort (Zipf 1949) and Menzerath-Altmann's law (Altmann 1980, 1988). Of course, these laws are modelling situations somewhat different from ours and their initial assumptions can be expressed by the differential equation

$$\frac{dY/dX}{Y} = b\frac{1}{X}$$

which after all leads to the same formula (16) with $b < 0$.

Returning to our experiment we shall prove the applicability of the allometric model to the dependence of polysemy (P) on frequency (Z). Using formula (16) and taking $Y = P$ and $X = Z$ we calculate the parameter a and b by means of logarithmic linearization and the method of least squares. The calculation yields the result:

$$\hat{P} = 9.0259\, Z^{-0.5934}.$$

Inspection of Table 4 does not leave any reasonable doubt about the power function providing a satisfactory fit for the observed and expected data. The residual sum of squares is quite small($\Sigma d_i^2 = 0.5883$) in comparison with two previous experiments (13.5356 and 3.7995; see Table 3). The distribution of residuals (d_i) may be considered approximately normal, and none of the standardized (normalized) residuals (d/s_d, where $s_d = \sqrt{d_i^2/(n-2)}$; cf. Draper & Smith 1981) has a magnitude in excess of 2.0[4]. The adjusted coefficient of determination (Grotjahn 1992: 155) is $R^2 = 0.9711$, calculated according to the formula

[4] The value of the standardized residual has approximately a 95% chance of lying within the range -2 and +2 if the correct model is fitted (cf. McPherson 1990: 457).

(17) $$R^2 = 1 - a\frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where $a = (n - 1)/(n - m - 1)$, n - number of observation, m - number of factors, y_i - observed values of the dependent variable, $\bar{y}$ - their arithmetic mean, $\hat{y}_i$ - expected (estimated) values of the dependent variable. The traditional interpretation of R^2 as a measure of the proportion of variance of the dependent variable attributable to sample regression is somewhat problematic in our case (with regard to the linearized model) and therefore we treat R^2 only as a descriptive measure showing the high degree of conformity between the observed and estimated values (R^2 = 0.9711; cf. Table 4). A significance test of the null hypothesis becomes trivial. However, if we calculate the F-test by means of formula (10), we get F = 336.02. For $df_1 = m$ and $df_2 = n\text{-}m\text{-}1$, i.e. 1 and 10 degrees of freedom, at the 0.001 level the critical F-value is 21.0. Indeed, the observed F-value exceeds the critical value by an order of magnitude. We can state that the null hypothesis will be rejected and good conformity between the observed and estimated data can be declared.

The allometric model can also be used for prediction. The extrapolation up to Z = 60 gives plausible results, for instance, if we take Z = 50, we get $\hat{P}$ = 1.08; if Z = 60, $\hat{P} \approx 1$, that is, rare words are monosemantic on average.

One more point is noteworthy. Good conformity between observed and expected values is also obtained when we use the feature F (mean frequency in a given frequency zone) instead of Z. The concrete values of F will be found in Table 1. Using formula (16) we calculate

$$\hat{P} = 0.9630\ F^{0.3921}.$$

Table 4 represents the results of the comparison of observed and expected values of P in its relation to concrete frequencies (F). The residual sum of squares Σd_i^2 = 1.1641. Extrapolation to F = 1 gives $\hat{P}$ = 0.963 ≈ 1 as expected; if F = 2, $\hat{P}$ = 1.26; if F = 3, $\hat{P}$ = 1.48, etc. The extrapolation to F = 10 gives $\hat{P}$ = 2.37 which is very near to the observed value P = 2.32.

As can be seen, the allometric law appears to be a satisfactory representation of the observed relationship between polysemy and frequency of occurrence (the latter being expressed both by frequency zones and by concrete mean

frequencies). The conformity of observed and expected values is so good that there is practically no need to include any additional factor in the non-linear regression equation, e.g. of the type ("multiplicative model"):

(18) $$Y = aX_1^b X_2^c X_3^d$$

where a,b,c are parameters and e is the basis of the natural logarithms.

Concerning the causal relations between frequency and other quantitative word characteristics under review, when frequency is considered the dominating factor, we also assume that these relations are governed by the allometric law of growth analytically expressed by the differential equation (15) and the power function (16).

Table 4
Regression analysis:
the relation between frequency zone (Z), mean frequency (F)
and polysemy (P)

Z	F	P	$\hat{P}$	$d = P\text{-}\hat{P}$		$\hat{P}$	$d = P\text{-}\hat{P}$	
			d/s_d			d/s_d		
1	413.58	9.46	9.03	0.43	1.77	10.22	-0.76	-2.23
2	89.26	6.03	6.19	-0.16	-0.66	5.60	0.43	1.26
3	55.01	4.55	4.97	-0.42	-1.73	4.63	-0.08	-0.23
4	37.98	4.50	4.25	0.25	1.03	4.01	0.49	1.44
5	29.34	3.66	3.76	-0.10	-0.41	3.62	0.04	0.12
6	24.31	3.32	3.41	-0.09	-0.37	3.36	-0.04	-0.12
7	20.38	3.43	3.14	0.29	1.20	3.14	0.29	0.85
8	17.47	2.75	2.92	-0.17	-0.70	2.96	-0.21	-0.62
9	15.49	2.80	2.73	0.07	0.29	2.82	-0.02	-0.06
10	13.40	2.58	2.58	0	0	2.66	-0.08	-0.23
11	11.96	2.50	2.45	0.05	0.21	2.55	-0.05	-0.15
12	10.69	2.32	2.34	-0.02	-0.08	2.44	-0.12	-0.35
	Σd_i			0.31			-0.11	
	Σd_i^2			0.5883			1.1641	
$s_d = \sqrt{\Sigma d_i/(n-2)}$				0.2425			0.3412	
Formulas			$\hat{P} = 9.0259\ Z^{-0.5434}$			$\hat{P} = 0.9630\ F^{0.3921}$		

In other words, we assume that the quantitative features, such as age and length of the word, change in full solidarity with frequency in such a way that the ratio of the relative growth rates is constant. With this in mind, calculations were made to ascertain the form of dependence between the features age (A) and word length (L) as resultative variables and frequency zone (Z) as the independent predictor variable. We compared the data of simple linear and allometric non-linear regressions.

The results of Table 5 indicate that the fit is actually very good for linear and non-linear relations between the frequency and the age of the word: the residual sums of squares are 0.0195 and 0.0363, respectively. But the analysis of the predictive (prognostic) power evidently favours the non-linear allometric model. For instance, for $Z = 20$ the linear function would give $\hat{A} = -0.12$, which makes no sense.

Table 5
Regression analysis:
the relation between frequency zone (Z) and the age of the word (A)

Z	A	$\hat{A}$	$d = A-\hat{A}$	d/s_d	$\hat{A}$	$d = A-\hat{A}$	d/s_d
1	0.91	0.82	0.09	2.04	1.04	-0.13	-2.16
2	0.77	0.77	0	0	0.78	-0.01	-0.17
3	0.70	0.72	-0.02	-0.45	0.66	0.04	0.66
4	0.66	0.67	-0.01	-0.23	0.58	0.08	1.32
5	0.60	0.62	-0.02	-0.45	0.53	0.07	1.66
6	0.54	0.57	-0.03	-0.68	0.49	0.05	0.83
7	0.46	0.52	-0.06	-1.36	0.46	0	0
8	0.44	0.47	-0.03	-0.68	0.44	0	0
9	0.43	0.42	0.01	0.23	0.42	0.01	0.17
10	0.37	0.37	0	0	0.40	-0.03	-0.50
11	0.33	0.32	0.01	0.23	0.38	-0.05	-0.83
12	0.35	0.28	0.07	1.58	0.37	-0.02	-0.33
Σd_i			0.04			0.01	
Σd_i^2			0.0195			0.0363	
$s_d = \sqrt{\Sigma d_i/(n-2)}$			0.0442			0.0602	
Formulas		$\hat{A} = 0.8680 - 0.0494\ Z$			$\hat{A} = 1.0442\ Z^{-0.4187}$		

Quite analogously, the linear and non-linear regression equations are both satisfactory approximating functions for the relation between frequency zone (Z) and word length (L)[5]. The residual sums of squares are 0.1613 and 0.1230, respectively (see Table 6). Again, the analysis of prognostic possibilities favours the non-linear model, but the difference is not very large. Suppose we want to predict mean word length by means of frequency zone. For $Z = 20$ the linear function gives $\hat{L} = 3.26$, and the non-linear function predicts $\hat{L} = 2.81$ syllables. Actually, the mean word length in this zone is 2.61. (In a previous work we have shown that good results in the analysis of the relation between frequency and word length can be obtained by using the logarithmic function of the type $y = a + b \ln x$; cf. Tuldava 1986).

Table 6
Regression analysis:
the relation between frequency zone (Z) and word length (L)

Z	L	$\hat{L}$	$d = L-\hat{L}$	d/s_d	$\hat{L}$	$d = L-\hat{L}$	d/s_d
1	1.77	1.87	-0.10	-0.79	1.71	0.06	0.54
2	1.85	1.94	-0.09	-0.71	1.92	-0.07	-0.63
3	1.93	2.02	-0.09	-0.71	2.05	-0.12	-1.08
4	2.32	2.09	0.23	1.81	2.15	0.17	1.53
5	2.30	2.16	0.14	1.10	2.23	0.07	0.63
6	2.16	2.24	-0.08	-0.63	2.30	-0.14	-1.26
7	2.48	2.31	0.17	1.34	2.36	0.12	1.08
8	2.40	2.38	0.02	0.16	2.41	-0.01	-0.09
9	2.37	2.45	-0.08	-0.63	2.46	-0.09	-0.81
10	2.41	2.53	-0.12	-0.94	2.50	-0.09	-0.81
11	2.66	2.60	0.06	0.47	2.54	0.12	1.08
12	2.62	2.67	-0.05	0.39	2.58	0.04	0.36
Σd_i			0.01			0.06	
Σd_i^2			0.1613			0.1230	
$s_d = \sqrt{\Sigma d_i/(n-2)}$			0.1270			0.1109	
Formulas		$\hat{L} = 1.7982 + 0.0730\ Z$			$\hat{L} = 1.7139\ Z^{0.1647}$		

[5] P.Zörnig et al. (1990) have pointed out that for exact approximation one has to consider the oscillation of word length as a function of word frequency.

A point of methodological interest is that due to the non-linear allometric relation between frequency, on the one side, and the features age, length and polysemy of the word, on the other, the real interdependence between the last mentioned features may also be non-linear and probably allometric, i.e. it can be expressed by means of the power function. Of course, in some cases the linear function fits well, as we have seen above, but comparing the results of the analysis, we see that the most obvious result is the difference in prognostic power when using the linear and the non-linear devices.

7. Conclusion

On the basis of the analysis of the causal relations between linguistic phenomena (frequency, polysemy, age and length of the word) it is possible to draw the following conclusions:

1. When analyzing causal relations in the language, it is expedient to apply a probabilistic (stochastic) approach which corresponds to the principles of the *quantitative-systematic analysis* of the language (Tuldava 1993). Correlation and regression analyses are the main methods of investigation of causal relations.

2. In the functioning of language a complicated set of multiple mutually interrelated features is operating. To get an objective picture of mutual relations between phenomena, the method of partial correlation can be applied. As each linguistic phenomenon is an effect (consequence) of a total influence of a set of causes, it is suitable to use the method of multiple correlation for the assessment of the power of total influence of multiple causes on the resultative feature.

3. The problem of the determination of the direction of the connection (cause → effect) is handled in each concrete case depending on the aims and tasks of the investigation where professional-theoretical viewpoints have been taken as a basis. In this way the dominating influence of the feature "frequency of occurrence", manifesting the effect of many external and concealed factors, was revealed in the system of features under discussion.

4. The construction of regression models allows us to give a quantitative description of deep connections between phenomena, to distinguish the significant factors which determine the variation of the dependent variables and to assess

their share of influence. The application of linear correlation and regression is justified in many cases; however, for predictive purposes it is necessary to move from simple linear models to more complicated, in particular non-linear, models.

5. Summarizing the concrete results of our investigation, we come to the conception of a definite mechanism of "allometric" relation between the linguistic features under review, when the ratio between relative growths is constant. It has been stated that the allometric growth model belongs to universal laws which are closely related to the notion of optimality (cf. Rosen 1967, chap. 5; cf. also Huxley 1932). It seems that the allometric model (in the form of power function) is not misspecified for the analysis of causal relations between linguistic phenomena. The conclusion made on the basis of our study of the Estonian language is strengthened by the fact that we find the same form of relation in question in other languages as well, as has been demonstrated in the fundamental works on language synergetics by R. Köhler (1986) and others. Structural affinity with another important law of linguistic self-regulation - Menzerath-Altmann's law - can be stated as a fact of interest. All this will justify the conclusion that in the case of the allometric model we are dealing with a general law of language according to which various linguistic phenomena are interrelated in a regular harmonious (although stochastic) way as the result of the self-regulation of the language.

References

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

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

Altmann, G., Schwibbe, M., Kaumanns, W., Köhler, R. & Wilde, J. (1988). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Stuttgart, Olms.

Arapov, M.V. & Cherc, M.M. (1983). *Mathematische Methoden in der historischen Linguistik.* Bochum, Brockmeyer.

Draper, N.R. & Smith, H. (1981²). *Applied regression analysis.* New York-London-Sydney, Wiley.

Ehrenberg, A.S.C. (1975). *Data reduction.* London-New York-Sydney-Toronto, Wiley.

Förster, E. & Rönz, B. (1979). *Methoden der Korrelations- und Regressions-*

analyse. Berlin, Die Wirtschaft.

Ginzburg, R.S., Khidekel, S.S., Knyazeva, G.Y. & Sankin, A.A. (1966). *A Course of Modern English lexicology.* Moscow.

Grotjahn, R. (1992). Evaluating the adequacy of regression models: Some potential pitfalls. *Glottometrika 13, 121-172.*

Hammerl, R. (1992). Überprüfung von Modellen zur Beschreibung der Relation zwischen Länge und Frequenz und zwischen Länge und Rangnummer von Lexemen in Textwörterbüchern. In: Saukkonen, P. (ed.), *What is language synergetics? Seminar on the International Language Synergetics Project.* Oulu, University Press.

Hammerl, R. & Rogalińska, A. (1992). Über die Untersuchung mehrdimensionaler sprachlicher Relationen. In: Saukkonen, P. (ed.), *What is language synergetics? Seminar on the International Language Synergetics Project.* Oulu, University Press.

Huxley, J. (1932). *Problems of relative growth.* London, Methuen.

Kaasik, Ü., Tuldava, J., Villup, A. & Ääremaa, K. (1977). Eesti tänapäeva ilukirjandusproosa autorikõne lekseemide sagedussõnastik (A frequency dictionary of lexemes of modern Estonian prose fiction). *Acta et Commentationes Universitatis Tartuensis 413, 5-140.*

Kang, K.M. & Seneta, E. (1980). Path analysis: an exposition. In: Krishnaiagh, P.R. (ed.), *Developments in Statistics 3, 217-246.*

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

Land, K.C. (1973). Identification, parameter estimation, and hypothesis testing in recursive sociological models. In: Goldberger, A.S. & Duncan, O.D. (eds.), *Structure equation models in the social sciences.* New York, Seminar Press.

McPherson, G. (1990). *Statistics in scientific investigation.* New York, Springer.

Miyajima, T. (1992). Relationships in the length, age and frequency of classical Japanese words. *Glottometrika 13, 212-229.*

Nowak, S. (1965). Causal interpretation of statistical relationships in social research. In: Nowak, S., *Studies in the methodology of the social sciences.* Warsaw.

Rothe, U. (1983). Wortlänge und Bedeutungsmenge: Eine Untersuchung zum Menzerathschen Gesetz an drei romanischen Sprachen. *Glottometrika 5, 101-112.*

Rosen, R. (1967). *Optimality principles in biology.* London, Butterworths.

Silnitsky, G.G., Andreev, S.N., Kuz'min, L.A. & Kuskov, M.I. (1990). *Sootnošenie glagol'nych priznakov različnych urovnej v anglijskom jazyke* (The relation between the verbal features at various levels in the English language). Minsk, Nauka i technika.

Storm, R. (1972[4]). *Wahrscheinlichkeitsrechnung, mathematische Statistik und statistische Qualitätskontrolle.* Leipzig, VEB Fachbuchverlag.

Suppes, P. (1970). *A probabilistic theory of causality.* Amsterdam, North-Holland.

Tuldava, J. (1977). Sagedussõnastik leksikostatistilise uurimise objektina (The frequency dictionary as an object of lexico-statistical analysis). *Acta et Commentationes Universitatis Tartuensis 413, 141-171.*

Tuldava, J. (1986). Dlina slova i raspredelenie slov po dline v tekste i slovare (On word length distribution in text and vocabulary). *Acta et Commentationes Universitatis Tartuensis 736, 150-166.*

Tuldava, J. (1993). *Probleme und Methoden einer quantitativen system- und textbezogenen Wortschatzforschung.* Hagen, Rottmann (in press).

Wiener, N. (1956). *I am a Mathematician.* Garden City, N.Y., Doubleday.

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

Zörnig, P., Köhler, R., Brinkmöller, R. (1990). Differential equation models for the oscillation of the word length as a function of the frequency. *Glottometrika 12, 25-40.*

3

On the Measurement of Correlation Between Qualitative Features in Linguistics: Contingency of Alternative Features

To determine and measure correlations (dependences) between various qualitative features of the object of study, many sciences make use of the coefficients termed coefficients of contingency. This paper is concerned with a specific type of relationship - contingency of alternative features as well as with ways of contingency identification, measurement and interpretation. [1]

1. 2 x 2 Table

To analyze contingency of alternative features we shall limit ourselves to two dichotomous variables, that is, each of the two can have two values: A_1 and A_2 (or A and *non-A*) and B_1 and B_2 (or B and *non-B*), respectively. Consequently, there are four possible configurations of their conjoint appearance: (A_1B_1), (A_1B_2), (A_2B_1) and (A_2B_2).

Table 1
2 x 2 contingency table

	B_1	B_2	Total
A_1	n_{11}	n_{12}	$n_{1.}$
A_2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

[1] For some recent works on the analysis of contingency tables in linguistics see, e.g., Grotjahn (1979, chap. 12), Altmann (1987), Schulz & Altmann (1988), Ivanyuk (1989), Andreev (1990), Hammerl & Rogalińska (1992), Silnitsky (1993).

We shall mark the frequency of occurrence of these configurations by n_{ij} (here: n_{11}, n_{12}, n_{21}, n_{22}). These data can be represented graphically in a four-field 2 x 2 table (Table 1).

In Table 1 $n_{i.}$ (i.e. $n_{1.}$ and $n_{2.}$) denotes row totals, $n_{.j}$ (i.e. $n_{.1}$ and $n_{.2}$) denotes column totals, and n is the grand total ($\Sigma_i \Sigma_j n_{ij}$). The corresponding probabilities will be expressed by p_{ij} (= n_{ij}/n), $p_{i.}$ (= $n_{i.}/n$), $p_{.j}$ (= $n_{.j}/n$) and $p_{..} = \Sigma_i \Sigma_j p_{ij} = 1$.

The figures (frequencies) in the contingency table of alternative features can be arrived at and presented mainly in two different ways (Case I and Case II).

Case I: Matched observations

Two dichotomous characteristics (features) are observed on n experimental units. For instance, n objects are analyzed and classified according to the presence or absence of the features A and B. To give an example, n verbs are analyzed for the presence or absence of the semantic feature termed "causative meaning" (A or *non-A*) and of the word-formation feature of derivation (B or *non-B*). The task is to identify *association* (correlation, dependence) between the features compared (A and B). Associated features tend to co-occur but this does not invariably mean that the features are in causal relationship. The interaction of cause and effect is to be proved by special investigations (see the article "On causal relations in language" in this issue).

Case II: Unmatched observations

One dichotomous characteristic is observed on two independent binomial samples (objects), each divided into two classes. For instance, we may analyze verbs (A_1) opposed to non-verbs (A_2), i.e. all the other parts of speech, as they relate to some phonetic characteristic, say "length of word", divided into monosyllabic (B_1) *vs* polysyllabic (B_2) words. Here the 2 x 2 table is used to compare A_1 with A_2 and to identify *differences* between them.

A special variant of Case II would be the model for "taste test", suggested by R.A.Fisher. According to this model two independent observers use a dichotomous characteristic to group the experimental units into two sets of predetermined size (cf. Nguyen & Rogers 1989: 293).

The dividing line between association and difference (Case I and Case II) is not always easy to establish. Depending on the aim of the investigation we would, for instance, equally well view the above-mentioned objects A_1 and A_2 ("verbs" and "non-verbs") as one dichotomous characteristic of "part of speech" (A) - in relation to the characteristic of "length of word" (B) divided into B_1 and B_2 to establish an association between the two features A and B observed on a

single sample of n words.

Contingency analysis with the help of contingency tables should be a useful and reliable tool in studies of

(i) identifying a relationship between the variables;

(ii) interpreting a relationship between the variables;

(iii) measuring the degree (strength, intensity) of a relationship between the variables.

This is done with the help of significance tests measuring the degree of *independence* in Case I and that of *homogeneity* in Case II, and special techniques for contingency measurement. For practical purposes both cases are covered by a single set of formulas.

This paper will mainly be concerned with Case I.

2. Hypothesis testing

To determine whether there is any relationship between two sets of variables (A and B) expressed in a 2 x 2 contingency table, we shall apply the chi-square test of significance (for other possible tests of significance see, e.g., Read & Cressie (1988)). We assume no significant association (or difference, as the case may be) between the sets of observations, and the chi-square test will tell us about the probability of the association (or difference) occurrence.

To test the significance level of an association between matched observations (Case I), the traditional null hypothesis is that of *independence* of the variables A and B. This translates to

$$H_o : p_{ij} = p_{i.}\, p_{.j} = (n_{i.}\, n_{.j})/n^2 \qquad \text{for } i, j = 1,2$$

vs.

$$H_1 \text{ (the alternative hypothesis)} : p_{ij} \neq p_{i.}\, p_{.j}\,.$$

When the chi-square test is used for assessing the significance of an observed *difference* between two independent, unmatched samples (Case II), the assumption is made that there is no significant difference between them, and the null hypothesis can be formulated as

$$H_o : p_{j/i} = p_j \text{ where } p_{j/i} = n_{ij}/n_{i.}\,; \qquad H_1 : \text{otherwise.}$$

The chi-square test can be performed using the formula

$$(1) \qquad X^2 = \sum_i \sum_j \frac{(n_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}$$

where $n\hat{p}_{ij}$ is the expected value (frequency).

Low values of X^2 indicate that the tentative assumption of insignificant association (or difference) cannot be denied. The higher the value of X^2, the smaller the likelihood is that the assumption of insignificant association is correct, and that the association (or difference) can be considered significant at a certain level of significance. As all 2 x 2 contingency tables have one degree of freedom (resulting from df = (r - 1)(c - 1) where r and c denote the number of rows and columns, respectively), the critical values of various levels are, e.g.

$$X^2_{0.05} = 3.84 \; ; \qquad X^2_{0.01} = 6.64 \; ; \qquad X^2_{0.001} = 10.83 \; .$$

To illustrate the application of significance criteria on matched observations (Case I) we shall take an example from a recent study in quantitative linguistics.

O. Golovinskaja (1987) studies and describes interrelations between various aspects of word derivation on the one hand and the syntactic and semantic characteristics of English verbal derivation on the other. She is especially interested in the relationship between the type of verbal derivation and the part of speech the words to which the verbal affixes are added belong to according to the Concise Oxford Dictionary of Current English (1977, 6th ed.). The derived verbs demonstrated the following cross-distribution of frequency of occurrence in the dictionary (see Table 2).[2]

The purpose of the investigation is to establish whether the two characteristics of "type of derivation" (A) and "part of speech the stem belongs to" (B) can be correlated or not. In other words: Is there an *association* between the two characteristics A and B ?

We set up the null hypothesis of two sets being independent, i.e. that there is no association between the characteristics. This hypothesis is tested by means of the chi-square test.

[2] The small group of verbs with both suffixes and prefixes has been joined with the group of suffixed verbs. All in all there were only 48 verbs of this kind (32 with noun stems and 16 with non-noun stems) (Golovinskaja 1987: 54).

Table 2
Absolute and relative frequencies of cross-classified data

Stem → Type of derivation ↓	Noun	Non-noun (other parts of speech)	Total
Suffix	498 (0.2056)	440 (0.1817)	938 (0.3873)
Non-suffix (Prefix)	477 (0.1969)	1007 (0.4158)	1484 (0.6127)
Total	975 (0.4025)	1447 (0.5975)	2422 (1.0)

The expected number (E) in each cell is given by

$$E(n_{ij}) = \frac{\textit{column total} \times \textit{row total}}{\textit{total number of obs.}} = \frac{R \times C}{n} \; .$$

Provided that our calculation of the first value of E is correct, the other three values can be found simply by subtraction from row (R) and column (C) totals. Table 3 presents the calculated values.

Table 3
The expected values for the data in Table 2

377.6	560.4	938
597.4	886.6	1484
975	1447	2422

The chi-square value calculated by formula (1) is:

$$X^2 = [(498\text{-}377.6)^2/377.6] + [(440\text{-}560.4)^2/560.4] +$$

$$+ [(477\text{-}597.4)^2/597.4] + [(1007\text{-}886.6)^2/886.6] = 104.87$$

which is much more than the critical chi-square value at the 0.001 level of significance (10.83) for one degree of freedom. Provided there is no association between the characteristics *A* and *B*, this value of X^2 (104.87) indicates that the observed frequencies in Table 2 are likely to occur in this configuration less than once in more than a thousand. Therefore the null hypothesis has been refuted and the association between *A* and *B* has proven to be statistically highly significant. That points to a dependence between the type of derivation of English verbs and the part of speech the affixed word stem belongs to.

As is known, there are several mathematically identical formulas for chi-square calculation. A special formula to be used with 2 x 2 tables is the following:

$$X^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}} \quad (2)$$

where $n_{1.}$ and $n_{2.}$ are the row totals and $n_{.1}$ and $n_{.2}$ are the column totals. The X^2 value for the data in Table 2 is

$$X^2 = \frac{2422[498(1007) - 440(477)]^2}{938(1484)975(1447)} = 104.87,$$

i.e. exactly the same as the one we arrived at with formula (1).

3. Modified tests

The chi-square test is said to be applicable to data of any kind of distribution. However, there are some restrictions. First, the chi-square test is fully appropriate for measurements on a continuous scale, but with discrete numbers it is somewhat inaccurate, especially when applied to 2 x 2 tables (cf. Langley 1968: 285). The chi-square test also tends to be inaccurate unless all the cells have an expected number of five or more and unless the total number of observations (n) is 50 or more. Yates' correction for discrete numbers and Fisher's exact test for small numbers of observations are more accurate and, therefore, recommendable.

The modification of the chi-square test, proposed by F.Yates (1934), lies in the subtraction of 0.5 from each absolute value of the difference between observed and expected numbers before squaring them, which splits the difference between the whole numbers. In reference to the data in Tables 2 and 3, the values of the modified chi-square test will be:

$$X^{2*} = [(120.4 - 0.5)^2/377.6] + [(120.4 - 0.5)^2/560.4] +$$

$$+ [(120.4 - 0.5)^2/597.4] + [(120.4 - 0.5)^2/886.6] = 104.00.$$

The values of Yates' chi-square test can be directly calculated by the formula

$$X^{2*} = \frac{n(|n_{11}n_{22} - n_{12}n_{21}| - 0.5n)^2}{n_{1.}n_{2.}n_{.1}n_{.2}}. \quad (3)$$

The result (104.00) is the same as by the former calculation.

The chi-square values calculated with the help of Yates' correction formula are always somewhat smaller than the values obtained with the unmodified formula; in our experiment 104.00 and 104.87 respectively. The difference is insignificant, but it may have an effect of statistical significance on the calculation of small values of X^2 (see below). It was pointed out by R.Langley (1968: 373) that Yates' chi-square test is inaccurate when two cells have an expected value ($\hat{n}_{ij}$) smaller than 5, but it could be used with discretion if only one cell has the expected value between 1 and 5. The accuracy of Yates' test increases with larger values of n.

With very small values in the cells of a 2 x 2 table (five or less) and with small samples (less than 50), the chi-square test (modified or unmodified) is less effective and precise. To establish the level of statistical significance in small-size samples, the "exact test" proposed by R.Fisher (1934) may be recommended. Fisher's test is calculated by means of the hypergeometric formula

$$P = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n_{11}!n_{12}!n_{21}!n_{22}!n!} \quad (4)$$

where $m!$ stands for m factorial (with $0! = 1$).

To provide for calculation ease factorial logarithms may be used. Formula (4) will then be translated to

$$P = exp\{[ln(n_{1.}!) + ln(n_{2.}!) + ln(n_{.1}!) + ln(n_{.2}!)] - [ln(n_{11}!) + ln(n_{12}!) + ln(n_{21}!) + ln(n_{22}!) + ln(n!)]\}. \quad (5)$$

The P criterion presents the probability (level of significance) in cases where the variables are presumed to be independent. For instance, when $P > 0.05$ the null hypothesis of independence may be accepted, and when $P < 0.05$ the null hypothesis is rejected and the association between the characterististics will be studied at the $P < 0.05$ level of significance. However, it is necessary to remember that Fisher's exact test is a one-sided test and for two-sided tests the probability has to be doubled.

An example of the application of Fisher's exact test is the following (Table 4).

Table 4
Cross-classified data

9	3	12
6	12	18
15	15	30

Formula (5) will give us a preliminary estimation of the probability level. If $P > 0.05$ the null hypothesis can be accepted. For the exact overall probability estimation some more calculating is necessary as Fisher's test is based on the examination of all possible extreme arrangements of data and the calculation of the probabilities of all the arrangements (for details see, e.g., Upton (1978, chap. 2.4)). We begin with the smallest value of the observations and rearrange the values so that the marginal values would remain unchanged (see Table 5).

Table 5
Extreme arrangements of the data in Table 4

10	2	12
5	13	18
15	15	30

(P = 0.00364)

11	1	12
4	14	18
15	15	30

(P = 0.00024)

12	0	12
3	15	18
15	15	30

(P = 0.00001)

Let us first calculate the probability for the data in Table 4. Using factorial logarithms

$$P = exp\{[\ ln\ (12!) + ln\ (18!) + ln\ (15!) + ln\ (15!)\ ] -$$

$$- [\ ln\ (9!) + ln\ (3!) + ln\ (6!) + ln\ (12!) + ln\ (30!)\ ]\} = 0.02633.$$

This is the value of the one-sided probability. The two-sided probability would be 2 x 0.02633 = 0.05266. Consequently, the null hypothesis of independence at the 0.05 level cannot be rejected, and further calculations would not be needed (as they would only add probabilities).

If we continue the calculation, we get the probabilities 0.00364, 0.00024 and 0.00001 for the "extreme arrangements" (see Table 5). By summing up all the values (including the initial value for Table 4), we arrive at the exact overall probability

$P = 0.02633 + 0.00364 + 0.00024 + 0.00001 = 0.03022$ for the one-sided and

2 x 0.03022 = 0.06044 for the two-sided criterion.

Fisher's exact test should be used in cases in which the total number of observations (n) is between eight and fifty. When n is less than eight, a 0.05 probability level cannot be reached. To avoid troublesome calculations, special tables for Fisher's exact test are available (see, e.g., Finney et al. (1963); Langley (1968); Rasch (1970)).

The superiority of Fisher's exact test in comparison with the ordinary chi-square test can be demonstrated on the data calculations of Table 4. The chi-square calculation by formula (2) results in $X^2 = 5.0$, significant at the 0.05 level ($P < 0.05$). The modified chi-square test with Yates' correction results in $X^{2*} = 3.47$ which is lower than the critical value $X^2_{0.05;\ 1} = 3.84$. This value as well as the value arrived at by Fisher's exact test is interpreted as a sign of absence of association (i.e. the null hypothesis of independence cannot be rejected).

4. Analysis of residuals

Where there is evidence of association between the variables it is recommended to construct and examine the set of *standardized residuals* based on the fit of the 2 x 2 model to establish the structure and the nature of the relationship between the objects under investigation (by searching for largest residuals, analyzing systematic patterns of plus and minus signs, etc.). Residuals of various types have

been proposed for this purpose (see, e.g., Haberman 1973; Santner & Duffy 1989; a special method for analyzing residuals in contingency tables - the so-called "sign scheme" analysis - is discussed in Řehák & Řeháková 1980).

The basic building blocks for the standardized residuals are the "raw residuals", defined as

$$(6) \qquad d^R_{ij} = n_{ij} - n\hat{p}_{ij}$$

where n_{ij} are the observed and $n\hat{p}_{ij} = \hat{n}_{ij}$ the expected values.

For 2 x 2 tables the most popular formula for calculating standard residuals is the formula for Pearson residuals, defined by

$$(7) \qquad d^P_{ij} = \frac{d^R_{ij}}{\sqrt{n\hat{p}}} = \frac{n_{ij} - n\hat{p}_{ij}}{\sqrt{n\hat{p}_{ij}}}$$

which satisfies $X^2 = \Sigma_{ij} (d^P_{ij})^2$. Therefore the Pearson residuals are often called "chi-components".

We shall use formula (7) to calculate the standardized residuals for the data in Table 2 (cf. also Table 3). Calculating d^P_{ij} for all four cells in the 2 x 2 table, the resulting values are:

$$d^P_{11} = \frac{498 - 377.6}{\sqrt{377.6}} = + 6.20; \qquad d^P_{12} = \frac{440 - 560.4}{\sqrt{560.4}} = - 5.09;$$

$$d^P_{21} = \frac{477 - 597.4}{\sqrt{597.4}} = - 4.93; \qquad d^P_{22} = \frac{1007 - 886.6}{\sqrt{886.6}} = + 4.04.$$

The values of the standardized residual calculation can the presented in tables similar to the 2 x 2 table presenting the initial data (see Table 6).

As can be seen from Table 6, the largest residual (+6.20) is in the first cell in the first row (n_{11}), associated with the joint occurrence of the features of "suffixation" and "noun stem" (e.g. *ideal-ize*, *function-ate*). Strong negative correlations are to be found between noun stems and prefixation (n_{21}) and non-noun stem and suffixation (n_{12}), especially between verb stems and suffixation (cf. Golovinskaja 1987).

The Pearson standardized residuals are approximately normally distributed with the mean value of 0 and the variance of 1. The magnitude of the absolute

values of Pearson residuals in Table 6 gives evidence that there is a significant difference between the observed and expected values and, consequently, there can be an association between the features examined. (With the normal distribution of the standardized residuals, the critical value at the 0.001 level is 3.29.)

Another variant of standardized residuals, presumably a better approximation to the normal distribution, is that of F.J.Anscombe (1953), defined by

Applied to our material, the calculation of Anscombe residuals gives results which are numerically somewhat different but, on the whole, analogous (Table 7).

Table 6
Pearson residuals under independence model for the data in Table 2

+ 6.20	- 5.09
- 4.93	+ 4.04

In this context it can be noted that the standardized residuals expressed by the raw residuals divided by the standard deviation are not an appropriate technique to compare the residuals in the cells of the 2 x 2 table, because of the resultant identical numerical values for all the cells. As shown in Haberman (1973) and Santner & Duffy (1989), the asymptotic variance in a 2 x 2 table, calculated on the basis of cell values, equals

$$(9) \quad Var(d^R_{ij}) = n\hat{p}_{i.}(1 - \hat{p}_{i.})\hat{p}_{.j}(1 - \hat{p}_{.j})$$

Table 7
Anscombe residuals for the data in Table 2

+ 5.92	- 5.28
- 5.10	+ 3.96

By replacing empirical relative frequencies for the probability values, the variance value for Table 2 can be calculated :

$$Var\,(d^R_{11}) = 2422\ (0.3873)\ 0.6127\ (0.4025)\ 0.5975 = 138.2208;$$

$$Var\,(d^R_{12}) = 2422\ (0.3873)\ 0.6127\ (0.5975)\ 0.4025 = 138.2208;$$

etc. Then the formula for the standardized residuals of this type, called "Haberman´s adjusted residuals" (Santner & Duffy 1989: 148), adopts the following expression:

(10) $$d_{ij}^{H} = \frac{d_{ij}^{R}}{\sqrt{Var(d_{ij}^{R})}}.$$

(8) $$d_{ij}^{A} = \frac{3[n_{ij}^{2/3} - (\hat{n}_{ij} - 1/6)^{2/3}]}{2\hat{n}_{ij}^{1/6}}.$$

Applying formula (10) to our material (Table 2), we get

$$d^{H}_{11} = 120.4 / (138.2208)^{1/2} = 10.2410;$$

$$d^{H}_{12} = - 120.4 / (138.2208)^{1/2} = - 10.2410;$$

etc. Comparing the values of d^{H}_{ij}, we can see that

$$d^{H}_{11} = d^{H}_{22} = -d^{H}_{12} = -d^{H}_{21}$$

and $(d^{H}_{ij})^2 = X^2$. Here $10.241^2 = 104.88$ which coincides with $X^2 = 104.87$ (see Section 2; the small difference is a round-off error).

Further it can be noted that the adjusted residuals d^{H}_{ij} are related to Pearson residuals by

(11) $$d_{ij}^{H} = \frac{d_{ij}^{P}}{[(1 - p_i)(1 - p_j)]^{1/2}}$$

and they coincide with the so-called z-scores discussed in detail in Řehák & Řeháková (1980).

5. Traditional measures of association

As noted in the previous sections, the chi-square test helps to identify an association between characteristics and to estimate the statistical significance of this association. However, it does not measure the *degree* of the association, i.e. its strength or intensity of connotative dependences. Now, to allow for a more differentiated and precise specification, a number of indices characterizing the strength of the relationship between qualitative alternative features are described

and known. Of the large number of coefficients measuring the interdependence there are two which are often used to characterize the association of alternative features in 2 x 2 tables (for Case I of contingency measurement; see Section 1). These are Yule's coefficient Q and the Bernoullian correlation coefficient Φ ("Phi-coefficient").[3] They are calculated according to the formulas:

(12) $$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}};$$

(13) $$\Phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}}.$$

As can be seen, these two formulas (12) and (13) look alike as their numerators are identical and only their denominators are different. Both coefficients function within the range of -1 and +1, with the extreme points pointing to full association (negative or positive) and the zero point signifying the absence of association (i.e. the evidence of independence of the features compared).

Q and Φ calculations are also similar in their allowing for the multiplying or dividing of all the frequencies in the 2 x 2 table by one and the same number without changing the outcome. So Q and Φ coefficients, calculated with the help of formulas (12) and (13), can be expressed in percentages of the total number of the observations (n) analyzed. However, coefficients Q and Φ differ in their ability to measure different aspects within the four-field table.

From the theoretical point of view it is important to note that coefficient Q is closely related to the concept of *odds ratio*. Odds can be defined by the expression (cf. Christensen 1990: 28):

$$Odds = \frac{p}{1 - p} = \frac{Pr(Event\ occurs)}{Pr(Event\ does\ not\ occur)}.$$

So, in our example (Table 2), the odds of suffixes (feature A_1) being combined with noun stems (feature B_1) *vs* becoming combined with non-noun stems (fea-

[3] The term "Bernoullian correlation coefficient" was suggested by G.Herdan (1966: 421). We shall not discuss the well-known coefficients of Goodman & Kruskal (t), Cramér (V) and Čuprov (T) in this paper as they are equivalent to Φ^2 while applied to 2 x 2 tables. For a discussion of the applicability of these coefficients to linguistic material see Schulz & Altmann (1988).

ture B_2) is p_{11}/p_{12}, or, replacing empirically observed values for the probabilities, we have n_{11}/n_{12} = 498/440 = 1.1318. We could say, "the odds of suffixes fixed to noun stems are 1.1318 to one". On the other hand, the odds of prefixes (feature A_2) being added to noun stems (B_1) against their being fixed to non-noun stems (B_2) are n_{21}/n_{22} = 477/1007 = 0.4737. So, the *odds ratio* is

$$\frac{p_{11}/p_{12}}{p_{21}/p_{22}} \tag{14}$$

or

$$\frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{1.1318}{0.4737} = 2.3893.$$

This means that the odds of the joint occurrence of A_1 and B_1 are nearly two and a half times as large as the odds of the joint occurrence of A_2 and B_1. The result, 2.3893, is fairly far removed from the target value of 1, which would mean equality of odds.

Now, the odds ratio (14) can be transformed to the *cross-product ratio*

$$C = \frac{p_{11}p_{22}}{p_{12}p_{21}} \tag{15}$$

and we can rewrite the Q coefficient as

$$Q = \frac{C - 1}{C + 1}. \tag{16}$$

Thus, we have demonstrated that coefficient Q, as a measure of the degree of association, is actually a function of the odds ratio (cf. Christensen 1990: 60). In this quality it reflects a specific aspect of the dependence between the characteristics examined, which we could define as "one-way dependence" between A_1 and B_1.

Coefficient Φ, however, evaluates two aspects simultaneously, i.e. besides measuring a one-way dependence, it also reflects the effect which the absence of factor A (i.e. A_2) exercises on factor B (i.e. B_2). Furthermore, coefficient Φ reflects another type of association between the variables. As is known, this co-

efficient belongs to the group of contingency measures which are based on the chi-square test, and formula (13) is mathematically identical with the formula

$$\Phi = \sqrt{X^2/n}\ . \tag{17}$$

As was pointed out by Olson (1985: 495), "it is important to recognize that the chi-square test for two-way tables does not test for differences between the levels of variable A or for difference between the levels of variable B, per se. Rather, it tests whether the proportions of observational units at the various levels of variable A remain unchanged regardless of which level of variable B is considered." So, the chi-square test is, in effect, like a test of correlation between the variables, and, consequently, coefficient Φ has the same quality. In fact, coefficient Φ may be treated as a modification of the Pearson product-moment coefficient of paired correlation which is calculated according to its mathematically strict formula

$$r_{xy} = \frac{n\Sigma xy - \Sigma x\ \Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}} \tag{18}$$

with the presence/absence of characteristics in the four-field table represented by 1 and 0. A small example could be helpful in making this connection clear (see Table 8) .

Table 8
Cross-classified data

x\y	1	0	Total
1	5	1	6
0	2	2	4
Total	7	3	10

We shall present the calculation of the values of r_{xy} in Table 9.

Using formula (18), we get

$$r_{xy} = \frac{10(5) - 6(7)}{\sqrt{[10(6) - 6^2][10(7) - 7^2]}} = 0.3563.$$

Table 9
r_{xy} value calculation

No.	x	y	x^2	y^2	xy
1.	1	1	1	1	1
2.	1	1	1	1	1
3.	1	1	1	1	1
4.	1	1	1	1	1
5.	1	1	1	1	1
6.	1	0	1	0	0
7.	0	1	0	1	0
8.	0	1	0	1	0
9.	0	0	0	0	0
10.	0	0	0	0	0
n = 10	6	7	6	7	5

The same result will be obtained when the calculation is made according to formula (13):

$$\Phi = \frac{5(2) - 1(2)}{\sqrt{7(3)6(4)}} = 0.3563.$$

Proceeding from the differences in the structure of coefficients Q and Φ, the values calculated on the data should differ, with $Q > \Phi$ as a rule. The values will be the same ($Q = \Phi = 1$) in the case of a full dichotomous dependence, i.e. when one feature can be fully explained through the other: each object under study having feature A will inevitably also have feature B and the other way round. The values are also equal ($Q = \Phi = 0$) in the case of frequency values reflecting absolute independence of the characteristics (e.g. Table 10).

Table 10
Absolute dependence and absolute independence

30	0	30
0	70	70
30	70	100

($Q = \Phi = 1$)

30	30	60
20	20	40
50	50	100

($Q = \Phi = 0$)

In the case of one-sided absolute dependence, meaning that all the objects with feature A_1 will also have feature B_1, whereas the objects with feature B_1 need not have feature A_1 only (e.g. Table 11 the first example), $Q = 1$ and $\Phi < 1$. As can be seen from the values in Table 11 (the second example) a moderate or weak Φ dependence does not exclude the possibility of a fairly strong Q dependence, i.e. $Q > \Phi$.

We shall calculate the values of both coefficients Q and Φ for our example (Table 2):

$$Q = \frac{498(1007) - 440(477)}{498(1007) + 440(477)} = 0.4099;$$

$$\Phi = \frac{498(1007) - 440(477)}{\sqrt{975(1447)938(1484)}} = 0.2081.$$

Table 11
Q and Φ dependences

30	0	30
20	50	70
50	50	100

($Q = 1$; $\Phi = 0.65$)

30	10	40
20	40	60
50	50	100

($Q = 0.71$; $\Phi = 0.41$)

Coefficient Φ as expected, does not present the interdependence between the features "type of verb derivation" and "part of speech the stem belongs to" in such a pronounced way as coefficient Q does (expressing "one-way" dependence). Now we can proceed to find the confidence limits of the coefficients.

Provided that the observed values are normally distributed (n being sufficiently representative), the standard deviation of Q is calculated according to the formula

$$\sigma_Q = \frac{1 - Q^2}{2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

The confidence limits are expressed through the formula $k\sigma_Q$, in which the value of k depends on the significance level chosen. The value of Q can be considered statistically significant if $Q > k\sigma_Q$. The standard deviation in the example is :

$$\sigma_Q = \frac{1 - 0.4099^2}{2}\sqrt{\frac{1}{498} + \frac{1}{440} + \frac{1}{477} + \frac{1}{1007}} = 0.0351.$$

As the empirically found Q value (0.4099) is larger than the three-fold value of the standard deviation (3 x 0.0351), then Q may be considered statistically significant at the level of $P < 0.0027$.[4] The confidence limits at that level are defined by 0.4099 ± 3(0.0351) or (0.3046; 0.5152).

The standard deviation of Φ can be calculated using the formula

$$\sigma_\Phi = \frac{1 - \Phi^2}{\sqrt{n}} \tag{20}$$

where n is the number of observations. The confidence limits for the Φ value are calculated using the standard deviation

$$\sigma_\Phi = \frac{1 - 0.2081^2}{\sqrt{2422}} = 0.0194.$$

The three-fold value of the standard deviation is 3 × 0.0194 = 0.0582, and the confidence limits at that level ($P < 0.0027$) are 0.2081 ± 0.0582 or (0.1499; 0.2663).

The confidence intervals of Q and Φ exceed the zero value, and thus we can conclude that the dependence between the characteristics is statistically significant. This corresponds to the estimate obtained by means of the chi-square test as described in Section 2.

Before concluding this section, there is another important feature of coefficient Φ calculation that is worth consideration. For practical purposes it may be necessary to compare the values of Φ related to different experiments. It would be incorrect to compare the coefficient values directly, as they can have different extreme values depending on the distribution of frequency values in the

[4] As two-fold and three-fold values of standard deviations are considered sufficient, then $k(P = 0.045) = 2.0$; $k(P = 0.0027) = 3.0$.

2 x 2 table. The extreme values of Φ (-1 or +1) will be achieved only in case two diagonal cells in the 2 x 2 table contain zeros. In all other cases the extreme values are smaller than one. There are special ways of calculating the extreme values (see, e.g., Clauß & Ebner 1970: 255). The extreme values help to correct the empirical Φ value by expressing it in its relation to the possible extreme (maximal or minimal) value: $\Phi_{corr} = \Phi/\Phi_{(max/min)}$. The corrected Φ values can be calculated directly by the formulas suggested by L.C.Cole (1949):

$$\Phi_{corr} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n(\min\, b,\, c) + (n_{11}n_{22} - n_{12}n_{21})} \quad \text{if } \Phi \geq 0; \tag{21}$$

$$\Phi_{corr} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n(\min\, a,\, d) - (n_{11}n_{22} - n_{12}n_{21})} \quad \text{if } \Phi < 0. \tag{22}$$

For example let us calculate the corrected Φ value for the experiment described in Table 2 and use formula (22) :

$$\Phi_{corr} = \frac{498(1007) - 440(477)}{2422(440) + [498(1007) - 440(447)]} = 0.2148.$$

The Φ_{corr} value compared with Φ = 0.2081, calculated by means of formula (13), does not manifest much difference (in our case). But to demonstrate the importance of correcting the Φ values, we shall present another example (Table 12).

Table 12
Comparison of Φ values

Experiment I

60	18	78
12	10	22
72	28	100

Φ = 0.2065
Φ_{corr} = 0.2424

Experiment II

60	10	70
21	9	30
81	19	100

Φ = 0.1836
Φ_{corr} = 0.2481

Judging by the uncorrected Φ values the dependence between the characteristics

may seem stronger in the first experiment (0.2065 and 0.1836 respectively). However, if the corrected Φ values are compared, it is quite obvious that the dependence in both experiments is approximately the same in strength (or in fact it is even slightly stronger in the second experiment). Of course, strictly speaking, the values of the coefficients should be compared with due consideration given to their confidence limits.

6. Informational measures

Alongside traditional measures of contingency, there are measure systems built up on the notions of information theory (see, e.g. Linfoot 1957; Kullback 1959). These measures of contingency have not been used to describe material in linguistics so far. However, as informational measures describe interrelationships between the objects under study from a slightly different angle, they can, in some cases, make for a deeper insight into the structure and nature of the interdependences of linguistic objects as well as for a more profound interpretation of the essence of the dependences. As informational measures do not depend on statistical distributions, they deserve special consideration as valid tools in measuring the interaction of the objects.

In this context one should remember that the concept of information invariably involves the concept of uncertainty, or *entropy*. The interdependence between information and entropy both measured in the same units is expressed by waning uncertainty (entropy) increasing the amount of information. In our study, it is important to remember that the notion of information can be associated with the notion of dependence (association). In practice, we can say that we speak of a dependence between the two variables *A* and *B* if the knowledge of *A* is capable of reducing the uncertainty of *B*. In other words, to measure a dependence between *A* and *B* means to measure the amount of reduced uncertainty of *B* within the system (*A*,*B*) against the amount of uncertainty of *B* without considering the system (*A*,*B*).

We shall demonstrate the application of informational measures on the case we used to illustrate the application of statistical contingency measures (Table 2) on the dependence between the "Type of verbal derivation" (*A*) and "Part of speech the stem belongs to" (*B*). For convenience sake, we shall present the probabilities (relative frequencies) in Table 2 and their notations once again in Table 13.

The basic criterion for the assessment of a dependence between the variables *A* and *B* is the measure of "shared information" denoted by *I*(*A*,*B*) and defined by

(23) $$I(A,B) = H(A) + H(B) - H(A,B)$$

where *H*(*A*) and *H*(*B*) are the entropies of *A* and *B* when considered independently, *H*(*A*,*B*) - the entropy of the joint event. They are calculated by means of the following formulas (cf. Kullback 1959, chap. 2.8):

Table 13
Probabilities in Table 2

	B	non-B	Total
A	0.2056 (p_{11})	0.1817 (p_{12})	0.3873 ($p_{1.}$)
non-A	0.1969 (p_{21})	0.4158 (p_{22})	0.6127 ($p_{2.}$)
Total	0.4025 ($p_{.1}$)	0.5975 ($p_{.2}$)	1.0 ($p_{..}$)

(24) $$H(A) = - \Sigma\, p_{i.}\, ln\, p_{i.}\,;$$

(25) $$H(B) = - \Sigma_j\, p_{.j}\, ln\, p_{.j}\,;$$

(26) $$H(A,B) = - \Sigma_i \Sigma_j\, p_{ij}\, ln\, p_{ij}\,.$$

Coefficient *I*(*A*,*B*) serves as measure of the degree (intensity) of the dependence between *A* and *B* and can also be calculated according to the formula

(27) $$I(A,\ B) = -\Sigma_i \Sigma_j p_{ij} \ln \frac{p_{ij}}{p_{i.} p_{.j}}.$$

For the example (Table 13) the entropies are calculated as follows:

$$H(A) = - (0.3873 \ln 0.3873 + 0.6127 \ln 0.6127) = 0.6675;$$
$$H(B) = - (0.4025 \ln 0.4025 + 0.5975 \ln 0.5975) = 0.6740;$$

$$H(A,B) = -\,(\,0.2056 \ln 0.2056 + 0.1817 \ln 0.1817 + 0.1969 \ln 0.1969 + 0.4158 \ln 0.4158 = 1.3200.$$

The amount of shared information is then calculated according to formula (23):

$$I(A,B) = 0.6675 + 0.6740 - 1.3200 = 0.0215,$$

or, according to formula (27):

$$I(A,B) = -[0.2056 \ln \frac{0.2056}{0.3873(0.4025)} + 0.1817 \ln \frac{0.1817}{0.3873(0.5975)} + 0.1969 \ln \frac{0.1969}{0.612(0.4025)} + 0.4158 \ln \frac{0.4158}{0.6127(0.5975)}] = 0.0215.$$

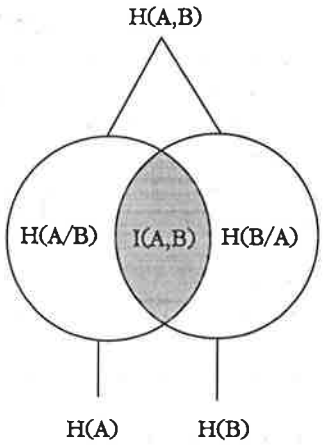

Figure 1. Relations between information and entropy within the system of *two* nominal variables A and B

The relations between the entropies and $I(A,B)$ are graphically presented in Fig.1 (adapted from Eliseeva & Rukavišnikov 1977: 96).

To test the statistical significance of the computed value of $I(A,B)$, i.e. the significance of interdependence (association) between A and B, the loglikelihood ratio statistic G^2 can be used:

(28) $$G^2 = 2nI = 2n(-\,\Sigma_i\Sigma_j p_{ij} \ln \frac{p_{ij}}{p_i p_j}).$$

In the case under study:

$$G^2 = 2(2422)\ 0.0215 = 104.15.$$

As is known, the statistical tests G^2 and X^2 are of rather similar functioning (cf. $X^2 = 104.87$ and $G^2 = 104.15$ in the example). Both have a distribution which is approximately chi-squared with $(r\text{-}1)(c\text{-}1)$ degrees of freedom. Applied to the 2 x 2 table $df = 1$ and the critical value at the 0.001 level is 10.81. Consequently, the value of $G^2 = 104.15$ allows us to state that the interdependence between A and B is statistically significant. However, according to the informational measure $I(A,B)$, the degree (intensity) of dependence is not very high.

The $I(A,B)$ value can be transformed into a standardized relative value through the ratio between $I(A,B)$ and the average of the entropies $H(A)$ and $H(B)$ (cf. Řehák & Řeháková 1973):

(29) $$R(A,B) = \frac{I(A,B)}{\frac{1}{2}[H(A) + H(B)]}.$$

The calculation of this *symmetrical* measure results in the following value:

$$R(A,B) = 2(0.0215) / (0.6675 + 0.6740) = 0.0321\ .$$

Another possible symmetrical measure of interdependence is the ratio between $I(A,B)$ and the entropy of the joint event $H(A,B)$ (cf. Rajski 1964):

(30) $$R'(A,B) = \frac{I(A,B)}{H(A,B)}.$$

In our example the value will be 0.0215/1.3200 = 0.0163. This coefficient varies within the limits of 0 and 1.

The informational measures based on the *conditional entropies* of A and B are of special interest:

(31) $$H(B|A) = H(A, B) - H(A) = -\sum_i \sum_j p_{ij} \ln \frac{p_{ij}}{p_{i.}},$$

measuring the entropy of B when A is known, and

(32) $$H(A|B) = H(A, B) - H(B) = -\sum_i \sum_j p_{ij} \ln \frac{p_{ij}}{p_{.j}},$$

measuring the entropy of A when B is known.

The values for the example will be:

$$H(B|A) = 0.6524 \quad \text{and} \quad H(A|B) = 0.6459.$$

On the basis of conditional entropies, the following *asymmetric* measures of dependence can be constructed:

(33) $$R(B|A) = \frac{H(B) - H(B|A)}{H(B)} = \frac{I(A, B)}{H(B)}$$

and

(34) $$R(A|B) = \frac{H(A) - H(A|B)}{H(A)} = \frac{I(A, B)}{H(A)}.$$

(Cf. Fig. 1.)

The measurement of dependence between A and B according to formulas (33) and (34) is the estimation of the relative amount of reduced uncertainty of one of the variables within the system of *(A,B)* against the amount of uncertainty of the variable considered independently. In other words, the coefficients signify the diminishing level of uncertainty of one of the variables under the influence of the other variable. The coefficients vary within the limits of 0 and 1. In our case *R(B/A)* = 0.0319 and *R(A/B)* = 0.0322. These results are close to the result calculated by means of formula (29) - of the symmetrical measure *R(A,B)* = 0.0321 - because of the nearly symmetrical disposition of the marginal values

(row and column totals) in our example (see Tables 2 and 13). In other cases the difference between the values of *R(B/A)* and *R(A/B)* could be more noticeable and it would be possible to establish the dominant direction of influence for the system *(A,B)* provided it is corroborated by the qualitative analysis.

As to the interpretation of the numerical values of the informational measures we would refer the reader to the investigation of Astola & Virtanen (1983). The authors propose to use a square root transformation of index (29) in order to get a better measure of dependence which would fulfill both the theoretical requirements (theoretical foundations for this derived index are presented in Astola & Virtanen 1983: 18 et seq.) and intuitive expectations set for a correlation coefficient. In our case *R(A,B)* = 0.0321 and the square root equals $0.0321^{0.5} = 0.1792$. This value comes near to the value obtained by using the Phi-coefficient and indicates the degree of the diminishment of the uncertainty in the system when one of the variables is known. Likewise, the square roots of *R(B/A)* and *R(A/B)* could be calculated in order to measure the oriented dependence of B on A, and vice versa.

Besides using the informational measures for prognostic purposes and for measuring interdependences and oriented dependences between linguistic objects, the coefficients based on conditional entropy can also be used for estimating partial and multiple associations (see, e.g., Taganov & Turgumbaev 1984, chap. 1.4). This approach is also used in classifying objects. For instance, the value of *R(B/A)* shows not only the reduction of the entropy of B under the influence of A, but, on the other hand, it also indicates how "informative" the variable B is (covering a large amount of shared information) in comparison with other variables in the given system. This can be done by breaking up the data of R x C table into a number of 2 x 2 tables.

* * *

Our aim in this paper was to examine various aspects of contingency analysis applied to 2 x 2 tables - beginning with the establishment of the type of contingency and the measurement of statistical significance (including the analysis of residuals) and ending with the practical application and interpretation of various measures of contingency to study linguistic material. We have seen that different measures correspond to different types of contingency: "one-way" association as a function of odds ratio (*Yule's Q*); interdependence as a special kind of product-moment correlation (*Phi-coefficient*); dependence based on the reduction of uncertainty of one of the variables provided the other is known (*informational measures*). We can conclude from this study that the appropriate method of relationship estimation is to be selected depending on the aims and

purposes of the investigation. In recent years much attention has been given to studying dependences and correlations between objects in linguistic research, especially in the field of language synergetics. The analysis and interpretation of linguistic relations with contingency table calculation provide new insights into the functioning of language and thus make a contribution to the elaboration of the theory of language as a self-organized system.

The present study has not exhausted all the possibilities of contingency analysis. There are a number of new effective methods of R x C contingency table analysis, particularly the methods based on log-linear models. These methods will be the topic of a separate study.

References

Altmann, G. (1980). *Statistik für Linguisten.* Bochum, Brockmeyer.

Altmann, G. (1987). Tendenzielle Vokalharmonie. *Glottometrika 8, 104-112.*

Andreev, S.N. (1990). Statističeskie mery svjazi kačestvennych priznakov (Statistical measures of connection between qualitative features). In: Silnitsky, G.G. et al. (eds.). *Sootnošenie glagol'nych priznakov različnych urovnej v anglijskom jazyke.* Minsk, Nauka i technika, 21-34.

Anscombe, F.J. (1953). Contribution to discussion of paper by H.Hotelling. New light on the correlation coefficient and its transform. *J. of the Royal Statistical Society B, 15, 229-230.*

Astola, J. & Virtanen, I. (1983). *A measure of overall statistical dependence based on the entropy concept.* Vaasa, University Press (Vaasan korkeakoulun julkaisuja, Tutkimuksia No. 91, Tilastotiede 12).

Clauß, G. & Ebner, H. (1970). *Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen.* Berlin, Volk und Wissen.

Cole, L. (1949). The measurement of interspecific association. *Ecology 30, 411-424.*

Christensen, R. (1990). *Log-linear models.* New York etc., Springer.

Eliseeva, I.I. & Rukavišnikov, V.O. (1977). *Gruppirovka, korreljacija, raspoznavanie obrazov (Grouping, correlation, recognition of images).* Moscow, Statistika.

Finney, D.J., Latschka, R., Bennett, B.M., Hsu, P., Pearson, E.S. (1963). *Tables for testing significance in a 2 x 2 table.* Cambridge, Cambridge University Press.

Fisher, R.A. (1934). *Statistical methods for research workers.* London, Oliver & Boyd.

Golovinskaja, O.E. (1987). *Proizvodnye affiksal'nye glagoly v sovremennom anglijskom jazyke (Derived affixed verbs in modern English).* PhD thesis. Minsk.

Grotjahn, R. (1979). *Linguistische und statistische Methoden in Metrik und Textwissenschaft.* Bochum, Brockmeyer.

Haberman, S.J. (1973). The analysis of residuals in cross-classified tables. *Biometrics 29, 205-220.*

Hammerl, R. & Rogalińska, A. (1992). Über die Untersuchung mehrdimensionaler sprachlicher Relationen. In: Saukkonen, P. (ed.). *What is language synergetics? Seminar on the International Language Synergetics Project.* Oulu, University Press, 70-85.

Herdan, G. (1966). *The advanced theory of language as choice and chance.* New York etc., Springer.

Ivanyuk, V.Yu. (1988). *Vlijanie semantiki glagola na realizaciju jego kategorii vremeni v sovremennom nemeckom jazyke. (The influence of verbal semantics on the category of tense in contemporary German).* PhD thesis. Černovcy.

Kullback, S. (1959). *Information theory and statistics.* New York, Wiley.

Langley, R. (1968). *Practical statistics for non-mathematical people.* London, Pan Books.

Linfoot, E.H. (1957). An informational measure of correlation. *Information and Control, 1. 85-89.*

Levitsky, V.V. (1989). *Statističeskoe izučenie leksičeskoj semantiki (Statistical investigation of lexical semantics).* Kiev, UMK.

Nguyen, H.T. & Rogers, G.S. (1989). *Fundamentals of mathematical statistics. Vol. II: Statistical inference.* New York etc., Springer.

Olson, C.L. (1985). *Essentials of statistics. Making sense of data.* Boston etc., Allyn & Bacon.

Rasch, D. (1970). *Elementare Einführung in die mathematische Statistik.* Berlin, Verlag der Wissenschaften.

Rajski, C. (1964). On the normed information rate of discrete random variables. In: *Trans. of the 3th Prague Conference on Information Theory.* Prague.

Read, T.R.C. & Cressie, N.A.C. (1988). *Goodness-of-fit statistics for discrete multivariate data.* New York etc., Springer.

Řehák, J. & Řeháková, B. (1973). Měření statistické závislosti nominálních znakú (Statistical measures for the dependence of nominal features). *Sociologiký časopis 4, 404-418.*

Řehák, J. & Řeháková, B. (1980). Analyse von Kontingenztafeln: Zwei Grundtypen von Aufgaben und das Vorzeichenschema. *Glottometrika 3, 1-28.*

Santner, T.J. & Duffy, D.E. (1989). *The statistical analysis of discrete data.*New York etc., Springer.

Schulz, K.-P. & Altmann, G. (1988). Lautliche Strukturierung von Spracheins-

heiten. *Glottometrika 9, 1-47.*
Silnitsky, G. (1993). Correlational system of verbal features in English and German. In: Köhler, T., Rieger, R.R. (eds.). *Contributions to Quantitative Linguistics.* Dordrecht etc., Kluwer, 409-420.
Taganov, I.N. & Turgumbaev, G.A. (1984). *Pričinnyj analiz složnych sistem* (Causal analysis of complex systems). Alma-Ata, Mektep.
Upton, G.J.G. (1978). *The analysis of cross-tabulated data.* New York, Wiley.
Yates, F. (1934). Contingency tables involving small numbers and the chi-square test. *J. of the Royal Statistical Society, Suppl.1, 217-235.*

4

An Attempt at Quantitative Analysis of the Style of Fiction

The present article analyzes the subsystem of statistical and stylistic parameters of texts of fiction with the aim of finding some regularities of interaction of these parameters. The methods of investigation are correlation analysis and factor analysis

1. Introduction and literary survey

In the last decades there has been a considerably increased interest in mathematics, including statistical methods, in the investigation of language, particularly in linguo-stylistics. Statistical methods have been taken into use in this area so widely that it is possible to speak about a new trend: *statistical stylistics*, also called *stylostatistics* or *stylometrics.* New methods of research make statistical stylistics a sensitive tool for the analysis of individual and functional styles (genres), including the style of fiction. The traditional qualitative methods will not disappear - they only get a good helpmate, a sufficiently precise instrument in the form of statistics. It is also useful to remember the apt remark of V.I. Perebejnos, "Statistical methods enable us not only to check up the correctness of our intuitive views about styles, but also to establish such regularities, which could not be found by other methods" (Perebejnos 1967).

Numerous concrete studies in the field of statistical stylistics have attracted attention, especially from the period of the sixties and the seventies, e.g. the works of J.B. Carroll (1960), G. Herdan (1960), C. Muller (1964), J. Thavenius (1966), H.H. Somers (1967), J. Mistrík (1967), F. Papp (1967), C.B. Williams (1970), N.B. Golovin (1971), M. Těšitelová (1972), J. Kraus (1972), B. Brainerd (1973), T. Zsilka (1974), R. Piotrowski (1975), C. Hassler-Göransson (1967), A. Šajkevič (1976), V. Moskovich & R. Caplan (1978) - to mention only some of the authors of that period. The collections *The Computer and the Literary Style* (1966), *Statistics and Style* (1969), *Mathematik und Dichtung* (1969), *Voprosy statističeskoj stilistiki* (1974), *The Computer in Literary and Linguistic Studies* (1976); the series *Prague Studies in Mathematical Linguistics* (1966-), *Studia metrica et poetica* (1976-), *Glottometrika* (1978-) with bibliographies, etc. are also noteworthy. The more recent works are very numerous;

among them we would mention the monographs of P. Thoiron (1980), P.M. Alekseev (1984), M. Těšitelová (1985), R. Köhler (1986), G. Altmann (1988), V.M. Arapov (1988), G. Ermolenko (1988), G. Martynenko (1988), L. Hřebíček (1992) and some special studies on text attribution (assignment of authorship) and textology, written by G. Kjetsaa et al. (1984, 1986), P. Vašák (1980, 1986), V. Levitsky (1989) and M. Marusenko (1990), where stylostatistical methods have been used.

2. Formulation of the problem and initial data

Statistical stylistic analysis of a text, including the analysis of the stylistic features of a work of fiction, is usually carried out by means of special stylistic parameters or "indicators", which are revealed by statistical research on grammatical phenomena and lexis as well as stylistic characteristics of the text studied. On the basis of stylostatistical parameters, the texts are compared with each other (or with a certain standard) and quantitative analysis is accompanied by the interpretation of content. The texts may differ on one parameter and be similar on another. The picture may become confusing and difficult to survey if a large number of stylostatistical parameters are analyzed. In such a case two questions arise: first, how to organize the work so that we could objectively establish the mutual connections between the parameters and the regularities of their interaction, and second, how to express all the information about the characteristics (parameters) in a more compact form, since it is natural to presume that a compact form reflects the more substantial, regular aspects of variation of the characteristics of different objects.

We make an attempt to solve this problem by means of correlation and factor analysis. The subject of our investigation is the author's monologue (narrative) in present-day Estonian fiction. As the original aim of our study was the comparative stylistic analysis of a small number of formally homogeneous texts (novels, published about 1970, non-conversational material only), seven samples - 5,000 words each - were taken from the works of seven authors (each sample consisted of five subsamples 1,000 words each from different places in the text). The numbers of the texts, their authors and titles are as follows:

1. AB - Aimée Beekman, Kartulikuljused (Potato-bells)
2. VG - Villem Gross, Pinginaabrid (Deskmates)
3. HK - Heino Kiik, Tondiöömaja (The Ghost's Nest)
4. JK - Jaan Kross, Kolme katku vahel I (Between Three Plagues)
5. LP - Lilli Promet, Primavera

6. VS - Veera Saar, Ukuaru (a placename)
7. HS - Herman Sergo, Põgenike laev (The Ship of the Fugitives).

Initial data for the present research are the frequencies of usage of parts of speech against the background of some formal lexico-statistical characteristics. A total of 12 parameters have been analyzed. We are interested in their stylistic background and their interdependence. Numerical data about the parameters observed are presented in Table 1.

Table 1
Initial data: stylostatistical parameters of text

Author and No. of the text / Parameters	AB 1	VG 2	HK 3	JK 4	LP 5	VS 6	HS 7
Frequency of nouns	35.9	30.4	34.8	32.8	30.4	26.0	36.8
Frequency of adjectives	5.4	7.4	5.2	7.4	5.5	4.9	5.8
Frequency of personal pronouns	3.6	4.0	2.5	3.7	7.1	6.8	1.5
Frequency of finite verbs	16.3	12.8	17.7	11.9	16.0	13.1	13.0
Frequency of independent adverbs	7.6	9.1	7.1	8.8	6.5	9.4	8.0
Frequency of pre-/ postpositions	3.7	2.5	2.4	3.9	2.9	3.7	3.6
Frequency of conjunctions	7.2	6.8	7.7	10.1	8.5	9.5	7.6
Frequency of content words (n + adj + v + adv)	80.2	77.4	79.2	75.0	73.8	71.7	79.4
Index of concentration	10.0	6.2	8.1	6.3	5.7	5.9	6.9
Entropy (H)	2.46	2.58	2.48	2.63	2.56	2.63	2.50
Index of frequent word-forms	11.0	12.0	12.9	13.8	14.0	15.2	10.6
Index of rare words	45.2	44.0	43.2	47.2	44.6	37.9	46.8

The choice of parts of speech relevant for our analysis was based on preliminary quantitative and qualitative analysis (Tuldava & Villup 1976). Particular attention was paid to the stylistic value of the features. The frequencies of the following grammatical and lexical classes were found: nouns, adjectives, personal pronouns, finite forms of verbs, so-called independent adverbs[1], postpositions and prepositions, conjunctions as well as the total of content words (i.e., nouns, adjectives, verbs and adverbs). A detailed analysis of these parameters would take up too much space, therefore we are going to deal briefly with only some of them.

It is customary to consider the frequency of *nouns* in the text (parameter 1) as indicator of the latent feature of "nominality" (i.e. nominal expression) of style. The gradation of nominality can be seen especially clearly if we compare different functional styles (genres). In Ukrainian (Tiščenko 1970), the average frequency of nouns in drama is 24.0%, in fiction 29.2%, in scientific and technical texts 39.4% and in sociopolitical texts 39.6%. However, considerable individual deviations from the average values may occur, particularly in fiction. It has also been ascertained that the frequencies of nouns in different texts by the same writer are rather homogeneous (Kožina 1972). This indicates that the parameter of noun usage characterizes to some extent the author's individuality. As our research (Tuldava & Villup 1976) has shown, the average frequency of nouns in author's monologue (narrative) in Estonian fiction (20 texts) is 31.7% with confidence limits <29.1%, 34.3%> at the 0.05 level. As can be seen from Table 1, some authors depart considerably from this level. For instance, the frequency of nouns in Text 6 (author VS) is 26.0%, while in Text 7 (HS) it is 36.8%.

Among other lexico-statistical characteristics the frequencies of *adjectives* (parameter 2) and finite forms of the *verb* (parameter 4) can be singled out as being important. A high percentage of adjectives in the text testifies to the "quality" of style, meaning that the author actively uses qualitative evaluation as a structural and stylistic element. A high percentage of verbs in the text, especially of the finite forms of the verb, is a marker of "activity" of style. Comparative analysis of functional styles allows us to state that frequent use of verbs as a general tendency is most characteristic of texts in drama and fiction. In Ukrainian, for example, the corresponding frequencies are 22.0 and 19.7%, while in scientific and sociopolitical texts the frequency of verbs is considerable lower, 13.9 and 11.1% respectively (Tiščenko 1970:217). The higher frequency

[1] In traditional Estonian grammar three kinds of adverbs are distinguished: independent (in the semantic and syntactic sense), modal and prefixal adverbs.

of verbs in the texts of fiction "testifies to their dynamism, to the fact that narration about events, the attitudes of the characters, etc., has a significant role in these texts" (Kožina 1972:140). Fluctuations in the frequency of verbs in the texts of fiction, in turn, reveal individual peculiarities of style in connection with the manner of description and narration.

The percentage of the *total of content words* (here: word forms of nouns, adjectives, verbs and adverbs) determines the degree of the text's "substantiality" (parameter 8). As can be seen from Table 1, in author's monologue of Estonian fiction this index fluctuates from 70 to 80%. The so-called index of concentration (parameter 9), first suggested by the French linguist P. Guiraud (1954), is a specifying characteristic which expresses the portion of cumulative frequency of the *50 most frequent content words* in the given text (the verb "to be" is not taken into account). A high value of the index shows that the author concentrates his attention on a relatively narrow range of words with full meaning which are used in the given text very often. This can testify to thematic compactness, to concentration on the main theme, in some cases also to stock phrases. If we turn to our study, we can see (Table 1) that Text 1 (AB) has the highest value of the index (10.0%), i.e. the 50 most frequent content words from the vocabulary of the given text cover as much as 10% of the text. As a comparison, in Text 5 (LP) the 50 most frequent content words cover only 5.7% of the text, i.e. the degree of concentration of autosemantic lexis is nearly half as small.

In general, a summary evaluation of the peculiarities of distribution of parts of speech can be expressed by the degree of *entropy* (parameter 10 in our study). Entropy is calculated according to the well-known formula of Shannon:

$$H = -\sum_i p_i \log_2 p_i$$

where H signifies the amount of entropy in bits, p_i is the probability of outcomes (in the present case the relative frequencies of different parts of speech in the text), log_2 is logarithm to the base 2. Entropy is understood as "indefiniteness", or "homogeneity of choice", which reaches its highest value if all probabilities in a system are equal. This means that the higher the value of H, the more homogeneous or similar the probabilities of occurrence of different parts of speech in the given text are. In the present study the correlation of entropy with other stylostatistic parameters (see section 3) is of primary interest.

The index of *frequent words* (parameter 11) has an outward similarity to the index of concentration (parameter 9) which was discussed above. The index expresses the cumulative frequency of the 10 most frequent word forms in the

given text in general. As the top part of any frequency dictionary of a text of fiction usually does not contain any content words (or there are only a few of them, most often proper names), the index of frequent words in its essence measures the concentration of function words, primarily conjunctions, prepositions, and some pronouns. Table 1 gives the values of the index in the form of percentages. For example, in Text 6 (VS) the ten most frequent word forms make up 15%. Text 7 (HS) has the lowest value of this index; in it the ten most frequent word forms cover only 10.6% of the text.

Parameter 12 or the index of *rare words* (also called the index of "exclusiveness") is known in stylostatistics as a formal measure of lexical variability and richness. The parameter expresses the proportion of word forms which have been used in a concrete text only with frequency 1 ("hapax legomena"). If a text includes a relatively large number of words with frequency 1, it may testify to the author's wish to find image-bearing expressions, choose rare or peculiar words, or avoid repetition of words. On the other hand, an insignificant share of words with frequency 1 is a sign of spontaneity of expression, great dependence on the content of thought, but sometimes a sign of banality and commonplaceness. Thus, the value of the index of rare words should be evaluated "not from the aesthetic position, but from the viewpoint of functional suitability and co-ordination" (Mistrík 1967:43). It is necessary to add that comparison of different texts on the basis of the index of rare words is possible only in the case of equal lengths of the texts or when using some mathematical transformation of the initial data (see, e.g., Tuldava 1993).

3. Correlation analysis

One way of gauging the association and interaction between the stylostatistical parameters under discussion is measuring the sample correlation between them according to Pearson's product-moment formula (see the article "On causal relations in language" in this issue). The values of the correlation coefficient (r) range from -1 to +1. A "plus" sign indicates that the two parameters tend to vary in the same direction; a "minus" sign means that one tends to increase as the other decreases. If there is no correlation, the correlation coefficient r = 0. The results calculated on the basis of the data of our study are given in Table 2 (to save space zeroes and decimal points as well as pluses have been omitted in the Table; for example the number 04 in the top left-hand corner should be read as +0.04). The absolute values of the correlation coefficient in our study must be above 0.66 to be statistically significant at the 0.10 level and 0.75 at the 0.05 level. A discussion of the principal results of our analysis follows.

Within the subgroup of parts of speech a strong positive correlation (r = +0.90) can be observed between the index of the total of content words (parameter 8) and the frequency of nouns (parameter 1). At the same time there is no significant correlation between the index of content words and the frequencies of verbs, adjectives and adverbs. From this a conclusion can be drawn that the feature "proportion of content words" in author's narrative depends first of all on the frequency of nouns in the text.

Table 2
Correlation between stylostatistical parameters

No.of param.	1	2	3	4	5	6	7	8	9	10	11	12	No.of param.
1	(n)	04	-85	31	-46	05	-46	90	67	-79	-82	76	1
2	(adj)		-24	-63	43	-03	03	04	-30	42	-15	54	2
3	(pron)			-01	03	03	44	-85	-49	57	77	-58	3
4	(v)				-81	-48	-40	-37	58	-72	-11	-08	4
5	(adv)					35	28	-33	-32	65	19	-33	5
6	(prep/postp)						55	20	08	24	04	11	6
7	(conj)							-76	-49	73	78	-18	7
8	(content words)								77	-86	-91	55	8
9	(concentration)									-81	-60	21	9
10	(entropy)										75	-32	10
11	(frequent word forms)											-61	11
12	(rare words)												12
Σ|r|	6.11	2.85	4.86	4.50	4.18	2.16	5.10	6.54	5.32	6.87	5.73	4.27	

There is a strong negative correlation (r = -0.87) between the frequency of nouns (parameter 1) and the frequency of personal pronouns. The same conclusion can be drawn from the analysis of texts of fiction in Russian (Golovin 1971) and texts of various functional styles in Latvian (Jakubaitis & Stūrite 1974). The tendency of repulsion between nouns and personal pronouns is explained by the fact that pronouns can replace nouns and thus become their competitors in the process of speech generation.

Sometimes an opinion has been expressed that nouns and verbs are interdependent and competitive parts of speech in a text. Our study shows, however, that - at least in the type of author's narrative studied by us - there is no significant correlation between the frequency of nouns (parameter 1) and that of finite

forms of verbs (parameter 4). Positing opposition between adjective and verb seems to be more justified. The data of our study reveal a tendency to an inverse relation between the frequencies of adjectives and verbs (r = -0.63). The same has been found in statistical investigations of many other languages (especially in texts of fiction), and the ratio between the frequencies of verbs and adjectives (or adjectives and verbs), called "Busemann's Coefficient", has been used in many investigations as an important feature of style (see, e.g. Antosch 1969; Fischer 1969; for a critical analysis, see Altmann 1978). Here, undoubtedly, the qualitative connection between the phenomena is revealed, and namely the fact that the adjective (epithet), although it is attributive in its syntactical character, is always predicative in its semantic nature. Like the verb, the adjective, too, reports something new about the object under discussion and consequently both parts of speech have the same or similar semantic functions and can compete on the concrete semantic level of text formation.

The data of our research do not reveal any significant correlation between the frequencies of nouns (parameter 1) and adjectives (parameter 2). An analogical study on material from the Russian language gave the same results (Golovin 1970). We can draw the conclusion that, on the whole, the frequency of occurrence of adjectives in the texts of fiction under review does not depend on the frequent use of nouns, but it manifests itself as an individual feature of style.

Turning to the correlations in the subgroup of formal lexico-statistical characteristics (parameters 9-12) and their connections with the subgroup of parts of speech, the special role of entropy (parameter 10) should be mentioned. This parameter seems to have the central place in the system of correlations of our 12 parameters. First, according to the sum of inter-correlations (6.87) it occupies first place among the parameters under discussion (see Table 2), i.e. on average, entropy correlates best of all with the other parameters. Second, entropy has approximately the same number of positive and negative correlations. Direct connection can be seen between entropy and the frequency of conjunctions, pronouns and adverbs as well as with the index of frequent word forms. We have already mentioned that entropy, in the present case calculated on the basis of distribution of frequencies of parts of speech, expresses the degree of homogeneity (uniformity) of distribution of different parts of speech in the text. Apparently, such equilibrium of parts of speech in the text depends first of all on the abundance of structural words. Negative correlation can be observed between entropy and the frequencies of nouns and verbs, also with the index of concentration of content words (parameter 9), in the latter case the correlation coefficient reaches -0.81. Thus, the measure of entropy as a characteristic of information theory has good diagnostic qualities and can - at least in the analysis of Estonian texts - be used as a formal parameter predicting the concentration of structural words (in the case of high values of entropy) or the concentration of content words (in the case of low values of entropy).

The index of concentration of content words (parameter 9) has moderate positive correlation with the frequencies of nouns and verbs. This reflects the fact that among the 50 most frequent content words nouns and verbs really dominate. The moderate negative correlation with the index of frequent words (parameter 11) shows that with the concentration of content words the share of function words with high frequency correspondingly decreases. Such a phenomenon is specific to Estonian and other languages of synthetic character. In the languages which use articles and prepositions as permanent companions of nouns, the indexes of concentration of content words and concentration of function words can correlate with each other. In Estonian, however, there are no articles while prepositions/postpositions play only a minor role with the developed system of cases.

Last, we should mention a significant correlation (r = + 0.76) between the index of rare words (parameter 12) and the frequency of nouns (parameter 1). The tendency towards a positive correlation between the index of rare words and the frequency of adjectives (parameter 2) can also be noted, while correlation between the index of rare words and the frequency of verbs (parameter 4) is missing. From this we can draw the conclusion that, in our case, general variability of lexis is realized mostly by nouns and adjectives, while verbs have a neutral role here.

As a result of correlation analysis some significant mutual connections (or lack thereof) between the stylostatistical parameters under discussion were revealed. The next step should be the systematization of correlations and development of hypotheses about the causal relations behind the correlations found between the parameters. Systematization and interpretation of the material can be attempted on the basis of thorough logical analysis of correlations with the help of some additional measures, e.g. the methods of partial and multiple correlation (cf. the article "On causal relations in language" in this issue). Another rational and economic way of systematic treatment of connections between the parameters is factorization of the available correlation matrix, i.e. the use of factor analysis.

4. Factor analysis

Factor analysis is a method meant for "compressing" the information contained in the correlation matrix. It is supposed that correlations between pairs of meas-

ured variables can be explained by the connections of the measured variables to a small number of non-measurable, but meaningful, variables, called *factors*. By correlation analysis we discovered that many parameters are interrelated and most probably express certain underlying characteristics of the phenomena. Our aim is to identify the underlying factors and define them as functions of the measured variables. Under such conditions it is necessary to group the parameters in such a way that their groups would reflect the essential features of the phenomena studied to the greatest degree. Factor analysis is exactly what is wanted to show how separate features which behave in a similar way are united into groups and to reveal the factors influencing the formation of these groups.

Factor analysis is based on the hypothesis that the parameters observed or measured are but external characteristics of the object or phenomenon studied, and that there actually exists a small number of internal (latent, not directly observable and measurable) parameters or qualities which determine the values of the observable parameters.

Numerous experimental studies, particularly in processing of psychological, sociological and other data have shown that the factors defined can as a rule be well interpreted as some basic internal characteristics of the objects investigated. It should be acknowledged that for several reasons, particularly because of a certain vagueness characteristic of interpretation and analysis of various models of factors and also because of the large amount of labor-consuming computation, factor analysis has for a long time been neglected in linguistic investigation. However, in recent years, as the opportunities to use computers have widened, interest in factor analysis has grown considerably and a number of successful attempts at factor analysis of linguistic data have been undertaken (see, e.g., Saukkonen 1990).

In our research factor analysis was carried out according to the method of common factors (cf. Harman 1976) at the Computing Center of Tartu University. The results of the calculations have been given in Table 3 which presents the so-called factor matrix, which has been obtained after rotation of the factors according to the "varimax" method. The table presents an array of loadings of four factors which characterize twelve initial parameters. For example, the coefficient 0.85 in the first column of the first line (Table 3) is the loading of parameter 1 (frequency of nouns), i.e. the positive correlation between the given factor and the parameter. The higher this correlation, the more the parameter is "filled" with the given factor and the more it can be considered the loading of this factor. The sum of squares of the factor loadings in each line is called "communality" and is signified by h^2. The maximum value of communality is 1. The communality of the given parameter shows to what extent the given parameter is described by the given factors. As can be seen from Table 3, the

majority of communalities of parameters have a value approaching one. Consequently, these four factors almost completely describe every single parameter. The sum of squares of factor loadings of each column expresses the contribution of the factor which can be transformed into a percentage[2]. In the last line of Table 3 we see that Factor 1 carries 42.2% of the variance, Factor 2 - 21.7%, Factor 3 - 14.0%, and Factor 4 - 11.5%. All four factors together explain 89.4% of the total variation, i.e. nearly all the information about stylostatistical parameters is contained in these four factors. Such compression of initial information can be considered wholly satisfactory. In fact, a number of additional factors can be elicited, which jointly cover the remaining percentage of total variation. But as the descriptive power of each of them separately is too small they can be left out of consideration (the computer program used by us gives only factors whose share in total variation exceeds 5%).

The decisive aspect in factor analysis is *interpretation* of the calculated factors, i.e. transition from quantitative to qualitative analysis. Interpretation is based on careful analysis of the system of coefficients of the factor matrix. As is known, the factor loadings (F_{ij}) indicate the correlation of parameters with the corresponding factors and in this way provide the basis for the content analysis of the factors. Comparing the factor loadings in the columns of the factor matrix, simultaneously taking into consideration the character of the parameters, we can posit certain hypotheses about the nature of the factors obtained. The analysis of factors usually ends with giving them a corresponding designation (name, label). In the interpretation and designation of factors only those parameters which have a relatively big loading must be taken into account. It is necessary to keep in mind that factor loadings, calculated on the basis of sample data, reflect the mutual connections only with a certain degree of approximation. In order to find how substantially the factor loadings differ from zero, we may calculate the standard deviation, for example by the approximation method of Halsinger and Harman (see Harman 1976:450). According to the data of our study the standard deviation is 0.3. We have decided to consider "big" the factor loadings whose absolute value is twice the standard deviation, i.e. 0.6.

Now let us proceed to the analysis of separate factors. First of all, we shall turn to the column of loadings of the first factor (see Table 3).

---

[2] The contribution of the factor is calculated according to the formulas:

$V = \sum_{i=1}^{m} F_j^2$ where V is the sum of columns of squared factor loadings F_j (the so-called eigenvalue);

$V' = (\sum_{i=1}^{m} F_j^2/m)\cdot 100$ - percent of variance among all the variables that is accounted for by the factor (*total variation).*

Table 3
Factor score matrix of stylostatistical parameters (varimax rotation)

Parameters	Factor loadings				Communality (h^2)
	F_1	F_2	F_3	F_4	
1 Nouns	0.85	0.32	0.41	0.13	1.00
2 Adjectives	-0.02	-0.58	0.65	-0.14	0.78
3 Pronouns	-0.80	0.10	-0.31	0.01	0.75
4 Verbs	0.22	0.87	-0.27	-0.29	0.96
5 Adverbs	-0.13	-0.93	-0.21	0.18	0.96
6 Pre-/postpositions	0.03	-0.23	-0.01	0.82	0.73
7 Conjunctions	-0.67	-0.09	0.08	0.66	0.90
8 Content words	0.96	0.16	0.16	-0.21	1.00
9 Concentration	0.77	0.35	-0.21	0.12	0.77
10 Entropy	-0.81	-0.57	0.10	0.17	1.00
11 Frequent word forms	-0.91	0.03	-0.23	0.15	0.90
12 Rare words	0.44	0.09	0.86	0.11	0.95
Eigenvalue (V)	5.07	2.61	1.68	1.38	-
% Total variation (V^t)	42.2	21.7	14.0	11.5	89.4 (%)

Factor 1. The table shows that the factor is characterized by both positive and negative factor loadings of different parameters ("bipolarity" of the factor). Parameters 1, 8 and 9 (frequency of nouns, total frequency of content words, and concentration of 50 most frequent content words) have big positive loadings. Parameters 3, 7, 10 and 11 (frequencies of personal pronouns and conjunctions, degree of entropy, concentration of function words) have negative values of coefficients. The general factor underlying such a complex of phenomena is likely to be connected with the character of concentration and dispersion in the text of two "opposite" groups of words - content and function words. As mentioned above, the parameter "frequency of content words" was most of all correlated with the parameter "frequency of nouns"; then from the viewpoint of linguistic analysis we might speak about the factor of "nominality" which, in the case of negative values of factor loadings, also has another polarity - "non-nominality". But one could also proceed further. In factor analysis, we should, as much as possible, try to find the essence of the phenomenon, the concealed regularities of the connections discovered by empirical research. In a sense the

factors can be regarded as *causes*, and the parameters observed (measured on objects) as *results*. Therefore we are interested in the conditions of creating a text that contains a group of content words, especially nouns, with high frequency, and, on the other hand, relatively few function words. Such a text is characterized by compactness of subject matter (which brings about concentration of content words), while the share of function words (including pronouns) decreases. At the same time the value of entropy (parameter 10) registers a relative imbalance (heterogeneity) in the distribution of parts of speech (on account of the concentration of content words) and, consequently, its value decreases (as compared to the average entropy of the system).

Proceeding from these preconditions, the first factor could be called the *factor of lexical concentration*. Concentration of lexis in the given sense depends on both the individuality of the author and on the choice of the subject matter and other extralinguistic reasons.

Factor 2. This is also a bipolar factor having substantial loadings with opposite signs. It has a big positive loading on parameter 4 (the share of finite forms of verbs in the text). Parameters 2, 5 and 10 (frequencies of adjectives and adverbs, entropy) have significant negative loadings. Proceeding from the nature of these parameters, this factor can be characterized as expressing two opposite stylistic categories: activity versus qualitativeness. Considering the higher value of factor loading on parameter 4 (frequency of finite verbs), this factor can be called the *factor of activity*. The loadings on this factor express the dominant role of verbal "dynamism", while significant negative loadings testify to the repulsion of non-verbs in the presentation of the subject-matter.

Factor 3. Parameter 12 (index of rare words) dominates here with factor loading as high as 0.86. Other parameters are neutral on this factor; only parameter 1 (frequency of nouns) reveals a slight tendency towards positive correlation with it. As the dominant index of rare words reflects the variety of vocabulary, the third factor could be called the *factor of lexical variety, or "richness"*, where the occurrence of rare nouns and adjectives plays the principal role.

Factor 4. This factor is characterized by significant loadings on parameters 6 and 7 (indices of occurrence of prepositions/postpositions and conjunctions in the text). The factor loadings are 0.82 and 0.66, correspondingly. What are the common characteristics of these two parameters? Let us assume that there is a certain analyticality of expression, or a string for concretization of spatial and temporal relations in speech. The contribution of this *factor of analyticality* to

the total communality is not large - only 11.5%. But still this interesting stylostatistical index ought to be studied; its real nature and meaning can be found only by additional qualitative analysis (cf. Tuldava 1978).

Thus, by factorization of initial data we have obtained the four main factors which determine the interrelations of elements in the given subsystem of stylostatistical parameters. The results of the analysis can be represented graphically, which gives a convenient visual spatial scheme for the interpretation of the material. Every parameter can be interpreted geometrically as a point in a space which has as many dimensions as the number of factors picked out. Practically, however, such spatial models can be constructed in two- or three-dimensional cases only. As an illustration, we give the graphic presentation of Factors 1 and 2, accounting together for 64% of the total variance (see table 3). In Figure 1, the coordinate axes represent Factors 1 and 2, and the coordinates of points (circles) are determined by the factor loadings of the stylostatistical parameters. Location of the points shows the grouping of the parameters into clusters, e.g. parameters 1, 8, 9 are grouped near axis F_1 with big positive factor loadings of Factor 1 and with relatively small loadings of factor 2; etc. Parameters 6 and 12 fall into the "neutral" zone with insignificant values of loadings of the first two factors. Thus, graphic presentation allows us to distinguish clearly between different groups (clusters) of parameters obtained on the basis of classification according to the two most important factors and to define the role of each parameter in the interpretation of factors.

Our computer program for factor analysis also allows us to determine the role of factors in classification of objects studied, in our case the texts under examination. The results of the calculations are given in Table 4 and presented graphically in Figure 2 (for the first two factors). The object values of the factors f_i are given in a standardized form, which shows the multiple of standard deviation from the mean value. It seems to be appropriate to consider the value $f_i = \pm 1$ to be sufficient to speak about the significant role of a factor with regard to the given object (text). For example, texts 1 and 7 have positive values of the first factor (1.3 and 1.1 correspondingly). This means that the first factor (factor of lexical concentration) is most strikingly expressed in these texts (authors A. Beekman and H. Sergo). This factor, as noted above, is closely correlated with parameters 1, 8 and 9 (frequencies of nouns and content words, concentration of content words). Further it can be seen in Table 4 that the authors of texts 5 and 6 (L. Promet and V. Saar - both women) have significant negative values of the first factor (-1.2 and -1.1). Their texts, studied in our present research, as a matter of fact, reveal low lexical concentration: from Table 1 we can see, on the one hand, a small proportion of nouns and a low

concentration of content words; on the other hand, significant concentration of function words (parameter 11) and relatively large shares of personal pronouns (parameter 3) and conjunctions (parameter 7).

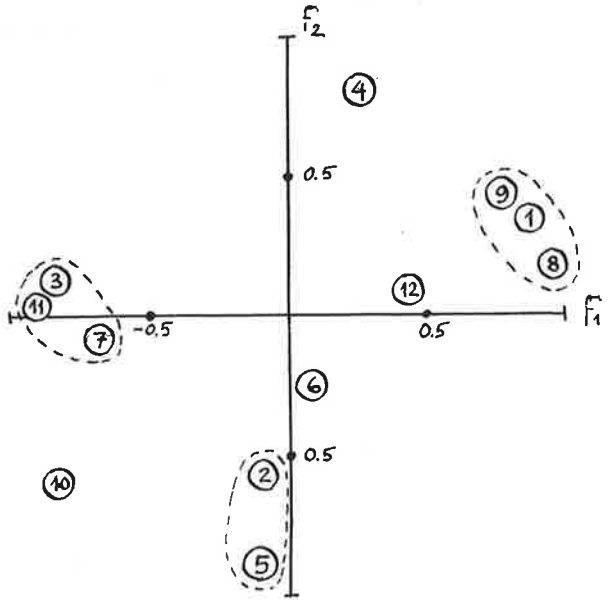

Figure 1. The presentation of 12 stylostatistical parameters in the space of two factor axes

Table 4
Object values of factors for seven texts

Texts (authors)	Factor values				Modified values			
	f_1	f_2	f_3	f_4	f_1'	f_2'	f_3'	f_4'
1. A. Beekman (AB)	1.3	0.3	-0.9	-0.5	+	0	0	0
2. V. Gross (VG)	0.1	-1.5	0.4	-1.8	0	-	0	-
3. H. Kiik (HK)	0.4	1.8	-0.6	0.9	0	+	0	0
4. J. Kross (JK)	-0.4	-0.4	1.0	1.3	0	0	+	+
5. L. Promet (LP)	-1.2	1.0	0.8	-1.4	-	+	0	-
6. V. Saar (VS)	-1.1	-0.6	-1.7	1.0	-	0	-	+
7. H. Sergo (HS)	1.1	-0.6	1.1	0.5	+	0	+	0

This way we could also examine and analyze the values of the remaining three factors expressing some substantive quantitative characteristics of the style of the texts. For better clarity we designate the significant values of f_{ij} with plus and minus signs and the insignificant values with zero and so get the modified values f_{ij}' (see Table 4). It can be seen that in the present case there are no exactly coinciding structures (series of plus and minus signs and zeroes) and all the texts under observation can be clearly differentiated on the basis of factor values. This kind of analysis can be used as one of the subsidiary methods of attribution of texts to one or another author. On the whole, the analysis of factor values of texts enables us to group the texts, measure their mutual closeness (or distance), and interpret the deviations from the mean value ("norm") of the given genre or subgenre. It should be noted that the analysis has not been based on the 12 initial factors. This became possible thanks to the significant reduction of dimension of the initial parametrical space with the help of the method of factor analysis.

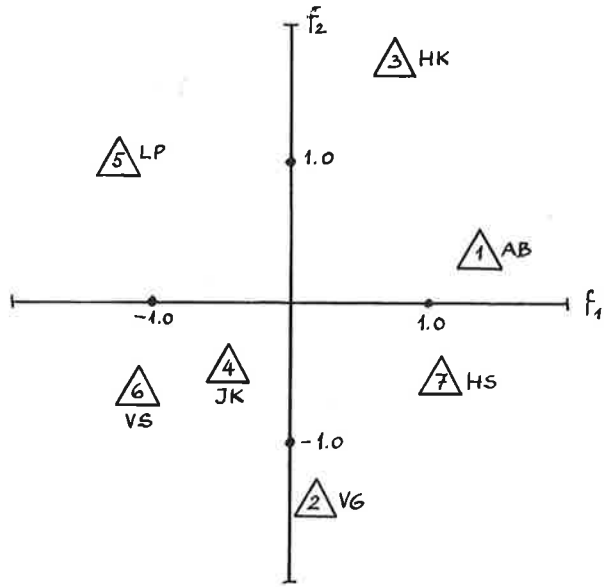

Figure 2. The distribution of texts in the space of two factor axes

5. Concluding remarks

The present study is an attempt at stylostatistical analysis of the general and the individual within one style - the author's narrative in contemporary Estonian fiction. As a basis of the analysis we have taken a series of lexicostatistical parameters, which are not observed in isolation, but from the viewpoint of correlation between them, and furthermore in connection with the formalized procedure of grouping the parameters into more general and abstract categories - factors. The main task here was the elaboration of a procedure which allows us to get an adequate classification system. Such analysis can be developed further, connecting the stylostatistical features with non-formal, qualitative characteristics of texts (subject matter, speed and acuteness of the unfolding of plots, etc.) as well as with evaluation of latent style characteristics (emotionality, intellectuality, verbosity, laconicism, expressiveness, etc.) obtained by questioning of informants and experts. (See the next article in this issue.)

References

Alekseev, P.M. (1984). *Statistische Lexikographie.* Bochum, Brockmeyer.

Altmann, G. (1978). Zur Anwendung der Quotiente in der Textanalyse. *Glottometrika 1, 91-106.*

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

Arapov, M.V. (1988). *Kvantitativnaja lingvistika (Quantitative linguistics).* Moscow, Nauka.

Antosch, P. (1969). The diagnosis of literary style with the verb-adjective ratio. In: Doležel, L. & Bailey, R.W. (eds.), *Statistics and Style: 57-67.* New York, Elsevier.

Brainerd, B. (1973). On the distinction between a novel and a romance. A discriminant analysis. *Computer and the Humanities 7, 259-270.*

Carroll, J.B. (1960). Vectors of prose style. In: Sebeok, T.A. (ed.), *Style in Language: 283-292.* Cambridge, Mass.

Doležel, L., Bailey, R.W. (eds.) (1969). *Statistics and Style.* New York, Elsevier.

Ermolenko, G.V. (1988). *Anonymnye proizvedenija i ich avtory. Na materiale russkich tekstov vtoroj poloviny XIX - načala XX v. (Anonymous works of literature and their authors. On the material of Russian texts of the second half of 19th and the beginning of 20th century).* Minsk, Universitetskoe Izdatel'stvo.

Fischer, H. (1969). Entwicklung und Beurteilung des Stils. In: Kreuzer, H. (ed.), *Mathematik und Dichtung, 3. Aufl.: 171-183.* München, Nymphenburger.

Glottometrika (1978 -). Ed. by G. Altmann et al. Bochum, Brockmeyer.

Golovin, B.N. (1971). *Jazyk i statistika (Language and statistics).* Moscow, Prosveščenie.

Golovin, B.N. (ed.) (1974). *Voprosy statističeskoj stilistiki (Problems of statistical stylistics).* Kiev, Naukova dumka.

Guiraud, P. (1954). *Les caractères statistiques de vocabulaire. Essai et méthodologie.* Paris, P.U.F.

Harman, H.H. (1976[3]). *Modern factor analysis.* Chicago, University of Chicago Press.

Hassler-Göransson, C. (1976). *Fyrtio författare i statistik belysning (Statistical analysis of forty authors).* Stockholm, Skriptor.

Herdan, G. (1960). *Type-token mathematics.* The Hague, Mouton.

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

Jakubaitis, T.A. & Stūrite, B.A. (1974). O statističeskoj odnorodnosti tekstov (On statistical homogeneity of texts). In: Golovin, B.N. (ed.), *Voprosy statističeskoj stilistiki: 43-54.* Kiev, Naukova dumka.

Jones, A. & Churchhouse, R.F. (eds.) (1976). *The Computer in Literary and Linguistic Studies.* Cardiff, The University of Wales Press.

Kjetsaa, G., Gustavsson, S., Beckman, B., Gil, S. (1984). *The authorship of the Quiet Don.* Oslo, Solum Forlag.

Kjetsaa, G. (1986). *Prinadležnost' Dostoevskomu (Attributed to Dostoevsky: The problem of attributing to Dostoevsky anonymous articles in* Time *and* Epoch) Oslo, Solum Forlag.

Kožina, M.N. (1972). *O rečevoj sistemnosti naučnogo stilja sravnitel'no s nekotorymi drugimi (About the systematic character of the language of scientific style in comparison with some other styles).* Perm', Izd. Universiteta.

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

Kraus, J. (1972). On the stylistical-semantic analysis of adjectives in journalistic style (A quantitative approach). *Prague Studies in Mathematical Linguistics 4,95-106.*

Leed, J. (ed.) (1966). *The Computer and Literary Style.* Kent, Ohio, Kent State University Press.

Levitskij, V.V. (1989). *Statističeskoe izučenie leksičeskoj semantiki (Statistical investigation of lexical semantics).* Kiev, UMK.

Martynenko, G. (1988). *Osnovy stilemetrii (Foundations of stylometrics).* Leningrad, Izd. Leningradskogo Universiteta.

Marusenko, M.A. (1990). *Atribucija anonimnych i psevdonimnych literaturnych proizvedenij metodami raspoznavanija obrazov (The attribution of anonymous and pseudonymous literary works by means of the method of the recognition of images).* Leningrad, Izd. Leningradskogo Universiteta.

Kreuzer, H. (ed.) (1969[3]). *Mathematik und Dichtung.* München, Nymphenburger.

Mistrík, J. (1967). Matematiko-statističeskie metody v stilistike (Mathematical-statistical methods in stylistics). *Voprosy jazykoznanija 3, 42-52.*

Moskovich, W., Caplan, R. (1978). Distributive-statistical text analysis: A new tool for semantic and stylistic research. *Glottometrika 1, 107-153.*

Muller, C. (1964). *Essai de statistique lexicale. L'illusion comique de Pierre Corneille.* Paris, Klincksieck.

Papp, F. (1967). O nekotorych količestvennych charakteristikach slovarnogo sostava jazyka (On some quantitative characteristics of the vocabulary of a language). *Slavica (Debrecen) 7, 51-58.*

Perebejnos, V.I. (ed.) (1967). *Statistični parametri stiliv (Statistical parameters of styles).* Kiev, Naukova dumka.

Piotrowski, R.G. (1975). *Tekst, mašina, čelovek (Text, machine, Man).* Leningrad, Nauka. (Translation: *Text, Computer, Mensch.* Bochum, Brockmeyer 1984).

Saukkonen, P. (1990). Interpreting textual dimensions through factor analysis: Grammatical structures as indicators of textual dimensions. *Glottometrika 11, 155-171.*

Somers, H.H. (1967). *Analyse statistique du style.* Louvain, Nauwelaerts.

Studia metrica et poetica (1976 -). Ed. by J. Põldmäe et al. Tartu University Press.

Šajkevič, A.Ja. (1976). Vydelenie klassov slov i paradigm posredstvom distributivno-statističeskogo metoda (Determination of word classes and paradigms by means of the distributive-statistical method). *Prikladnaja lingvistika 18, 96-134.*

Těšitelová, M. (1972). On the so-called vocabulary richness. *Prague Studies in Mathematical Linguistics 3, 103-120.*

Těšitelová, M. and coll. (1985). *Kvantitativní charakteristiky současné češtiny (Quantitative characteristics of the present-day Czech language).* Praha, Academia.

Thavenius, J. (1966). Kvantitativa metoder i stilistiken (Quantitative methods in stylistics). In: Hallberg, P. et al (eds.), *Litteraturvetenskap: 37-62.* Stockholm.

Thoiron, P. (1980). *Dynamisme du texte et stylostatistique: Elaboration des index et de la concordance pour Alice's Adventures in Wonderland. Problèmes, méthodes, analyse statistiques de quelques données.* Geneva, Slatkine.

Tiščenko, V. (1970). Častota častin movi v riznich funkcional'nych stiljach sučasnoj ukrainskoj movi (Frequency of parts of speech in different functional styles of present-day Ukrainian). In: *Pitannja strukturnoj leksikologii.* Kiev, Naukova dunka.

Tuldava, J. (1978). Sõnavormide esinemus eestikeelses tekstis (The occurrence of word forms in Estonian texts). *Acta et Commentationes Universitatis Tartuensis 446, 107-126.*

Tuldava, J. (1993). *Probleme und Methoden einer quantitativen system- und textbezogenen Wortschatzforschung.* Hagen, Rottmann (in press).

Tuldava, J. & Villup, A. (1976). Sõnaliikide sagedusest ilukirjandusproosa autorikõnes (Statistical analysis of the parts of speech in Estonian fiction). *Acta et Commentationes Universitatis Tartuensis 377, 61-106.*

Vašák, P. (1980). *Metody určování autorství (Methods of establishing authorship).* Praha, Academia.

Williams, C.B. (1970). *Style and vocabulary: Numerical studies.* London, Griffin.

Zsilka, T. (1974). *Stilisztika és statisztika (Stylistics and statistics).* Budapest, Akadémiai Kiadó.

5

A Comparison of Subjective and Objective Characteristics of Style

The article gives a survey of the results of a psychometric experiment where informants (experts) evaluated the stylistic qualities of works of fiction. On the basis of correlation and factor analyses an attempt was made to establish the interrelations between subjective and objective characteristics of style.

1. Expert evaluations of style

When carrying out the experiment on style evaluation by qualified informants (experts), we proceeded from the assumption that there are certain generally acknowledged standards according to which it is possible to evaluate qualitative features of fictional style. This could concern primarily the evaluation of such "normative literary stylistic features of general speech" (Kožina 1972:99) as clarity, simplicity, expressiveness, laconicism, etc. For our research we chose 15 characteristics of style (see Table 1) which are most often used by literary critics in their articles to express opinions and judgements about works of fiction. Ten experts (qualified professional critics and philologists) were asked to evaluate the "intensity" of these characteristics of style in author's narrative in seven works of present-day Estonian fiction (novels published about 1970, non-conversational material only; for the full list of texts see the article "An attempt at quantitative analysis of the style of fiction" in this issue).

The experts received questionnaires and were asked to evaluate each novel according to the method of "successive intervals" (Frumkina & Vasilevič 1971). A scale of five categories was used, arranged in order of intensity of the characteristics (features of style) from the smallest to the greatest. Every category was given a number (score, mark) which corresponded to its intensity:

1 - very weak
2 - weak
3 - average
4 - strong
5 - very strong.

The experts evaluated the style of the texts independently, not knowing the judgements of the other experts. They were free to reason about the meaning of the characteristics proposed for analysis. When asked afterwards what their motives of judgement were, most of them spoke of a "general impression" underlying the process of evaluation of the concrete stylistic features, adding that in some cases they considered the features to be based on binary opposition, for instance, concrete (-abstract), exact (-vague), dynamic (-static). Some of the experts explained that their decision depended on the analysis of the language of the texts, for instance "emotionality" was connected with frequent use of words denoting emotion, expressing feelings as against matter-of-fact referential words used in narration. An important role was assigned to the concrete subject matter treated in the book, so the characteristic "dynamism" was brought into relation to combat in combination with corresponding lexical means. One expert pointed out that the proposed stylistic features should be divided into two separate groups: subjective and objective characteristics, the first ones being objects of pure personal estimation and the second ones expressing general stylistic qualities (such as "popularity of language", "verbosity, "fluency of language").[1]

The results of the expert judgements were presented in the form of tables separately for each characteristic and for each text. As an example we present expert estimations of the intensity of stylistic features in Text 1 (Table 1). For a more compressed presentation of the data received, usually the average values of estimates for the given characteristic or text were calculated. However, it is important to stress that on an ordinal scale the intervals need not be equal, and the experts' evaluations should be regarded as rank values. In such a case, in order to reveal the central tendency of evaluation, not the arithmetic mean is calculated but rather the median (*Md*), i.e. the value of the estimate which is situated in the middle of the ordered list of values (arranged from the smallest value to the largest, or vice versa). With an even number of values, as in the present case when we have the evaluations from 10 experts, the median is calculated as the average value of the two central values. For example, the evaluations of "emotionality" of Text 1 are distributed as follows (see Table 1): 3 4 4 4 4 5 5 5 5 5. In this ranged series the median is situated between "4" and "5", i.e. $Md = (4+5)/2 = 4.5$.

Comparing the *Md* values with the corresponding arithmetical means ($\bar{x}$) in Table 1 we see that they do not differ much from each other. Only one obvious difference can be noted, namely in the evaluation of the characteristic "intellectuality" ($Md = 4$, $\bar{x} = 3.3$). We see, e.g., that one expert (No. 3) has assigned the mark "1" to this characteristic, whereas most of the other experts have given the mark "4" to the same characteristic. An analysis of such extreme values can be of interest when investigating expert judgement in detail (e.g., the problem of competency or originality of their judgements), but in our case we are foremost interested in obtaining average values as indicators of *group decision,* assessing that the experts were highly qualified professionals. In principle, the median should be preferred as it is not affected by the size of extreme values (occasional or not occasional). Apart from this, we can use medians in case of doubts over the correct distributional assumption.

[1] Cf. Doležel (1969) who distinguished between three classes of stylistic features: subjective, objective and subjective-objective features (cf. also Enkvist 1974; Grotjahn 1979).

Table 1
Expert estimations on the intensity of stylistic features (characteristics) in Novel 1 (AB)

Chracteristics (i)	Experts (j) 1 2 3 4 5 6 7 8 9 10	Σx_i	$\bar{x}_i$	s_x	v_x	Md
1. Emotionality	5 4 5 5 5 4 5 4 3 4	44	4.4	0.70	0.16	4.5
2. Intellectuality	2 3 1 4 4 4 4 4 3 4	33	3.3	1.06	0.32	4.0
3. Concreteness	5 4 3 3 3 3 4 3 4 4	36	3.6	0.70	0.19	3.5
4. Thoughtfulness	3 3 1 2 3 4 4 4 2 3	29	2.9	0.99	0.34	3.0
5. Eventfulness	3 4 4 4 2 3 4 4 3 3	34	3.4	0.70	0.21	3.5
6. Expressiveness	4 4 3 4 5 5 5 3 3 5	41	4.1	0.88	0.21	4.0
7. Saturation with details	5 4 5 4 5 5 5 3 4 5	45	4.5	0.71	0.16	5.0
8. Dynamism	3 3 2 3 4 3 4 3 3 3	31	3.1	0.57	0.18	3.0
9. Verbosity	3 4 2 3 3 2 2 3 3 2	27	2.7	0.67	0.25	3.0
10. Clarity	3 3 5 5 4 5 5 3 3 4	40	4.0	0.94	0.24	4.0
11. Readability	3 4 5 5 4 4 4 3 3 4	39	3.9	0.74	0.19	4.0
12. Laconicism	2 1 4 3 3 3 3 3 2 4	28	2.8	0.92	0.33	3.0
13. Popularity of language	3 2 1 2 2 2 2 2 2 2	20	2.0	0.47	0.24	2.0
14. Fluency of language	3 4 5 5 4 3 4 3 3 3	37	3.7	0.84	0.23	3.5
15. Artistic perfection	4 4 2 3 4 4 5 3 3 4	36	3.6	0.84	0.23	4.0
Total	51 48 55 60 44 / 51 55 54 48 54	520	3.47	-	-	3.5

x_{ij} - data on scores (marks); $\bar{x}$ - arithmetic means; s_x - standard deviations; $v_x = s_x/\bar{x}$ - coeefficient of variation, Md - median

2. Test of concordance

In a study of expert judgement, when group decision is the aim of the investigation, it is recommended to establish the degree of concordance (agreement, conformity) within the group of experts. The appropriate procedure to be used is the distribution-free (non-parametric) test of concordance, proposed by M. Kendall (1955).[2] The test is closely related to the method of rank correlation analysis. We start with assigning rank orders to estimates (see Tables 1 and 2). Each column will be ranked independently of the other columns, so that there will be m sets of ranks (here: m = 10), each set ranging from 1 to n (n = 15). If two or more estimates (marks) in a column have the same value, we give such estimates an average rank. For instance, in the first column of Table 1 the two smallest estimates ("2") share the 1st and the 2nd place in the rank and each is given a rank value of (1+2)/2 = 1.5. The three largest estimates ("5") share the 13th, 14th and 15th place and each is given a rank value of (13+14+15)/3 = 14.

Kendall's coefficient of concordance is calculated using the following formula:

$$(1) \qquad W = \frac{12 \sum_{i=1}^{n} d_i^2}{m^2(n^3 - n) - m\sum_{j=1}^{m} T_j}$$

where W is the coefficient of concordance;

$\sum_{i=1}^{n} d_i^2$ denotes the sum of squares of the deviations;

$$d_i = \sum_{j=1}^{m} R_{ij} - \frac{\sum_j \sum_i R_{ij}}{n}; \quad i = 1,2,...,n; \quad j = 1,2,...,m$$

R_{ij} - rank values;
m - number of experts;
n - number of observations (style characteristics);

[2] A preliminary estimation of concordance may be made on the basis of standard deviations and coefficients of variation of the arithmetic means. As can be seen from Table 1, the values of the coefficients of variation (v_x) do not exceed the magnitude 0.35, i.e. 35%, which can be regarded as normal dispersion in a sample.

$m\sum_{j=1}^{m} T_j$ correction for ties;

$$T_j = \sum_{k=1}^{z_j} (t_k^3 - t_k), \; k = 1,2,...,z;$$

Z_j - number of groups of equal ranks in column j;
t_k - number of equal ranks in group k.

If there are no tied ranks, then the correction factor equals 0. The coefficient W itself varies from 0 to 1. The higher the value of W, the better the agreement in the group of experts.

As an example we demonstrate the application of the concordance test on the material of Text 1 (*AB*). First we have to calculate the correction factor for tied ranks. In the first column (see Table 2) we have four groups of tied ranks and we calculate: $T_1 = (2^3 - 2) + (8^3 - 8) + (2^3 - 2) + (3^3 - 3) = 540$. In a similar manner we calculate $T_2 = 720$, $T_3 = 180$, etc. The total $\Sigma_{j=1}^{m} T_j = 5256$.

Table 2
Rank values (R_{ij}) of marks assigned to 15 stylistic features of Novel 1 (AB) by 10 experts, computation of totals (ΣR_{ij}), deviations (d_i) and squares of deviations (d^2_i)

Char (i)	Experts (j)										ΣR_j	d_i	d^2_i
	1	2	3	4	5	6	7	8	9	10			
1	14	11	13	13.5	14	10	13	13.5	8.5	10	120.5	40.5	1640.25
2	1.5	4.5	2	9.5	9.5	10	7	13.5	8.5	10	76.5	-4	16
3	14	11	7.5	5	4.5	5	7	6.5	14.5	10	85.0	5	25
4	6.5	4.5	2	1.5	4.5	10	7	13.5	2	4.5	56.0	-24	576
5	6.5	11	9.5	9.5	1.5	5	7	13.5	8.5	4.5	76.5	-3.5	12.25
6	11.5	11	7.5	9.5	14	14	13	6.5	8.5	14.5	110.0	30	900
7	14	11	13	9.5	14	14	13	6.5	14.5	14.5	124.0	44	1936
8	6.5	4.5	5	5	9.5	5	7	6.5	8.5	4.5	62.0	-18	324
9	6.5	11	5	5	4.5	1.5	1.5	6.5	8.5	1.5	51.5	-28.5	812.25
10	6.5	4.5	13	13.5	9.5	14	13	6.5	8.5	10	99.0	19	361
11	6.5	11	13	13.5	9.5	10	7	6.5	8.5	10	95.5	15.5	240.25
12	1.5	1	9.5	5	4.5	5	3	6.5	2	10	48.0	-32	1024
13	6.5	2	2	1.5	1.5	1.5	1.5	1	2	1.5	21.0	-59	3481
14	6.5	11	13	13.5	9.5	5	7	6.5	8.5	4.5	85.0	5	25
15	11.5	11	5	5	9.5	10	13	6.5	8.5	10	90.0	10	100
	120	120	120	120	120	120	120	120	120	120	1200.0	0	11473

Using formula (1) we get the result

$$W = \frac{(12)11473}{10^2(15^3 - 15) - 10(5256)} = 0.486.$$

In order to determine whether the computed value of W is statistically significant, we can apply the chi-square test. On condition that $n > 7$, the magnitude $m(n-1)W$ follows the χ^2 distribution with $df = n-1$ degrees of freedom. In the case of tied ranks the formula is as follows

$$(2) \qquad \chi^2 = \frac{12\sum_{i=1}^{n} d_i^2}{mn(n+1) - \sum_{j=1}^{m} T_j/(n-1)}.$$

The calculation yields the result

$$\chi^2 = \frac{12(11473)}{10(15)16 - 5256/14} = 68.0.$$

For the number of degrees of freedom in question, which equals the number of observations minus one, that is 15 - 1 = 14, the chi-square required at the 0.001 level of significance is 36.12. Since $\chi^2 > \chi^2_{0.001;14}$, we can conclude that the calculated degree of concordance is statistically significant. In other words, the observed differences in the evaluation of 15 style characteristics by 10 experts may be regarded as mere variants of one and the same population.

The ascertainment of statistical homogeneity, as demonstrated above, gives us the confidence that the results of the experiment truly express the group opinion of the experts. Analogous results were obtained when analyzing the concordance of expert judgement on the remaining texts under examination. The number of experts (10) can also be considered sufficient for this kind of experiment.[3] All this allows us to state that the average estimates (medians) are sufficiently representative of the expert judgement under review and can be

[3] In an analogous experiment by J.B. Carroll (1960) the number of experts was 8. For more details about the methodology of expert judgement see, e.g. Amara & Lipinski (1972), Bešelev & Gurvič (1980).

used as typical (standard) values in further analysis.

Table 3
Average (median) scores given by experts and sums of average scores of stylistic features in seven texts

No. of style characteristic	Texts (No. and author)							Total score
	1 AB	2 VG	3 HK	4 JK	5 LP	6 VS	7 HS	
1. Emotionality	2	2	3	3	4.5	4	3	21.5
2. Intellectuality	3	3	3	4	4	2	2	21
3. Concreteness	4	4	4	4	3.5	3	4	26.5
4. Thoughtfulness	3.5	2	2	4	3	3	2	20
5. Eventfulness	3.5	3	4	4	3.5	3	4	25.
6. Expressiveness	3	2	3	4	4	3	3	22
7. Saturation with details	4	3	4	4	5	3	3	26
8. Dynamism	3	3	4	3	3	3	3	22
9. Verbosity	3	4	4	3	3	3	4	24
10. Clarity	3	4	3	3	4	4	4	25
11. Readability	3	4	4	2	4	4	4	25
12. Laconicism	3	3	2	2	3	2	2	20
13. Popularity of language	3	2	4	3	2	4	4	22
14. Fluency of language	3	3	4	3	3.5	4	3.5	24
15. Artistic perfection	3	3	3	4	4	3	2	22

3. Interaction between characteristics of style

The general results of expert evaluations of 15 style characteristics are presented in Table 3 which gives the average values (medians) of "intensity" of characteristics for all seven novels. For example, novels 5 and 6 by two female writers: Lilli Promet (Md = 4.5) and Veera Saar (Md = 4) scored the highest marks for "emotionality". "Emotionality" does not necessarily exclude "intellectuality" as proved by Lilli Promet (her style achieved the mark "4" also for "intellectuality"). If we sum up the results, it is revealed that according to the group opinion of our experts "concreteness" (sum of medians 26.5) gets the

highest score in the sample of present-day Estonian fiction under discussion. It is followed by "saturation with details" (sum of medians 26.0), "eventfulness" (25.0), "clarity" (25.0) and "readability" (25.0). The characteristics estimated lowest by our experts were "laconicism" and "thoughtfulness" (sum of medians 20.0 in both cases).

In order to find the possible interrelations between the evaluations of style characteristics and to check the correspondence of the model of interaction with our intuitive ideas, correlation analysis according to Pearson's product-moment method was carried out (see Table 4). The absolute values of the correlation coefficient (r) exceeding 0.65 are statistically significant at the 0.10 level and those exceeding 0.74 at the 0.05 level.

Table 4
Correlation matrix of subjective characteristics of style
(zeros and decimal points are omitted)

No. of characteristic	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Emotionality	11	-75	28	26	65	39	-03	-41	41	28	-24	09	80	37
2. Intellectuality		26	59	43	59	81	00	-38	-38	-52	38	-68	-20	89
3. Concreteness			-13	41	-18	04	24	51	-51	-35	11	-13	-64	-18
4. Thoughtfulness				33	75	55	09	-71	-65	-75	-26	04	00	75
5. Eventfulness					69	60	33	05	-37	-25	-28	14	18	18
6. Expressiveness						73	-09	-65	-26	-48	-19	-04	24	66
7. Saturation with details							17	-47	-35	-24	35	-42	22	73
8. Dynamism								47	-47	24	-35	42	44	-09
9. Verbosity									17	51	-17	20	00	-66
10. Clarity										68	17	-20	31	-26
11. Readability											11	10	65	-48
12. Laconicism												-84	-31	26
13. Popularity of language													37	-58
14. Fluency of language														00
15. Artistic perfection														

The strongest correlations could be found between the characteristics "intellectuality" and "artistic perfection" (r = 0.89), "emotionality" and "fluency of language" (r = 0.80). Strong correlation between the expert estimations of "intellectuality" and "artistic perfection" can be explained by subjective factors

(the experts' "taste") as well as by objective reasons (specific features of the texts observed).

Correlation analysis also gave a number of other fully acceptable positive or negative correlations between the characteristics as could be expected. For example, we could confirm the existence of a positive correlation between "readability", on the one hand, and "clarity" and "fluency of language" on the other hand. A tendency to negative correlation could be observed between "artistic perfection" and "verbosity" (r = -0.68). Strong negative correlation was observed between "popularity of language" and "laconicism" (r = -0.84), "readability" and "thoughtfulness" (r = -0.75), "emotionality" and "concreteness" (r = -0.75), "popularity of language" and "intellectuality" (r = -0.68).

On the basis of significant positive correlations between the characteristics we constructed a visual model of interaction between the style characteristics in the sample texts observed (see Fig. 1). It can be seen from the scheme that several closely connected groups of characteristics are formed, e.g. the characteristics "intellectuality", "thoughtfulness", "expressiveness" and a few others are grouped around "artistic perfection". The characteristic "emotionality" is directly connected with "expressiveness", "fluency of language", "readability" and "clarity".

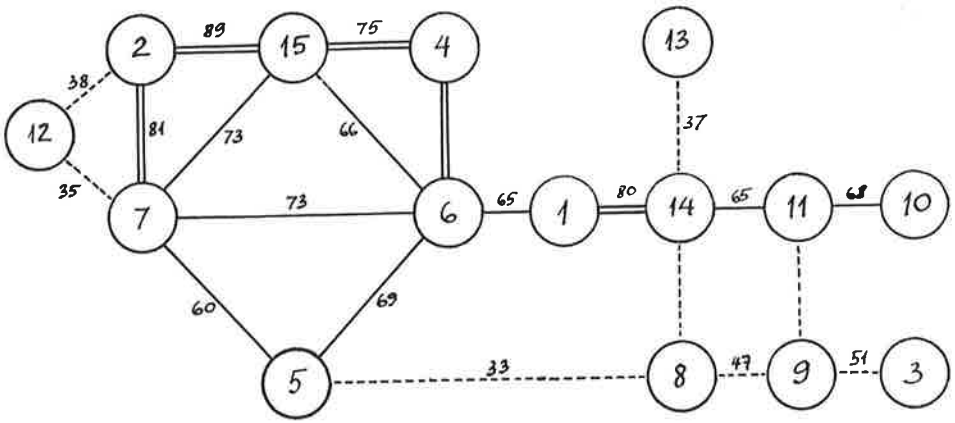

Figure 1. Interaction of stylistic features in present-day Estonian fiction: the numbers in the circles indicate the style characteristics (see Table 1) and the numbers over lines the strength of correlation

The characteristics which, according to the results of our experiment, have no significant positive correlation with other characteristics can be connected

with the general model via these characteristics with which they have the strongest positive connection (marked by a dotted line in Fig. 1).

Thus, by means of correlation analysis we have established positive and negative connections between the characteristics, which in general correspond to the (intuitively) expected relations between stylistic features. This testifies to the fact that experts' estimations can serve as a basis for constructing a model of interaction between the style characteristics of a work of fiction. In addition, the results of our investigation can be used to establish the interrelations between the subjective and objective characteristics of style.

4. Comparison of subjective and objective characteristics of style

As we also have at our disposal the data on some quantitative linguistic (stylostatistic) parameters important for stylistic analysis of texts studied (see the article "An attempt at quantitative analysis of the style of fiction" in this issue), we can compare the objective linguistic parameters with the experts' estimation of qualitative features of style. As objective parameters we shall observe such stylostatistic characteristics as frequencies of different parts of speech, the indices of the concentration of the 50 most frequent content words and the concentration of frequent word forms (mostly function words), proportion of rare (nonce) words, or "hapax legomena" (the so-called index of exclusiveness) and the degree of entropy which was calculated on the basis of distribution of parts of speech in the text.

In addition to these parameters we also used such characteristics as average length of words, average length of sentences and the index of objective difficulty (complexity) of the text which is calculated according to the formula (Tuldava 1975):

$$R = \bar{i} \ln \bar{j} \tag{3}$$

where R is the index of text complexity, $\bar{i}$ the average word length in syllables, $\bar{j}$ the average sentence length in word forms and *ln* the natural logarithm.

In total, we have selected 15 objective characteristics of the text which are compared with the subjective estimations of 15 stylistic features (see Table 5).

At the first stage we performed correlation analysis of all the characteristics studied in the seven texts.

Table 5
Factor matrix of subjective and objective characteristics of style
(varimax rotation)

Characteristics	Significant factor loadings (F ≥ |0.4|)			
	F_1	F_2	F_3	F_4
Subjective characteristics				
1. Emotionality	0.63			
2. Intellectuality			0.85	-0.50
3. Concreteness	-0.89			
4. Thoughtfulness			0.80	
5. Eventfulness	-0.47		0.73	
6. Expressiveness			0.88	
7. Saturation with details		0.46		
8. Dynamism		0.45		
9. Verbosity	-0.57		-0.48	
10. Clarity			-0.54	
11. Readability		0.63	-0.61	
12. Laconicism				-0.92
13. Popularity of language				0.93
14. Fluency of language	0.48	0.65		0.48
15. Artistic perfection	0.49		0.78	
Objective characteristics				
1. Frequency of nouns	-0.94			
2. " adjectives		-0.77		-0.47
3. " pronouns	0.90			
4. " verbs		0.87		
5. " adverbs		-0.81	-0.50	
6. " pre- and postpositions		-0.45		0.40
7. " conjunctions	0.58	-0.42	0.51	0.45
8. " content words	-0.91			
9. Concentration of 50 most fre quent content words	-0.52			
10. Entropy (distribution of parts of speech)	0.70	-0.64		
11. Concentration of frequent word forms	0.92			
12. Hapax legomena	-0.80		0.48	
13. Word length	-0.54			-0.77
14. Sentence length		-0.91		
15. Index of text complexity		-0.98		

The strongest cross-correlations we could observe were the positive correlations between "concreteness" and "frequency of nouns" (r = 0.83), "emotionality" - "frequency of pronouns" (r = 0.67), "dynamism" - "frequency of verbs" (r = 0.66), "eventfulness" - "proportion of rare words" (r = 0.66), etc.

To present the general impression in a more compact way and to find some internal regularities of interrelations between the subjective and objective style characteristics, we made use of factor analysis (for the principles of factor analysis, cf. the article "An attempt at quantitative analysis..." in this issue). We determined the four main factors where the contribution of each of them is more then 10 per cent of the total variation. All four factors together account for about 80 per cent of the total variation of characteristics. The results of the factor analysis are given in Table 5.

Along with the calculation of factor loadings (F_i) for style characteristics, we also determined the role of factors in the classification of texts by calculating the individual values of factors (f_i) for each text separately (see Table 6) which have also been presented graphically as "style profiles" (Carroll 1960) of the texts studied (see Fig. 2).

The analysis of the factor matrix allows us to draw the following conclusions. The factors have a bipolar character. Consequently, if we wish to, we can speak about eight groups of relevant characteristics, which essentially describe the style of the novels observed.

Table 6
Individual factor values of seven texts

Text (author)		Factor values			
		f_1	f_2	f_3	f_4
1. Aimée Beekman	(AB)	-0.6	0.4	-0.1	-0.7
2. Villem Gross	(VG)	-0.1	-0.7	-1.1	-1.6
3. Heino Kiik	(HK)	-0.3	0.9	0.3	0.8
4. Jaan Kross	(JK)	0.0	-1.7	1.5	0.3
5. Lilli Promet	(LP)	0.7	1.2	1.1	-0.8
6. Veera Saar	(VS)	1.8	-0.2	-1.0	0.9
7. Herman Sergo	(HS)	-1.4	0.0	-0.7	1.0

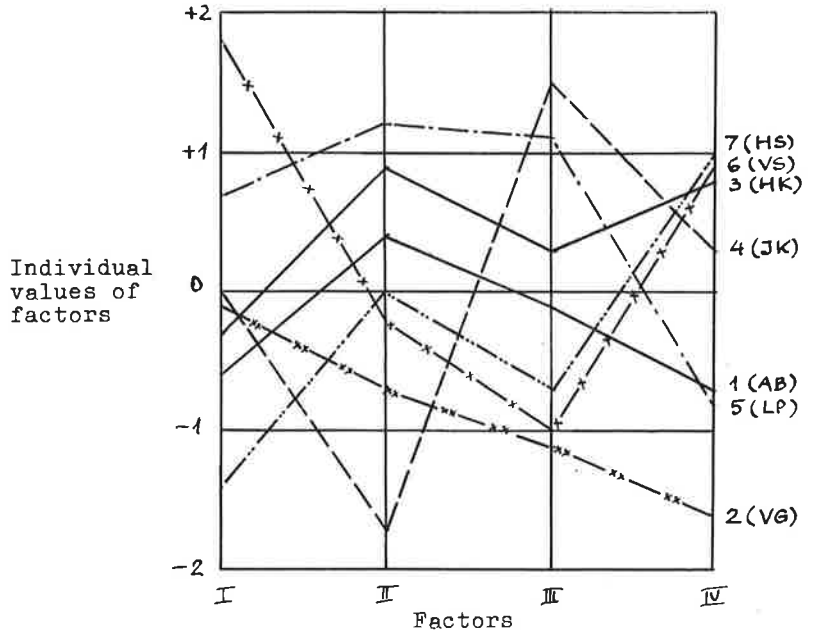

Figure 2. Style profiles of seven texts

The first and the strongest factor (explaining 25 per cent of the total variation) connects the characteristics "emotionality", also "fluency of language" and "artistic perfection", with the frequency of pronouns and conjunctions, entropy and concentration of frequent word forms in the text. This complex factor, which provisionally could be called the factor of *emotionality*, is markedly expressed in text 6 (the individual value of f_i is as high as 1.8; see Table 6). The same factor is also quite strongly expressed in Text 1 (f_i = 0.7). The respective authors are two female writers Veera Saar (VS) and Aimée Beekman (AB). The bipolar, opposite variant of the first factor (with negative factor loadings) connects the subjective characteristic of "concreteness" (to a smaller degree also "eventfulness" and "verbosity") and the objective linguistic parameters - frequency of nouns, concentration of the 50 most frequent content words, the proportion of rare word forms and average length of the word. This factor is most strongly expressed in text 7 (f_1 = -1.4). As for the other authors, the first bipolar factor seems to be neutral ($f_1 < |0.7|$), i.e. it has no differentiating

role in the comparison of styles[4].

The second factor could be called the factor of *readability*; it connects on the one hand the subjective characteristics of "readability", "fluency of language" and, to some extent, "dynamism", and the objective characteristic of frequency of verbs on the other hand. This factor of readability (i.e. easy reading) is naturally opposed by negative factor loadings expressing the objective characteristics of "text complexity", "sentence length", "frequency of adverbs" and "frequency of adjectives". This negative variant of the second factor has a high significance in Text 4 (f_2 = -1.7), to a smaller degree in text 2. It is noteworthy that the index of text complexity, calculated according to formula (3), correlates well with the subjective estimations of readability of the text (the correlation is negative). We can also note the correctness of the intuitive assumption that "dynamism" of style is connected with frequent use of verbs in the narrative.

The third factor is *intellectuality*. It unites the characteristics "intellectuality", "thoughtfulness", "eventfulness", as well as the evaluative characteristics "artistic perfection" and "expressiveness". The objective characteristics connected with it are frequency of conjunctions and proportion of rare words (index of exclusiveness). This factor is characteristic of Texts 4 and 5. The bipolar variant with negative factor loadings can be characterized by "verbosity", "readability" and "clarity" and the objective parameter of frequency of adverbs, and it is typical of Texts 2, 6 and 7.

The fourth factor may be called, after its dominant characteristic, *popularity of language.* Another style characteristic connected with it is "fluency of language". Here we find high frequency of pre- and postpositions and conjunctions. At the same time we can notice the relatively low frequency of adjectives and a negative correlation with the length of the word, i.e. the usage of short words is peculiar to that style. According to the individual values the fourth factor plays a significant role in Texts 3, 6 and 7. The negative loadings of the fourth factor indicate the characteristics "laconicism" and "intellectuality", which are connected with the stylostatistic parameters of frequency of adjectives and length of the word. This kind of combination of characteristics is typical of Text 4 (a historical novel by Jaan Kross).

The complex factors express certain internal regularities which in their various combinations characterize individual styles to a significant degree. At that, we should keep in mind that in their essence the factors are independent of each other ("orthogonal" factors - in the terminology of factor analysis). Therefore,

[4] In this study with 30 characteristics and critical values of F_i and f_i approximately estimated as |0.4| and |0.7|, correspondingly.

one and the same text (author) can be simultaneously characterized by several different factors. The author's individuality is revealed in the combination of individual factor values - in the so-called *style profile* (see Fig. 2). For example in Text 5 (Lilli Promet) both the stylistic factors of "emotionality" and "readability" along with the objective features like high frequency of pronouns and verbs as well as concentration of frequent word forms and use of short words and short sentences play a significant role. At the same time the third factor of "intellectuality" is also notable in this text, combining such objective characteristics as abundance of rare words (a characteristic of "richness of vocabulary") and a great proportion of pre- and postpositions (which might show the "analyticism" of thought and language).

5. Conclusion

The statement of connection between subjective and objective characteristics of style may lead us to think about regular mutual interdependence between subjective and objective characteristics. This does not mean, however, that we could draw direct conclusions about the qualitative features of style depending on the existence and strength of certain stylostatistic parameters of texts. Nevertheless, the results of the research into subjective and objective characteristics and their interrelations can be of help to the comprehensive qualitative analysis of style of a work of fiction and solving the problems connected with the study of typology of styles. The problem, on the whole, deserves further investigation on enlarged material (new texts and additional sets of characteristics) and other languages, as well.

References

Amara, R. & Lipinski,A. (1972). Some views on use of expert judgement. In: Technological Forecasting and Social Change, No. 3.

Bešelev, S.D. & Gurvič, F.G. (1980). *Matematiko-statističeskie metody ekspertnych ocenok* (Mathematical-statistical methods of expert evaluations). Moscow, Statistika.

Carroll, J.B. (1960). Vectors of prose style. In: Sebeok, T.A. (ed.), *Style in Language*. Cambridge, Mass.: 283-292.

Doležel, L. (1969). A framework for the statistical analysis of style. In: Doležel, L. & Bailey, R.E. (eds.), *Statistics and Style.* New York, Elsevier, 10-25.

Enkvist, N.E. (1974). *Stilforskning och stilteori* (The study of style and theory of style). Lund, Uniskol.

Frumkina, M.R. & Vasilevič, A.P. (1971). Polučenie ocenok verojatnosti slov psichometričeskimi metodami (The assessment of word probability estimates by psychometric methods). In Frumkina, M.R. (ed.), *Verojatnostnoe prognozirovanie v reči.* Moscow, Nauka, 7-28.

Grotjahn, R. (1979). *Linguistische und statistische Methoden in Metrik und Textwissenschaft.* Bochum, Brockmeyer.

Kendall, M. (1955). *Rank correlation methods.* New York, Hafner.

Kožina, M.N. (1972). *O rečevoj sistemnosti naučnogo stilja sravnitel'no s nekotorymi drugimi* (On the systematicity of the scientifc style in comparisonwith some others). Perm', Izd. Universiteta.

Tuldava, J. (1975). Ob izmerenii trudnosti teksta (On measuring text difficulty). *Methodica IV. Acta et Commentationes Universitatis Tartuensis 345, 102-120* (English translation in press).

6

On the Lexical Connection of Texts

The present paper deals with some possible methods for the estimation and measurement of lexical connection (closeness, similarity) of texts on *vocabulary level*, i.e. not considering the frequencies of words in the texts. The problem of the lexical connection on the *text level*, i.e. using the frequencies of occurrence of words in the texts, will be the subject matter of a separate discussion (see the next article in this volume).

1. The necessity for the estimation and measurement of the lexical connection of texts may appear in the statistical investigation of individual styles or literary genres, when comparing frequency lists of words, solving problems of automatic classification of texts, etc. In this paper we are interested in the comparison of two texts, proceeding from the lists of words, not taking into account their frequencies in the texts. Depending on the tasks set by the investigator all the words (word-forms or lexemes) or certain classes of words of concrete texts may be compared.

To illustrate the method the present paper compares monosyllabic word-forms of the Russian language appearing in one and the same text but in various contexts: in the author's monologue ("text *A*") and in the direct speech of the characters ("text *B*") of the novel "Na tichom brege" (*On the quiet bank*) by Boris Polevoj (Moscow 1966). The size of the sample (length of text) in both cases is 2,000 word occurrences (one-syllable word-forms).

2. In text *A* there were 211 different word-forms, in text *B* 186. Thus, the initial data are the following:

$$N_A = 2000, \qquad V_A = 211;$$
$$N_B = 2000, \qquad V_B = 186.$$

The symbol N with the subscript denotes the size of the corresponding text and V the size of the vocabulary.

Comparing the lists obtained of one-syllable word-forms we calculate the number of word-forms which are common to both texts (C) and the number of word-forms which appear only in the vocabulary of text A (A_0) and only in the vocabulary of text B (B_0). The analysis of our material gives the following results:

$C =$ 96
$A_0 =$ 115
$B_0 =$ 90
Total 301

The total (301) is equal to the number of monosyllabic word-forms in the joint (combined) vocabulary, the size of which we denote by $V_{A,B}$ when

$$V_{A,B} = C + A_0 + B_0 = V_A + V_B - C.$$

These interrelationships can be visualized graphically (Fig. 1) where C is the overlap in word-forms in V_A and V_B, i.e. the common part of the two vocabularies.

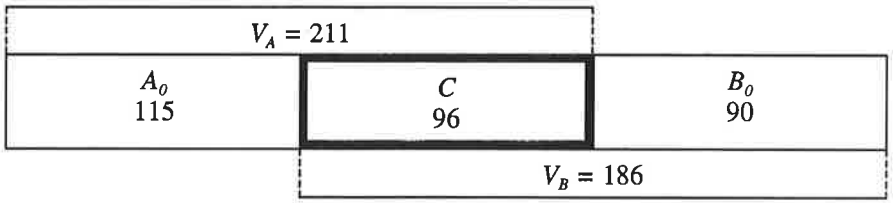

Figure 1. Graphical representation of joint vocabulary

3. Proceeding from the above, it is possible to calculate the *indices of connection* (or correlation) between the comparable texts. We shall use three of the best known indices. The first of these expresses the ratio of the common part (C) to the size of the joint vocabulary:

(1) $$R = \frac{C}{V_{A,B}} = \frac{C}{V_A + V_B - C}$$

where R is the index of lexical connection. The index may obtain values from 0 to 1. It is evident that $R = 0$ if $C = 0$, i.e. when there are no common words in the vocabularies compared; when both vocabularies exactly coincide ($V_A = V_B = C$) then $R = 1$. In our example the calculation of the index of lexical connection yields

$$R = \frac{96}{211 + 186 - 96} = 0.319.$$

Another possibility for constructing an index of lexical connection is to calculate the ratio of the common part (C) to the average size of the two vocabularies compared:

(2) $$R = \frac{C}{(V_A + V_B)/2} = \frac{2C}{V_A + V_B},$$

which, when used with the data of our experiment, yields

$$R = \frac{2 \cdot 96}{211 + 186} = 0.484.$$

Statistically interpreted, it means that on average the common part constitutes 48.4% of the size of each vocabulary taken separately.

Instead of the arithmetical mean it is also possible to use the geometrical mean as a basis, especially in the case when the vocabulary sizes differ greatly. The formula of the index of lexical connection will then take the following form:

(3) $$R = \frac{C}{\sqrt{V_A \cdot V_B}}.$$

In our example the calculation of the index using formula (3) gives $R =$ 0.485. The difference is insignificant compared with the calculation using formula (2) because the sizes of vocabularies ($V_A = 211$ and $V_B = 186$) do not differ much from each other.

The indices calculated by using formulas (2) and (3) may obtain values from 0 to 1, as for the index calculated using formula (1). The calculated values of the indices are not directly comparable; nevertheless, they are all in positive correlation with each other, the difference being only in the rate of increase or decrease of numerical values.

The drawback of all these indices is the fact that their numerical values depend on the size of sample. Therefore, in a pairwise comparison of several dif-

ferent texts, even using one and the same formula, it is necessary to compare the data of samples of approximately equal sizes, or to use some special testing procedures for comparison (see below).

Another drawback of the indices mentioned above is the fact that we cannot simply judge by the index values whether they are statistically significant or not. For the purpose of calculating the degree of significance a special method will be proposed.

4. As mentioned above we have to take into consideration the sample size when calculating the indices of lexical connection of texts. The question of *confidence limits* of the indices arises.

A somewhat dubious procedure would be to consider the common part of the comparable vocabularies (C) as a mere proportion of the joint vocabulary (as was proposed in the original version of the article) and correspondingly to determine the standard deviation of the index using the well-known formula

$$\sigma_p = \sqrt{\frac{p(1 - p)}{n}}. \tag{4}$$

If we do this, we obtain the following result for index R when using formula (1):

$$\sigma_R = \sqrt{\frac{R(1 - R)}{V_{A,B}}} = \sqrt{\frac{0.319 \cdot 0.681}{301}} = 0.03.$$

A more satisfactory solution might be to interpret the indices as measures of correlation (contingency). The standard deviation is then calculated using the following formula (Herdan 1966:156):

$$\sigma_R = \frac{1 - R^2}{\sqrt{n}}. \tag{5}$$

For index R when using formula (1) the standard deviation would have the value

$$\sigma_R = \frac{1 - 0.319^2}{\sqrt{310}} = 0.05.$$

At the 0.05 level of significance the confidence limits of index R are determin-

ed by $R \pm 1.96\sigma_R$.

This calculation is correct under the assumption of a normal distribution of the data (for n sufficiently large).

5. In order to establish a *criterion of significance* when measuring the lexical connection between two texts (on the vocabulary level), the following method may be recommended.

Under the null hypothesis we tentatively assume that the observed differences between the distributions of word frequencies in the comparable texts could have arisen purely by chance and both distributions came from one statistical population. We should keep in mind that we are comparing the lists of monosyllabic word-forms in the author's monologue (text A) and the direct speech of the characters (text B) of the same novel.

To begin with, we take our two texts, A and B, and consider them to be combined into a whole - a *joint text*. The combination of the corresponding vocabularies gives us the *joint vocabulary*, taking into account the eventual overlap in the words of the two vocabularies (see Figure 1).

We proceed with constructing a theoretical model for the joint text where the word frequencies would be distributed according to the binomial formula (cf. Muller 1968):

$$(p + q)^m. \tag{6}$$

Here m denotes the word frequency, p and q are the probabilities with which some randomly taken word-form from the joint text would belong either to (sub)text A or B:

$$p = \frac{N_A}{N_A + N_B}; \qquad q = \frac{N_B}{N_A + N_B}; \qquad p + q = 1.$$

The next step is to calculate the theoretically expected number of words with a given frequency, occurring only in one of the texts (as part of the joint text), using the following fomulas:

$$A_0' = \sum f_m p^m \qquad (\textit{for text } A), \tag{7a}$$

$$B_0' = \sum f_m q^m \qquad (\textit{for text } B), \tag{7b}$$

where A'_0 and B'_0 are the theoretically expected numbers of different word-forms, peculiar to text A or to text B; f_m is the number of word-forms with the given frequency (m).

For the calculation of A'_0 and B'_0 it is necessary to make a table for the distribution of the observed word frequencies in the joint text starting with the smaller frequencies. On the basis of such a table the values of $f_m p^m$ and $f_m q^m$, and the corresponding sums are calculated. In our example $p = q = 0.5$ (as $N_A = N_B = 2000$); consequently, A'_0 and B'_0 are of equal size. The calculation of A'_0 is given in Table 1.

As the Table shows, it is sufficient to calculate the values of $f_m p^m$ for the frequencies up to $m = 7$ because the addition of new frequencies (in our case) does not change the result. The theoretically expected number of different monosyllabic word-forms occurring only in text A is 84 units. The same number of non-overlapping word-forms could be expected in text B. As these theoretical values are obtained on the basis of the empirical distribution of word frequencies in the joint text (the values of f_m in Table 1), we may in the calculation of the theoretically expected number of overlapping word-forms proceed from the empirical size of the joint vocabulary ($V_{A,B} = 301$):

$$C' = V_{A,B} - (A'_0 + B'_0) = 301 - (84 + 84) = 133.$$

Table 1
Calculation of the expected number of words occurring only in text A

m	f_m	$f_m p^m$		
1	139	$139(0.5)$	=	69.5
2	42	$42(0.5)^2$	=	10.5
3	24	$24(0.5)^3$	=	3.0
4	5	$5(0.5)^4$	=	0.4
5	9	$9(0.5)^5$	=	0.2
6	7	$7(0.5)^6$	=	0.1
7	5	$5(0.5)^7$	=	0.0
...	...	...		...
-	-	$\Sigma f_m p^m = 83.8 \approx 84$		

Thus, the theoretical distribution of overlapping and non-overlapping monosyllabic word-forms in the joint vocabulary would be:

$$C' = 133; \quad A'_0 = 84; \quad B'_0 = 84.$$

These theoretical data can be compared with the observed data and the hypothesis of the identity of distributions can be tested by the chi-square test (Table 2).

Using the formula

$$X^2 = \sum \frac{(O - E)^2}{E} \tag{8}$$

the value of X^2 works out to be 22.1 with $k - 1 = 3 - 1 = 2$ degrees of freedom. Reference to the χ^2-table (e.g. Owen 1962) shows this to have a probability of less than 0.001 ($\chi^2_{0.001;2} = 13.82$), i.e. the observed differences could arise by chance less than once in 1,000 times. We therefore reject our initial null hypothesis and accept instead that the observed differences are statistically highly significant. Consequently, the observed word frequency distribution of the joint text (combination of two separate texts) does not correspond to the distribution that could be expected if the (sub)texts came from one homogeneous text.

Table 2
Calculation of the chi-square value

Parts of the vocabulary	Observed (O)	Expected (E)	O - E	$\frac{(O - E)^2}{E}$
C	96	133	-37	1369/133 = 10.3
A_0	115	84	+31	961/84 = 11.4
B_0	90	84	+ 6	36/84 = 0.4
$k = 3$	301	301	0	$X^2 = 22.1$

Conclusion: from the statistical point of view, the author's monologue and the speech of characters in the novel examined cannot be regarded as belonging to the same population with regard to the occurrence of monosyllabic word-forms.

References

Herdan, G. (1966). *The advanced theory of language as choice and chance.* Berlin-Heidelberg-New York, Springer.

Muller, Ch. (1968). *Initiation à la statistique linguistique.* Paris, Larousse.

Owen, D.B. (1962). *Handbook of statistical tables.* Reading, Mass., Addison Wesley.

7

A Statistical Method of Comparison of the Lexical Composition of Two Texts

In this paper we shall treat the problem of the lexical connection of texts in greater detail than we did in the paper *On lexical connection of texts*. This will allow us to introduce one or two new topics and methods by analyzing the lexical composition of texts and measuring their connection (correlation) on the text level, i.e. taking into account the fact that the words have different frequencies of occurrence in the comparable texts.

1. At present there are two main approaches to the comparison of the vocabularies of two texts. In the first case the closeness or distance of the texts with regard to their lexical composition is measured considering the degree of coincidence (overlap) of the vocabularies or the correlation between the word frequency distributions. Special indices are most often used for this purpose.

The second approach is characterized by an entirely different view of the notions of connection or correlation between vocabularies. Here the investigator does not proceed from the requirement of coincidence or the degree of identity of vocabularies but from the assumption of the probability distribution of word frequencies in a "joint text" according to the method proposed by C. Muller (1968). From this we can conclude that for the comparison of vocabularies of two texts an experiment of combining the vocabularies into a whole should be carried out, taking into account the repeated use of words in both texts. The observed word frequency distribution will be compared with a theoretical model and the hypothesis of homogeneity will be tested by a statistical criterion. Beside this, it is possible to carry out qualitative stylistic analysis of the lexical composition of the comparable texts.

2. It will be assumed that in the case of homogeneity of the vocabularies of two texts A and B the distribution of subfrequencies of words in the joint text is characterized by the binomial coefficient $(p + q)^m$, where p and q are the probabilities of words belonging either to text A or to text B, if the words are taken at random from the joint text; m denotes the frequency of occurrence of a word in the joint text. The probabilities p and q are calculated on the basis of the sizes of the comparable texts ($p + q = 1$).

In the most convenient case, when the comparable texts are of equal size ($N_A = N_B$) and, consequently, $p = q = 0.5$, the distribution of word frequencies in the joint text would be the following:

- the words with frequency 1 ($m = 1$) may be distributed with equal probability (0.5 and 0.5) either in (sub)text *A* or (sub)text *B*;

- the words with frequency 2 in the joint text are distributed according to the formula $(p + q)^2$, i.e. in the proportions 0.25 + 0.5 + 0.25: the probability of not occurring in text *A* (or text *B*) is 0.25, the probability of occurring equally - once in text *A* and once in text *B* - is 0.5, and the probability of occurring twice in text A (or text B) is 0.25;

- the words with frequency 3 are distributed according to the formula $(p+q)^3$, i.e. the combination of subfrequencies for text *A* or text *B* would be 0 - 3 (zero-occurrence in one of the texts and frequency 3 in the other text) with probability 0.125, the combination 1-2 with probability 0.375, the combination 2 - 1 with probability 0.375, and the combination 3 - 0 with probability 0.125; etc. (see Barlow's Tables: Table V).

3. With the purpose of illustrating the method, we have taken two random samples of equal size ($N_A = N_B$ = 5,000 word occurrences) from the author's monologue (i.e. non-conversational material) of two modern Estonian novels:

text *A*: the novel *Kartulikuljused* ("Toneless bells") by Aimée Beekman (1968),
text *B*: the novel *Primavera* by Lilli Promet (1971).

The analysis showed that in text *A* there were 1768 and in text *B* 1564 different words (lexemes). When comparing the distributions of word frequencies (the "lexical spectra") of the two texts (see Table 1) we can see essential differences in the distribution of words with frequencies 1 and 2 ($m = 1$, and $m = 2$) and in the totals (V_A and V_B). These data themselves disclose certain individual differences of style ("richness of vocabulary"), but a more subtle analysis requires the examination of word frequency distributions in the individual texts as well as in the joint text. The size of the joint text is, naturally, 5,000 + 5,000 = 10,000 word occurrences, and the joint vocabulary constitutes 2809 different words (taking into account the overlap of the two vocabularies).

4. The next stage in the experiment will be the construction of a theoretical model of the distribution of subfrequencies (for text A and text B) in the joint text. As mentioned above we proceed from the assumption of the distribution of word frequencies according to the binomial formula $(p + q)^m$. The theoretical model will then be compared with the empirical model where the subfrequencies are calculated on the basis of the data obtained (Table 2). In our example the subfrequencies have been calculated in the limits from $m = 1$ to $m = 10$, which involve 2684 words of the total 2809, i.e. 96.6% of the joint vocabulary. In this way the most frequent words, which in their majority are grammatical words (conjunctions, pre- and postpositions, auxiliary verbs) and pronouns, are not considered. In the present case we did not include them, but in principle, they may be involved, for example, by combining them in a group. It is also possible to deal with different classes of words separately.

Table 1
Distribution of word frequencies in text A and text B and in the joint text

m	Text A N_A = 5000 f_m	Text B N_B = 5000 f_m	The joint text $N_{A,B}$ = 10000 f_m
1	948	860	1808
2	300	226	406
3	120	109	163
4	72	64	89
5	57	55	66
6	38	31	39
7	40	35	42
8	27	28	30
9	21	16	21
10	20	18	20
Sum	1643	1442	2684
> 10	125	122	125
Total	1768 (V_A)	1564 (V_B)	2809 ($V_{A,B}$)

N_A and N_B - text sizes; V_A and V_B - vocabulary sizes
f_m - the number of words with frequency m
m - word frequency

The comparison of the theoretical and the empirical models of the distribution of subfrequencies in the joint text will reveal some interesting differences, especially in the distribution of words with small frequency. From Table 2 it can be seen that the words with frequency 1 are distributed in the following way: the observed number in text *A* is 948, and in text *B* 860. As the expected number of words with frequency 1 is 904 for both texts, we can con-clude that

in text *A* the observed number (948) exceeds the expected number by 44 units, whereas 44 units are missing in text *B* (860 to 904). The words with frequency 2 in the joint text (406 words) should be distributed as 101 - 204 - 101 (101 words twice in Text *A*, 204 words once in both texts, 101 twice in text *B*) but in reality their distribution is 178 - 120 - 108. In text *A* there are 77 words with frequency 2 that exceed the expected number (101) in the joint text and can therefore (approximately) be considered as peculiar (specific) words of text *A*. In text *B* 7 words may be peculiar (108-101). The expected equilibrium (one occurrence in both texts) fails with 84 units to reach the expected number 204.

5. In order to calculate the degree of statistical homogeneity and work out an adequate measure of lexical connection (correlation) between the two texts, we shall apply the chi-square test. For Table 2 we calculate the chi-square as X^2 = 380.87 (see Table 2). For the number of degrees of freedom $df = k\text{-}1 = 38 - 1 = 37$ (after having gathered small subfrequencies into groups) the chi-square required at the 0.001 level of significance is 73. This shows that the differences in the distribution cannot be attributed to pure chance and, consequently, there is a significant difference between the lexical compositions of texts *A* and *B*.

For the comparison of the results of different experiments the values of X^2 cannot serve as measures of lexical connection because they change depending on the number of observations. The intensity of relationship can be ascertained by transforming the chi-square into a modification of Pearson's coefficient of contingency:

$$(1) \qquad K = 1 - \sqrt{\frac{X^2}{n + X^2}}$$

where K stands as the index of lexical connection, n denotes the number of observations (here $n = 2684$, as we have included words with frequencies 1 to 10 only; see Table 1). The index K will obtain the value 1 in the case if $X^2 = 0$ (i.e. when the theoretical and empirical models fully coincide), and is nearing the value 0 with the increasing of the value of X^2. In our case

$$K = 1 - \sqrt{\frac{380.87}{2684 + 380.87}} = 0.647.$$

Table 2
Distribution of subfrequencies in the joint text and calculation of the chi-square value

Frequency	Number f_m	Distrib. A - B	Probability	O	E	O - E	$(O-E)^2/E$
m = 1	1808	1 - 0	1/2	948	904	+44	2.14
		0 - 1	1/2	860	904	-44	2.14
m = 2	406	2 - 0	1/4	178	101	+77	58.70*
		1 - 1	2/4	120	204	-84	34.59*
		0 - 2	1/4	108	101	+7	0.49
m = 3	163	3 - 0	1/8	54	20	+34	57.80*
		2 - 1	3/8	35	61	-26	11.08*
		1 - 2	3/8	31	61	-30	14.75*
		0 - 3	1/8	43	20	+23	26.45*
m = 4	89	4 - 0	1/16	25	6	+19	60.17*
		3 - 1	4/16	14	22	-8	2.91
		2 - 2	6/16	14	33	-19	10.94*
		1 - 3	4/16	19	22	-3	0.41
		0 - 4	1/16	17	6	+11	20.17*
m = 5	66	5 - 0	1/32	11⌉	2⌉		
		4 - 1	5/32	10⌋	10⌋	+9	6.75
		3 - 2	10/32	15	21	-6	1.71
		2 - 3	10/32	14	21	-7	2.33
		1 - 4	5/32	7⌉	10⌉		
		0 - 5	1/32	9⌋	2⌋	+4	1.33
m = 6	39	6 - 0	1/64	8⌉	1⌉		
		5 - 1	6/64	6⌋	4⌋	+9	16.20*
		4 - 2	15/64	5	9	-4	1.78
		3 - 3	20/64	7	12	-5	2.08
		2 - 4	15/64	3	9	-6	4.00
		1 - 5	6/64	9⌉	4⌉		
		0 - 6	1/64	1⌋	1⌋	+5	5.00

Table 2. (Continued)

m = 7	42	7 - 0	1/128	7	0		
		6 - 1	7/128	2	2	+5	2.78
		5 - 2	21/128	5	7		
		4 - 3	35/128	8	12	-4	1.33
		3 - 4	35/128	6	12	-6	3.00
		2 - 5	21/128	7	7		
		1 - 6	7/128	5	2	+5	2.78
		0 - 7	1/128	2	0		
m = 8	30	8 - 0	1/256	2	0		
		7 - 1	8/256	4	1		
		6 - 2	28/256	3	3	0	0.00
		5 - 3	56/256	2	7		
		4 - 4	70/256	7	8	-1	0.13
		3 - 5	56/256	1	7	-6	5.14
		2 - 6	28/256	6	3		
		1 - 7	8/256	1	1	+7	12.25*
		0 - 8	1/256	4	0		
m = 9	21	9 - 0	1/512	5	0		
		8 - 1	9/512	0	0		
		7 - 2	36/512	2	2		
		6 - 3	84/512	3	3	+5	5.00
		5 - 4	126/512	3	5	-2	0.80
		4 - 5	126/512	3	5	-2	0.80
		3 - 6	84/512	3	3		
		2 - 7	36/512	1	2		
		1 - 8	9/512	1	0	0	0.00
		0 - 9	1/512	0	0		
m = 10	20	10 - 0	1/1024	2	0		
		9 - 1	10/1024	0	0		
		8 - 2	45/1024	1	1		
		7 - 3	120/1024	2	2		
		6 - 4	210/1024	4	4	+2	0.57
		5 - 5	252/1024	2	5	-3	1.80
		4 - 6	210/1024	2	4		
		3 - 7	120/1024	4	2		
		2 - 8	45/1024	3	1		
		1 - 9	10/1024	0	0		
		10 -0	1/1024	0	0	+2	0.57
r = 10	-	-	n = 2684	k = 38			X^2 = 380.87*

Index *K* may be used in the comparison of more than two texts on condition that the models of the distribution of subfrequencies have a similar number of groups (*r*) but may have a different number of observations (*n*). When comparing models with a different number of groups it is necessary to take into consideration the number of groups in the comparable pairs of models. With some approximation the index K_r (a modification of Čuprov's coefficient for contingency tables) may be used as a measure of lexical connection in the following form:

(2) $$K_r = 1 - \sqrt{\frac{X^2}{n\sqrt{r-1}}}$$

where *n* denotes the number of observations and *r* the number of groups. In our example $n = 2684$ and $r = 10$ (see Table 2). The calculation of index (2) gives the result:

$$K_r = 1 - \sqrt{\frac{380.87}{2684\sqrt{10-1}}} = 0.783.$$

This index may have the maximum value 1, which is reached in the case of $X^2 = 0$, i.e. when the theoretical and empirical models fully coincide, and is nearing the value 0 with the increasing of X^2 value.

Both coefficients - *K* and K_r - have been calculated on the basis of the chi-square, but the concrete numerical values of *K* and K_r are not directly comparable. One has to use only one of them when comparing the data of several experiments. However, the chi-square test can always be used for assessing the statistical significance of observed differences between the comparable distributions.

In practice it is not always possible, especially in the case of a small sample size, to compare the frequency distributions of vocabularies fully. Then one can use the procedure of comparing certain groups of vocabularies, e.g. common (overlapping) and individual (non-overlapping) parts of the vocabularies, as was demonstrated in the paper on the measurement of lexical connection of texts on the vocabulary level.

6. The proposed method may be used for *qualitative lexico-stylistic analysis* of concrete word usage in the model of the joint text. As an example we shall compile an incomplete list of words with frequency 5 (Table 3). It can be seen

that in the case of a 5-0 distribution, i.e. words with 5 occurrences in text *A* and zero-occurrence in text *B*, there are 11 such words in text *A*, but the theoretically expected ("admitted") number is only 2. Consequently, 11 - 2 = 9 words with frequency 5 in text *A* may tentatively be qualified as words peculiar to text *A*, and only 2 words might belong to the stratum of overlapping words according to the binomial division of subfrequencies in the joint text. Peculiar to text *A* might be, e.g. words denoting *earth, furrow, cow, whip, (to) draw, (to) harness* (see Table 3), which suggest the rural topic of the novel. On the other hand, there are 9 words in the column of 0 - 5 distribution, i.e. words which are missing in text *A* but occur five times in text *B*. As the expected number of such words is 2, we can conclude that at least 7 words of 9 might be characteristic exclusively of text *B*. Indeed, such words as (in English translation) *town, mirror, letter, (to) lie* (on a bed), *to cry* (weep) point to a different subject matter as compared with text *A*. The analysis may be continued, considering the other columns of the table. Our method does not allow us to establish which words should be considered peculiar to one or another text; for such an analysis a special "filter" should be set up as a result of the comparison of data from many different texts. At this stage we can state that the method will give additional information that can be used for objective stylistic analysis of the lexical composition of texts.

7. Finally, one more question has to be answered. When calculating the theoretical values of the distribution of subfrequencies in the model of the joint text, we proceeded from the observed distribution of word frequencies (f_m values in Table 2). The question arises whether the word frequency distribution in the joint text, which is artificially combined on the basis of two individual texts, would have the same statistical qualities as the distribution from a "natural" homogeneous text. As is well known, frequency distributions in texts are considered to be governed by some general (however approximate) laws of probabilistic nature. The word frequency distribution ("lexical spectrum") has usually been connected with Zipf's law or the law of lognormal distribution (Carroll 1967) or the law of Waring-Herdan (Herdan 1964; Muller 1968), or some others (cf. Tuldava 1986). In order to ascertain whether the empirical results are in satisfactory concordance with some of the theoretical distributions, we shall compare our observation and theory on the basis of Zipf's model and the Waring-Herdan distribution (Table 4).

Table 3
The distribution of words with frequency 5 (m = 5) in the joint text

A - B 5 - 0	A - B 4 - 1	A - B 3 - 2	A - B 2 - 3	A - B 1 - 4	A - B 0 - 5
muld 'earth'	värav 'gate'	taevas 'heaven'	algus 'beginning'	huvi 'interest'	linn 'town'
vagu 'furrow'	äär 'edge'	jutt 'story'	rahvas 'people'	voodi 'bed'	peegel 'mirror'
lehm 'cow'	lähedus 'nearness'	peremees 'landlord'	lôug 'chin'	trepp 'stairs'	kiri 'letter'
piits 'whip'	jälgima '(to) watch'	jagama '(to) divide'	jôud 'strength'	poeg 'son'	aru 'reason'
vedama '(to) draw'	viskama '(to) throw'	lôikama '(to) cut'	juus 'hair'	jätma '(to) leave'	lamama '(to) lie'
rakendama '(to) harness'	lahtine (adj.) 'open'	loopima '(to) fling'	lendama '(to) fly'	sulgema '(to) close'	nutma '(to) cry'
kahmama '(to) grab'	magus 'sweet'	riputama '(to) hang'	môjuma '(to) affect'	üksnes 'only'	nôudma '(to) demand'
kobama '(to) fumble'	kaugel 'far away'	raske 'heavy'	uurima '(to) examine'	-	ehitama '(to) build'
pitsitama '(to) squeeze'	liiga 'too'	mingi (pron.) 'some'	külm 'cold'	-	kohta (prep.) 'about'
iseenese '(of) himself'	too (pron.) 'that'	siiski 'still'	soe 'warm'	-	-
kannul 'at heels'	-	...	...	-	-
(O) 11	10	15	14	7	9
(E) 2	10	21	21	10	2
(O-E) +9	-	-6	-7	-3	+7

O - observed number, E - expected number
-ma is infinitive suffix

We calculated the theoretical values for Zipf's model using the formula

(3) $$f_m = Cm^{-\alpha}$$

where f_m denotes the number of words with frequency m; C and α are constants. From the data table provided, we can see that, on the whole, the observed distribution of word frequencies in the joint text may obey Zipf's law, but we also

Table 4
Observed number of words f_m with frequency m and expected values calculated using the formulas

$f'_m = Cm^{-\alpha}$ (*Zipf*)

$f''_m \Rightarrow f_{i+1} = \frac{a + i - 1}{x + 1} f_i$ (*Waring - Herdan*)

m	Text A		Text B		Text A + B		
	f_m	f'_m	f_m	f'_m	f_m	f'_m	f''_m
1	948	878	860	773	1808	1563	1808
2	300	273	226	237	406	405	417
3	120	137	109	119	163	184	180
4	72	85	64	73	89	105	99
5	57	58	55	50	66	68	62
6	38	43	31	37	39	48	42
7	40	33	35	28	42	35	30
8	27	26	28	23	30	27	23
9	21	22	16	18	21	22	18
10	20	18	18	15	20	18	14
>10	125	-	122	-	125	-	116
Σ	1768	-	1564	-	2809	-	2809
C		878		773		1563	
α		1.685		1.701		1.948	

see a considerable deviation from the theoretical value in the region of words with frequency 1. On a graph with bi-logarithmic scale we would see here a bend upwards from the straight line (which, in principle, could be corrected by adding a new constant, analogous to Mandelbrot's constant B, e.g. in the form of $f_m = C(m + B)^{-\alpha}$). This deviation is quite in accordance with the fact that the joint text has been composed of two independent texts with their own vocabularies and differences especially in the region of rare words - hence the accumulation of *hapax legomena* (words with frequency 1) in the joint text. Normally, the relative amount of rare words will decrease with increasing text length, but here we see the reverse: the percentage of words with frequency 1 is 53.6 in text *A*, 54.0 in text *B*, but 64.4 in the joint text *A+B* (see Table 1).

8. Taking into account the fact that there is an increased number of words with frequency 1 in the joint text, we shall try to ascertain whether the observed distribution of word frequencies as a whole is in satisfactory concordance with the Waring-Herdan distribution. This distribution can be calculated on the basis of the given amount of words with frequency 1 (f_1) in addition to the text length (N) and vocabulary size (V). The distribution represents a decreasing series of numbers; it was originally constructed by the English mathematician Edward Waring (1734-1798) and applied to linguistic processes by Gustav Herdan (1964).

There are several ways of calculating the Waring-Herdan distribution (for details see Altmann & Zörnig 1992: 167-175). We shall use one of them and start with fixing the initial data in our experiment (cf. Table 1):

$$N = 10{,}000; \qquad V = 2{,}809; \qquad f_1 = 1{,}808.$$

As the next step we have to construct the "Waring series" with the parameters a and x:

$$(4) \qquad \frac{x - a}{x} + \frac{(x - a)a}{x(x + 1)} + \frac{(x - a)a(a + 1)}{x(x + 1)(x + 2)} + \ldots$$

$$\ldots + \frac{(x - a)a(a + 1)(a + 2)}{x(x + 1)(x + 2)(x + 3)} \ldots \frac{(a + n - 1)}{(x + n)} = 1.$$

The concrete values of a and x can be calculated by means of intermediate values M and Q:

$$M = V/N = 2809/10000 = 0.2809;$$
$$Q = (1 - f_1/V)^{-1} = (1 - 1808/2809)^{-1} = 2.8058.$$

And

$$a = (Q - M - 1)^{-1} = (2.8058 - 0.2809 - 1)^{-1} = 0.6558;$$
$$x = aQ = 0.6558(2.8058) = 1.8401.$$

Formula (4) allows us to calculate the decreasing series of probabilities (p_i) of the occurrence of words with frequences 1, 2, ..., n. For example:

$$p_1 = \frac{x - a}{x} = \frac{1.8401 - 0.6558}{1.8401} = 0.6436;$$

$$p_2 = \frac{(x - a)a}{x(x + 1)} = \frac{a}{x + 1}p_1 = \frac{0.6558}{2.8401}0.6436 = 0.1486;$$

$$p_3 = \frac{(x - a)a(a + 1)}{x(x + 1)(x + 2)} = \frac{a + 1}{x + 2}p_2 = \frac{1.6558}{3.8401}0.1486 = 0.0641;$$

etc.

Multiplying every probability by the total number of words in the vocabulary of the joint text ($V = 2809$), we get the theoretical distribution of word frequencies (f''_m values in Table 4):

$$0.6436(2809) = 1807.9 \approx 1808;$$
$$0.1486(2809) = 417.4 \approx 417;$$
etc.

The Waring series can be calculated directly as a series of absolute values $f_1, f_2,...$ by means of the recurrence formula

$$(5) \qquad f_{i+1} = \frac{a + i - 1}{x + i}f_i$$

where i denotes the frequency of words ($i = 1,2,...$).

Starting with the given $f_1 = 1808$ and calculating the following numbers in the series, we get

$$f_2 = \frac{0.6558 + 1 - 1}{1.8401 + 1}1808 = 417.46 \qquad (\approx 417);$$

$$f_3 = \frac{0.6558 + 2 - 1}{1.8401 + 2}417.46 = 180.00 \qquad (= 180);$$

$$f_4 = \frac{0.6558 + 3 - 1}{1.8401 + 3}180 = 98.77 \qquad (\approx 99);$$

etc.

The results coincide with those calculated by means of formula (4).

In order to test the goodness of fit between prediction and observation we shall use the chi-square test. For the number of degrees of freedom in question

($df = k - 4 = 11 - 4 = 7$) the chi-square requires at the 0.05 level $\chi^2 = 14.1$. The result of the test is $X^2 = 13.8$, which does not reach the critical value. This shows that the differences between the theoretical and empirical distributions are not greater than what might be expected by pure chance.

Thus, on the whole, our results are in good conformity with the Waring-Herdan model: with increasing frequency the number of words of that frequency decreases, but the decrease slows down as we proceed along the frequency scale according to the degressive Waring series. C. Muller (1976) has explained this process of "uneven transition" in the following way: "Each time the speaker pronounces or writes down a word, he transfers one lexical unit item from frequency i to frequency $i + 1$; at the same time, he reduces the subset f_i by 1 and increases the subset f_{i+1} by 1. Experience shows that each subset f_i tends to gain more than it loses. Consequently the transitions from f_i to f_{i+1} are more numerous than between $f_{i+1}$to f_{i+2}." And if the Waring series gives a good description of the relations between successive f_i values in a linguistic distribution, it can be accepted as an adequate model for the transition from f_i to f_{i+1} (cf. formula 5), and, accordingly, as a model for the word frequency distribution. Our own experience shows that the application of the Waring-Herdan model gives good results for texts up to $N = 200,000$ in various languages (Tuldava 1986).

We can therefore conclude that the observed word frequency distribution in the joint text, composed of two independent subtexts, may be qualified in our experiment as quite a "normal" linguistic distribution, and the experiment based on the joint text and the joint vocabulary, as described above, may be fully justified.

References

Altmann, G & Zörnig, P. (1992). *Diskrete Wahrscheinlichkeitsverteilungen II.* Bochum, Brockmeyer.

Barlow's Tables (1947[4]), edited by L.J. Comrie.

Carroll, J.B. (1967). On sampling from a lognormal model of word-frequency distribution. In: Kučera, H. & Francis, W.N. (eds.), *Computational analysis of present-day American English.* Providence, R.I., Brown University Press: 406-424.

Herdan, G. (1964). *Quantitative linguistics.* London, Butterworths.

Muller, C. (1968). *Initiation à la statistique linguistique.* Paris, Larousse.

Muller, C. (1976). Some recent contributions to statistical linguistics. *Statistical Methods in Linguistics 1976, 136-147.*

Tuldava, J. (1986). O častotnom spektre leksiki teksta (On the frequency spectrum of text and vocabulary). *Acta et Commentationes Universitatis Tartuensis 745, 139-162.*

8

On the Relation between Text Length and Vocabulary Size

The question of how to express the dependence of the size of vocabulary on the length of the text is of practical importance to the resolution of a series of problems in quantitative linguistics, for example, in order to determine the sample size needed to establish a given degree of saturation of vocabulary, or to investigate the stylostatistical characteristics of individual or functional style. The problem is also connected with the theoretical explanation of some important aspects of text generation. In this paper a simple method for calculating vocabulary size as a function of text length is discussed. The statistical fit and the possibility of extrapolation are illustrated on the material of several languages.

1. Introduction

A great number of attempts have been made to construct an appropriate mathematical model which would express the dependence of the size of vocabulary on the size of the text. Some scholars would try to derive the formulas from theoretical consideration, for example, by basing their considerations on the hypothesis about the lognormal distribution of words in a text (e.g. Carroll 1967) or by deducing the relation between the size of the text and the size of vocabulary from some other important quantitative characteristics of the text such as Zipf's law (cf. Kalinin 1964; Orlov 1976) or by determining the relation axiomatically on the basis of a priori natural combinatorial considerations (Krylov 1985), etc. In practice various empirical formulas are often used. They are usually simpler and describe the growth of vocabulary with sufficient exactitude in a number of cases. The first empirically constructed formulas were suggested as early as in the 1940s and 1950s, for example, by Chotlos (1944), Kuraszkiewicz (1958), Guiraud (1959), Somers (1959), and later by Müller (1971), Tuldava (1974), Nešitoj (1975) and others. In recent years new attempts have been made to analyze the relation between the size of the text and size of vocabulary against the background of some inner factors of text generation. For instance, the dynamics of vocabulary growth is viewed as a result of the interaction of several linguistic and extralinguistic factors which in an integral way are governed by the principle of "the restriction of variety" of the vocabulary, resulting in a sys-

tem of functions based on the allometric law of growth (cf. Tuldava 1980).

In this article we shall examine the problem from another point of view with the aim of developing a simple model for practical use which is constructed on the assumption that definite stages (phases) characterized by regular interdependence between text length and vocabulary size exist in the process of text generation.

2. Some definitions and conventions

As a preliminary we should agree on the meaning of some notions and terms. Formally a text is a linear sequence of linguistic units which in everyday usage are called words, in quantitative linguistics also *occurrences*, *recurrences*, *repetitions*, *usages* or *"tokens"* (cf. Herdan 1960; Williams 1970). Where there is no doubt, the term *word* is often used without qualification (e.g. "the number of words in a sentence"). The size (length) of the text, i.e. the total number of word occurrences is in our study designated by the symbol *N*. The size of the corresponding vocabulary is the sum total of different words which occur in the given text and which are usually presented in the form of a list (often in the form of a frequency dictionary). The vocabulary units may be of two kinds. First, we can compile a list of *word forms*, i.e. words in such a form as they actually occur in the text. Second, we can unite different forms under a common denominator ("lemma"), usually under the so-called principal form (for nouns the common case or nominative, for verbs the infinitive). Such "lemmatized" units of vocabulary are called *lexemes*. The size of the vocabulary consisting of word forms is designated (in our study) by the symbol *V*, the size of the vocabulary of lexemes - by the symbol *L*. Vocabulary units (both word forms and lexemes) may be called *vocabulary words* or "*types*".

These data, in particular the relations between the different numerical characteristics *N*, *L* and *V*, describe the structure of the text from the quantitative point of view. The ratio *L/N* (or *V/N*) expresses the relation between the number of different lexemes (or word forms) and the number of word occurrences in a given text. This ratio is called "the index of diversity (variety)" of the vocabulary or simply the "type-token ratio" (abbreviated as "TTR"). The larger the value of this index, the more different words have been used by the writer or speaker in the text (speech) studied. The inverse ratio *N/L* (or *N/V*) gives us the average frequency of words (lexemes or word forms) in the text. The bigger this relation, the less lexical variety there is in the text. Thus, both the index of diversity (TTR-index) and the average usage per vocabulary word (average word frequency) enable us to make some preliminary conclusions about the distribution and concentration of words in a text. The relation between the number of lexemes and word forms (*L/V* or *V/L*) defines the degree of analyticism/ syntheticism of the language (see the next article in this issue).

At that it should be kept in mind that the above-mentioned quantitative characteristics can by no means serve as criteria by which one can judge the aesthetic and other qualities of style. By themselves, they only reflect some structural, and perhaps, functional peculiarities of the text. But as "the quantitative and qualitative sides of human speech are correlated and interdependent" (Piotrowski 1968), the quantitative characteristics may serve as signals calling the researcher's attention to the qualitative features of individual or functional styles (genres) that may remain unnoticed by simple observation.

3. Text phases

First of all, it should be emphasized that a direct comparison of quantitative text characteristics of different texts (or different parts of a text) is possible only in the case of equal sizes of texts or text fragments. The relation between the size of text and the size of vocabulary does not remain unchanged at any value of text size. The size of the text (*N*) increases faster than the size of the vocabulary (*L* or *V*). With the growth of text size the words which have occurred in the text before will recur more and more often, whereas the increase of "new" words slows down. Generally, we can distinguish between several phases or stages in the process of text generation:

(1) The text begins inevitably by $N = L$ (or V) $= 1$. It may happen that in the very beginning, and in the case of a small size of the text, for instance, if $N < 50$, all the words in the text are different, and thus $N = L$ or $N = V$.[1] However, the process of word recurrence will begin very soon and, on the whole, we can speak of the *initial* phase of text generation which is dominated by an intense accumulation process of the most frequent words - mostly "function words" and topical "key words". This initial phase usually ends - depending on the type of language - at a text length of about 2,000 - 3,000 words.

(2) When the vocabulary is saturated with function words and topical key words, the process reaches the following, *medial* phase of text generation. Now the accumulation of vocabulary items happens mainly on account of words of

[1] An example where $N = V$: *"My purpose in writing this book is clear from its title: I want to show how various aspects of language can, and for many reasons must, be set within a quantitative-systemic model".* (A preface.)

medium frequency (defined separately for various texts or genres). For homogeneous texts this phase continues approximately up to N = 20,000 - 30,000 words (cf. Allén 1970).

(3) The moment will arrive when nearly all words of high and medium frequency have appeared in the text, and the further growth of vocabulary occurs only on account of rare words (provided we have to do with a homogeneous text; see below). This *final* phase of text generation can last quite long depending on the type and language. Now the point is that when the text size grows very large, a moment may arrive when all the active vocabulary of the speaker or writer (or a group of them) is exhausted and the appearance of new words will stop or be insignificantly small. An appropriate analytical model for homogeneous texts could therefore include an asymptote pointing to the *limit* of the individual vocabulary.

The several phases of text generation may differ from each other in details depending on the type of the text, genre and language, but they are probably similar in principle: in the multi-phase character of text formation.

4. Experiment

As an example we present the data from A. Zacharova's experiment (1967) obtained by studying the oral speech of a group of seven-year-old Russian children. Table 1 gives the numerical data of this experiment which illustrates the growth of the vocabulary of lexemes when the length of the text increases. In columns 3, 4 and 5 of Table 1 we can see the gradual decrease of the number of "new" words (denoted by ΔL) and the changes in the relations of L/N and N/L. The numerical values of L/N, the index of lexical diversity (variety), is steadily decreasing when the length of the text (N) increases, but the values of N/L, the average word frequency, is steadily increasing with the increase of text length.

As has been pointed out by Tuldava (1980), the relation between L/N (or V/N) and N in the initial stages of text formation is governed by the allometric law expressed in the form of a power function. For an adequate analysis of the relation in the following text phases, the allometric formula has to be modified by logarithmization.

In this study we shall take the inverse relation, i.e. the average word frequency N/L (or N/V), as a starting point for the construction of a formula which allows linguistic interpretation of its parameters.

We start with the practically and theoretically corroborated thesis of a regular interdependence between the average word frequency and the text size in certain

text intervals, corresponding to various text phases.[2] Inspection of column 6 in Table 1 reveals the fact that in the range of N = 3,000 to N = 12,000 the relation between the increase of N/L and the increase of N remains practically constant, i.e.

$$\frac{\Delta (N/L)}{\Delta N} = const.$$

This means that there is a linear relation between N/L and N in the given text interval, which can also be seen on Figure 1: the graph rises steeply to a nuclear point (at N = 3,000) and thereafter it continues less steeply in a straight line up to N = 12,000. The nuclear point apparently divides the initial phase from the medial phase of text formation.

Expressing the linear relation between N/L and N by the equation

$$N/L = \alpha + \beta N \quad (1)$$

Table 1
Dynamics of vocabulary growth in children's speech
(Russian)

N	L	ΔL	L/N	N/L	Δ(N/L)/ΔN
1000	386	368	0.368	2.717	5.8 (10^{-3})
2000	607	239	0.304	3.295	3.8 (10^{-3})
3000	817	210	0.272	3.672	3.0 (10^{-3})
4000	1007	190	0.252	3.972	3.2 (10^{-3})
5000	1164	157	0.233	4.296	3.0 (10^{-3})
6000	1305	141	0.218	4.598	3.0 (10^{-3})
7000	1428	123	0.204	4.902	3.2 (10^{-3})
8000	1530	102	0.191	5.229	3.3 (10^{-3})
9000	1617	87	0.180	5.566	3.6 (10^{-3})
10000	1716	79	0.172	5.828	3.5 (10^{-3})
11000	1782	66	0.162	6.173	3.2 (10^{-3})
12000	1849	67	0.154	6.490	

[2] Cf. Williams (1970: 92) who examines the relation between N/L and N as a possible basis for model construction, and Ejiri & Smith (1992) who use the linear relation between the logarithms of N/L and N as the basis for the construction of a formula for the text-vocabulary relation.

where α and β are constants, we obtain the formula for the dependence of vocabulary size on text size

$$L = \frac{N}{\alpha + \beta N} \tag{2}$$

or, in a more convenient form,

$$L = \frac{aN}{N + b} \tag{3}$$

where $a = 1/\beta$ and $b = \alpha/\beta$. Formula (3) represents the well-known Tornquist function of the first type (cf. Förster & Rönz 1979, chap. 5.1) defined in a general form as

$$y = \frac{ax}{x + b}$$

where both constants a and b are asymptotes (see Figure 2).[3]

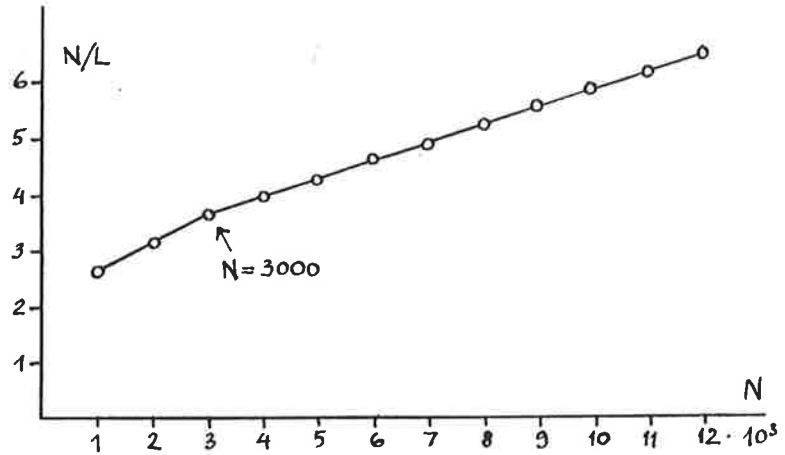

Figure 1. The relation between increase in text length (N) and increase in average word frequency (N/L)

[3] From the mathematical point of view Tornquist's formula is a particular case of the so-called "fraction-linear" function y = (ax + d)/(cx + b) when $d = 0$ and $c = 1$. (Cf. Markovič 1972:91.)

The constants can be interpreted as follows:

Constant a is an asymptote pointing to the limit of y if x increases infinitely:

$$y = \frac{ax}{x + b} = \frac{ax}{x(1 + b/x)} = \frac{a}{1 + b/x};$$

as $\lim_{x \to +\infty} b/x = 0$, then

$$\lim_{x \to +\infty} y = a.$$

Constant b is an asymptote which determines the form of the function, e.g. if a remains unchanged, a small value of b would mean a steep rise in the initial phases followed by a moderate rise later on; a large value of b would mean a slow rise from the very beginning.

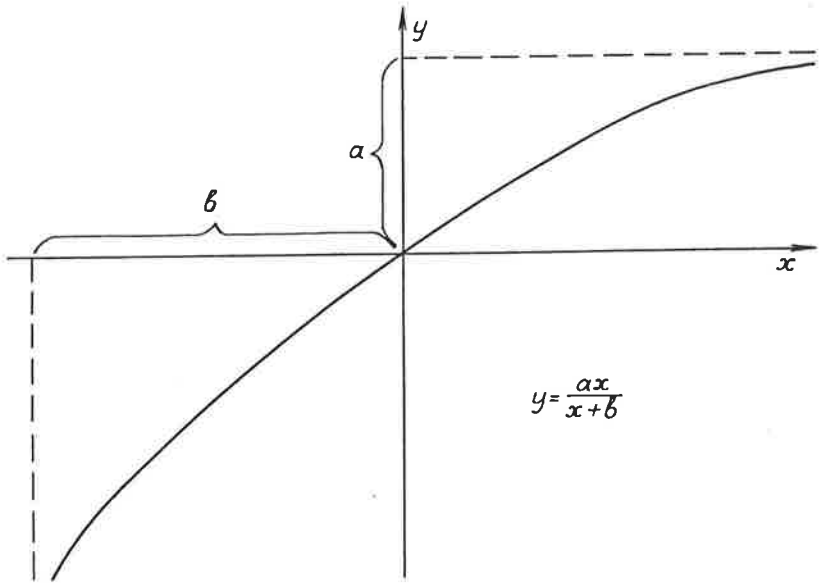

Figure 2. Interpretation of the parameters a and b in the Tornquist function of the 1st type

The Tornquist function has been widely used in various fields of science, particularly in economics, as an analytic means for expressing "adapted growth rate"; for instance, when analyzing the relation between income (x) and expenditure (y). (Cf. Allgemeine Statistik 1967, chap.6.5.2).

5. Preliminary results

Applying the Tornquist function to our experimental data, we have to start with the calculation of the expected values of constants α and β in the range of N = 3,000 to N = 12,000 by means of the method of least squares. The calculation according to formula (1) yields the result:

$$\alpha = 2.7243 \quad \text{and} \quad \beta = 0.00031308 .$$

The next step is to find the values of a and b in formula (3):

$$a = 1/\beta = 1/0.00031308 \approx 3194 \quad \text{and}$$

$$b = \alpha/\beta = 2.7243/0.00031308 \approx 8702 .$$

Thus, the formula for calculating the expected values of vocabulary size (L') is the following:[4]

$$L = \frac{3194\ N}{N + 8702}.$$

The fitting of the function to observed data is presented in Table 2. As expected, the fit is actually very good. Even without statistical testing it seems clear that the dynamics of the dependence of vocabulary size on text size in the

[4] Formula (3) means also a linear relation between 1/L and 1/N:

$$L = \frac{aN}{N + b} \rightarrow \frac{1}{L} = \frac{N + b}{aN} = \frac{N}{aN} + \frac{b}{aN} = 1/a + b/a(1/N).$$

Taking 1/L = Y, 1/a = A, b/a = B and 1/N = X, we get the linear function Y = A + BX. The calculation on our material gives the result: a = 3205 and b = 8754 (the small difference between the results of the calculations is due to round-off error).

given text is governed by the law of "adapted growth rate". Alternatively, or additionally, we can state that there is a linear relation between average word frequency (N/L) and text size (N) in the same text interval. As a hypothesis we would like to consider the linear relation between N/L and N in a given text interval, corresponding to a certain phase in text formation, a formal (statistical) characteristic of a *homogeneous text* (or *text fragment*) whereas a deviation from linearity would be regarded as a "deviation from homogeneity". Such a hypothesis is supported by analogous argumentation in linguo-statistical practice, for instance, by the thesis of homogeneity of text structure on the basis of "Zipf's size" (Orlov 1982) which leads to the rank frequency distribution of words according to Zipf-Mandelbrot's formula of the type $y = a(x+c)^{-b}$ when $b = 1$. As a fact of coincidence, the Tornquist formula (3) can be expressed in the form

(4) $$L/N = a(N + b)^{-1}$$

which is formally identical with the "canonical" Zipf-Mandelbrot's formula mentioned above with $b = 1$.

Table 2
Fitting of function (3) to observed data

N	L	L'	d = L - L'
3000	817	819	- 2
4000	1007	1006	+ 1
5000	1164	1166	- 2
6000	1305	1304	+ 1
7000	1428	1424	+ 4
8000	1530	1530	0
9000	1617	1624	- 7
10000	1716	1708	+ 8
11000	1782	1783	- 1
12000	1849	1851	- 2
n = 10	-	-	$d_i = 0$

Returning to our experiment, we can state that according to the results of the analysis of children's speech within the limits of

$$3{,}000 < N < 12{,}000$$

the dependence of L on N can quite adequately be described by means of Tornquist's function of the first type, i.e. by formula (3). This also means that there is a linear relation between the average frequency of word occurrence (N/L) and the size of the text (N). Furthermore, the dependence between the index of lexical diversity or TTR-index (L/N) and the size of the text (N) can be expressed by a power function which is formally identical with Zipf-Mandelbrot's law in its simplest form (with the power factor $b = 1$).

If we presume that within the limits of $N = 3{,}000 - 12{,}000$ the form and the parameters of function (3) are mostly determined by the replenishment of the vocabulary with "new" words of medium frequency (certainly in interaction with the appearance of rare words, although at this stage the former effect dominates over the latter), we can conclude that in this "medial phase" of text formation there is full evidence for lexical homogeneity of the text examined from the statistical point of view.

6. Interpretation

According to Tornquist's function, the limit of the function expressed by the asymptote a allows us to estimate the potential size of the vocabulary on condition that the text under examination remains homogeneous beyond the limits of the observed text interval. In our case $a = 3194$ (see above), which means that the potential limit of the vocabulary of a group of seven-year-old Russian children, according to the experimental data, may be approximately 3,200 lexemes (with certain confidence limits). Of course, constant a in formula (3) need not be interpreted as the real limit of the vocabulary, but simply as a stylostatistical parameter expressing the tendency of vocabulary growth ("potential richness of the vocabulary") of the given text. The other constant in the formula (b), or, to be more exact, its relation to constant a, i.e. the ratio a/b, may be interpreted as an index pointing to the form of vocabulary growth: the greater the ratio a/b, the more intensive is the vocabulary growth in the initial stages of text formation. Thus, both constants, separately or in combination, can serve as differentiating stylistic characteristics of texts (for a comparative analysis of 20 texts see Tuldava 1977).

For further interpretation of function (3) the so-called indices of *elasticity* may be of interest (cf. Allgemeine Statistik 1967, chap. 6). Two indices can be used.

The index of "absolute elasticity" (δ) is based on the first derivative of function (3):

$$\delta(N) = \frac{dL}{dN} = \frac{ab}{(N + b)^2} \tag{5}$$

and the index of "relative elasticity" (ε) is defined by

$$\varepsilon(N) = \frac{dL/L}{dN/N} = \frac{b}{N + b}.^{5} \tag{6}$$

The indices of elasticity show the dynamics of vocabulary growth with increasing text length at various stages of text formation.

The index of absolute elasticity (5) indicates the relation between the increases of L and N in absolute values: if N increases by x units, then L increases by δx units. For instance, in our example (see Table 2) the absolute elasticity at the point $N = 3{,}000$ equals

$$\delta(3{,}000) = \frac{3194(8702)}{(3000 + 8702)^2} = 0.203.$$

That is, if at that point N increases by 100 units, then we can prognosticate that L increases by $0.203(100) = 20.3$ units on average. The increase of L will slow down with increasing text length. At the point $N = 10{,}000$ we get

$$\delta(10{,}000) = \frac{3194(8702)}{(10000 + 8702)^2} = 0.08,$$

i.e., here the increase of N by 100 units would add only eight new vocabulary units. Extrapolating (assuming homogeneity beyond the limits of the observed text interval), we get

$$\delta(20{,}000) = 0{,}03;\ \delta(30{,}000) = 0{,}02;\ \text{etc.}$$

The index of relative elasticity (6) expresses the dependence of L on N at various stages in relative values. If N increases by a small quantity, e.g. by 1 %

[5] The equation can be rewritten as

$$(dL/dN)(N/L) = \frac{ab}{(N + b)^2}\left\{\frac{N}{(aN)/(N + b)}\right\} = \frac{b}{N + b}.$$

of its size, then L increases by ε%. In our example, at the point N = 3,000, the relative elasticity is

$$\varepsilon(3{,}000) = \frac{8702}{3000 + 8702} = 0.74,$$

i.e., if N increases by 1%, L increases by 0.74 %. At the point N = 10,000 the relative increase of L would be 0.47 %; at the point N = 20,000 = 0.30 %; at the point N = 30,000 = 0.22 %; etc.

The indices of absolute and relative elasticity may be used for prognostication (extrapolation as well as interpolation) of vocabulary growth at various stages of text formation. The indices can also be used for solving stylostatistical tasks, e.g. while comparing lexical structures of various texts.

7. Other examples

It would be interesting to see to what extent some other texts correspond to the requirements of statistical homogeneity as defined in Section 5. As illustrative material we shall examine texts which can be considered to be "nominally" homogeneous, in the first place texts written by one author. Of the existing empirical material we have chosen the data of the whole of the story "The Queen of Spades" by A.Puškin (after Nadarejšvili & Orlov 1982), in Russian. Table 3 presents the data on vocabulary growth (in lexemes) compared with the increase of text length in the range of N = 2,000 to N = 6,861 (total text size). As can be seen from Table 3, there is good conformity between the observed and the expected values of L: the sum of deviations (residuals) $\Sigma d_i = 0$ and standardized residuals d_i / s_d do not exceed 2.0.

We can state that this story of A.Puškin can be considered formally (statistically) homogeneous with regard to the lexical structure of the text. The linear relation between the average word frequency (N/L) and the size of text (N) is valid for the text interval beginning with N = 2,000. Qualitative analysis will show that in this story the vocabulary is distributed very evenly and the author does not move far away from the main topic of the story.

In some other cases, when the individual text is not quite homogeneous in its contents (e.g. due to the rise of new themes in the course of text generation), the rate of increase of vocabulary becomes more rapid and the relation between N/L and N approaches a non-linear relation. Consequently, the original function (3) does not fit very well and we have to modify it. The best way is to add an

adjustment factor (c) which gives us Tornquist's function of the second type (cf. Förster & Rönz, chap. 5.1):

(7) $$L = \frac{a(N + c)}{N + b}$$

Table 3
Fitting of function (3) to the data of A.Puškin's "The Queen of Spades" (Russian)

N	L	L'	d = L - L'	d_i / s_d
2000	787	786	+ 1	+ 0.08
4000	1348	1347	+ 1	+ 0.08
6000	1752	1766	- 14	- 1.07
6861	1928	1916	+ 12	+ 0.92
$L' = \frac{4693\ N}{N + 9940}$			$\Sigma d_i = 0$	$\Sigma d_i^2 = 342$
			$s_d = (\Sigma d_i^2/(n-2))^{1/2} = 13.08$	

where a, b and c are constants. In this case there is a linear relation between $(N+c)/L$ and N (instead of between N/L and N).

The calculation of parameter c requires iterative solution (a feasible solution should give non-systematic deviations, $d_i/s_d < 2.0$, and $\Sigma d_i \approx 0$). Practically, parameter c can be considered a measure of the degree of deviation from linearity in the relation between N/L and N, and, by definition, it may be considered a measure of deviation from statistical homogeneity of the text with regard to the distribution of words in the process of text generation. In relative values the degree of deviation can be expressed by means of index K defined by

(8) $$K = \frac{c}{N_1 - N_n}$$

where N_1 denotes the initial and N_n the final text size under examination.

As an illustration we shall compare the results calculated by formulas (3) and (7) on the material of two samples from Russian: the first 10,000 word occurrences in the novel "Resurrection" and in the story "The Cossacks" by L.Tolstoy

(the initial data have been taken from Nadarejšvili & Orlov 1982). Table 4 presents the fitting of the two formulas to observed data.

At first we shall examine the results calculated by formula (3) on the texts in the range of N = 2,000 to N = 10,000 (i.e. in the medial phase of text formation). We see that in spite of low chi-square values (0.87 and 4.22, respectively, while the critical value at the 0.05 level and for four degrees of freedom is 9.49), the distribution of plus and minus signs of the residuals (d_i) shows systematic deviation from the expected values of L. We have to make the conclusion that the growth rate of vocabulary in these texts is not quite uniform as expected on the assumption of full "homogeneity" in the medial phase of text formation (presuming a linear relation between N/L and N). (Nevertheless, as the deviations are quite small, we can use the results for approximate estimation of

Table 4
"Resurrection" and "The Cossacks" by L.Tolstoy (Russian)
(After Nadarejšvili & Orlov 1982)

	"Resurrection"			"The Cossacks"		
N	L	L' (d_i)	L'' (d_i)	L	L' (d_i)	L'' (d_i)
500	282	-	288 (-6)	274	-	274 (0)
1000	489	-	478 (+11)	438	-	432 (+6)
2000	824	812 (+12)	824 (0)	732	709 (+23)	733 (-1)
4000	1409	1416 (- 7)	1403 (+6)	1233	1298 (-65)	1284 (-51)
6000	1863	1883 (-20)	1870 (-7)	1814	1796 (+18)	1775 (+39
8000	2237	2255 (-18)	2253 (-16)	2253	2224 (+29)	2216 (+37)
10000	2587	2559 (+28)	2575 (+12)	2582	2592 (-10)	2612 (-30)
Σd_i		- 5	0		- 5	0
$s_d = \sqrt{\frac{\Sigma d_i^2}{n-2}}$		23.8	11.3		44.8	35.9
X^2		0.87	0.60		4.22	3.93
$K = \frac{c}{N_1 + N_n}$			- 0.021			- 0.035
Formulas		$L' = \frac{5543\ N}{N + 11630}$			$L' = \frac{7728\ N}{N + 19813}$	
		$L'' = \frac{6200(N + 200)}{N + 14562}$			$L'' = \frac{10297(N + 330)}{N + 30727}$	

the tendencies of vocabulary growth in the two texts compared: the asymptote a for "The Cossacks" equals 7728, as against 5543 for "Resurrection", i.e. the first text is characterized by a larger potential vocabulary; the results of the calculation can also be used for truthful interpolation within the limits of the text interval examined.)

We get much better results when we apply the modified (corrected) Tornquist formula (7) to our data (see Table 4). The formula fits well not only for the medial phase of text formation but also when the initial phase is included (beginning with N = 500). The chi-square test shows good conformity between the observed and the expected values of L for both texts (X^2 equals 0.60 and 3.93 respectively, while the critical value is 12.59 at the 0.05 level and with six degrees of freedom). The plus and minus signs of the residuals are distributed evenly with $\Sigma d_i = 0$, and the standardized residuals d_i/s_d do not exceed 2.0.

Comparing the two texts on the basis of index K (showing deviation from "homogeneity"; see formula 8), we can see that the sample taken from "The Cossacks" seems to be less homogeneous with regard to the distribution of words in the process of text generation (K = 330/(500-10000) = - 0.035 for "The Cossacks" and K = 200/(500-10000) = -0.021 for "Resurrection"). A plausible explanation of this difference lies in the fact that the text fragment of "The Cossacks" consists of two contrasting episodes with somewhat differing vocabularies; the sample from "Resurrection" also includes contrasting episodes, but these alternate in small portions, which, on the whole, leads to a more uniform distribution of words (cf. Nadarejšvili & Orlov 1982: 75-76).

The modified Tornquist formula (7) seems to be an adequate model for describing the relation between text length and vocabulary size in general. It is appropriate not only for correcting the relation between the average word frequency and the text size within the limits of the medial phase of text formation but also for modelling vocabulary growth beyond the text phases, whereas the initial formula (3) expresses a particular case of the relation - when there is real uniformity of vocabulary growth in comparison with the increase of text length ("homogeneity" by definition).

The appropriateness of formula (7) can be demonstrated on several examples taken from investigations of texts from various languages, with regard to both lexemes (L) and word forms (V). Some further examples:

Table 5 presents calculations by means of formulas (3) and (7) on the Polish poem "Pan Tadeusz" by A.Mickiewicz (after Sambor 1970). The modified formula gives very good results. The same procedure has been applied to the analysis of German newspapers (after Billmeier 1968) within the limits of N = 15,000 to N = 60,000, and Latvian newspapers (Jakubaite 1969) within the limits of N = 50,000 to N = 200,000; see Tables 6 and 7. Finally, in Table 8

English and Rumanian texts on electronics (after Alekseev 1968 and Ešan 1966) have been examined. All examples show substantial overall agreement with the proposed analytical model of text-vocabulary relation.

Table 5
"Pan Tadeusz" by A.Mickiewicz (Polish)
(After Sambor 1970)

N	V	V'	(d_i)	V''	(d_i)
6587	2257	2130	(+ 127)	2255	(+ 2)
12172	3434	3504	(- 70)	3447	(- 13)
23861	5428	5587	(- 159)	5400	(+ 28)
48255	8026	8132	(- 106)	8060	(- 34)
64510	9250	9160	(+ 90)	9223	(+ 17)
Σd_i			- 118		0
s_d			147.9		28.3
X^2			15.76*		0.37
K			-		- 0.031
Formulas		$V' = \frac{19277N}{N + 90461}$		$V'' = \frac{16731(N+1800)}{N+55649}$	

Table 6
Fitting of functions (3) and (7) to German
newspapers (after Billmeier 1968)

N	L	L'	(d_i)	L''	(d_i)
15000	4280	4220	(+ 60)	4268	(+ 12)
30000	7472	7625	(- 153)	7535	(- 63)
45000	10460	10430	(+ 30)	10375	(+ 85)
60000	12832	12781	(+ 51)	12866	(- 34)
Σd_i			- 12		0
s_d			123.5		79.0
X^2			4.21		1.35
K			-		- 0.038
Formulas	$L' = \frac{39468\,N}{N + 125283}$			$L'' = \frac{50996(N+1700)}{N+184554}$	

Table 7
Fitting of functions (3) and (7) to Latvian
newspapers (after Jakubaite 1969)

N	L	L'	(d_i)	L''	(d_i)
50000	7065	6862	(+ 203)	7050	(+ 15)
100000	9834	10121	(- 287)	9892	(- 58)
150000	11976	12025	(- 49)	11909	(+ 67)
200000	13389	13273	(+ 116)	13413	(- 24)
Σd_i			- 17		0
s_d			215.6		65.8
X^2			15.36*		0.79
K			-		- 0.146
Formulas		$L' = \frac{19277N}{N+90461}$		$L'' = \frac{23767(N+22400)}{N+194073}$	

Table 8
Fitting of function (7) to observed data in English and Rumanian texts
on electronics (after Alekseev 1968 and Ešan 1966)

N	English V	English V'	English (d_i)	Rumanian V	Rumanian V'	Rumanian (d_i)
5000	5399	5407	(- 8)	6785	6791	(- 6)
100000	7853	7812	(+ 41)	10281	10228	(+ 53)
150000	9361	9416	(- 55)	12477	12563	(- 86)
200000	10582	10560	(+ 22)	14292	14253	(+ 39)
Σd_i			0			0
s_d			51.3			76.7
X^2			0.59			0.98
K			- 0.080			-0.053
Formulas	$V' = \frac{17425(N+1200)}{N + 149812}$			$V' = \frac{24784(N+8000)}{N + 161688}$		

* * *

As mentioned before, a number of methods exist for expressing the relationship between the length of the text and vocabulary size. The value of the models discussed in this article (Tornquist's functions of the 1st and 2nd type) consists in their simplicity and in the feasibility of linguistic interpretation of the parameters of the corresponding functions. The fact that these models can be used for studying different languages shows that the assumptions underlying them - especially regular connection between the average word frequency and the length of the text in certain phases of the text and the possibility of determining the limits of vocabulary in the case of a homogeneous text - are universal to a certain extent as they are valid for many languages.

References

Alekseev, P.M. (1968). Leksičeskaja i morfologičeskaja statistika anglijskogo pod'jazyka elektroniki (Lexical and morphological statistics of the English sublanguage of electronics). In: Piotrowski, R.G. (ed.). *Statistika reči.* Leningrad, Nauka.

Allén, S. (1970). Vocabulary data processing. In: *The Nordic Languages and Modern Linguistics.* Reykjavik.

Allgemeine Statistik (1967). Autorenkollektiv. Berlin, Verlag Die Wirtschaft.

Billmeier, G. (1968). Über die Signifikanz von Auswahltexten. In: Moser, H. (Hrsg.). *Forschungsberichte des Instituts für deutsche Sprache 2.* Mannheim, 126-171.

Carroll, J.B. (1967). On sampling from a lognormal model of word-frequency distribution, In: Kučera, H. & Francis, W.N. (eds.). *Computational Analysis of Present-day American English.* Providence, R.I., Brown University Press, 406-424.

Chotlos, J.W. (1944). Studies in language behaviour. A statistical and comparative analysis of individual written language samples. *Psychological Monographs 56, 75-111.*

Ejiri, K. & Smith, A.E. (1993). Proposal of a new 'constraint measure' for text. In: Köhler, R. & Rieger, B.B. (eds.). *Contributions to quantitative linguistics. Proceedings of the First International Conference on Quantitative Linguistics, QUALICO 1991.* Dordrecht etc., Kluwer, 195-211.

Ešan, L.I. (1966). *Opyt statističeskogo opisanija naučno-techničeskogo stilja rumynskogo jazyka (An attempt at statistical description of scientific-technical style in Rumanian).* Cand. thesis. Leningrad.

Förster, E. & Rönz, B. (1979). *Methoden der Korrelations- und Regressionsanalyse.* Berlin. Verlag Die Wirtschaft.

Guiraud, P. (1959). *Problèmes et méthodes de la statistique linguistique.* Dordrecht, Reidel.

Herdan, G. (1960). *Type-token mathematics.* The Hague, Mouton.

Jakubaite, T. (ed.) (1969). *Latviešu valodas biežuma vārdnica II: 1. (Frequency dictionary of the Latvian language).* Riga, Zinātne.

Kalinin, V.M. (1964). O statistike literaturnogo teksta (On statistics of literary texts). *Voprosy jazykoznanija, 123-127.*

Krylov, Ju.K. (1985). K voprosu o dinamike narastanija ob' jema slovarja slučajnoj vyborki i svjaznogo teksta (On the growth of vocabulary size in random samples and connected texts). *Acta et Commentationes Universitatis Tartuensis 711, 55-66.*

Kuraszkiewicz, W. (1958). Statystyczne badanie słownictwa polskich tekstów XVI wieku (Statistical investigation of the vocabulary of XVIth century Polish texts). In: Zwoliński, P. (ed.). *Z polskich studiów slawistycznych.* Warszawa, Państwowe wydawnictwo naukowe, 240-257.

Markovič, E.S. (1972). *Kurs vysšej matematiki (A course in higher mathematics).* Moscow, Vysšaja škola.

Müller, W. (1971). Wortschatzumfang und Textlänge. *Muttersprache 81, No. 4, 266-276.*

Naderejšvili, I.S. & Orlov, Ju.K. (1982). Die Methode der vollständigen Textfixierung durch eine linguistisch-statistische Analyse. In: Orlov, Ju.K., Boroda, M.G., Nadarejšvili, I.Š. (eds.). *Sprache, Text, Kunst (Quantitative Analysen).* Bochum, Brockmeyer, 56-81.

Nešitoj, V.N. (1975). Dlina teksta i ob'jem slovarja (Text length and vocabulary size). In: *Metody izučenija leksiki.* Minsk, BGU, 110-118.

Orlov, Ju.K. (1982). Dynamik der Häufigkeitsstrukturen. In: Guiter, H. & Arapov, M.V. (eds.). *Studies on Zipf's law.* Bochum, Borckneyer, 116-153.

Piotrowski, R.G. (1968). *Informacionnye izmerenija jazyka (Informational measuring of language).* Leningrad, Nauka.

Sambor, J. (1970). Analiza stosunku "type-token", czyli objetości słownictwa (W) i długości tekstu (N). (An analysis of the "type-token ratio" or the relation between the size of vocabulary and the length of text). *Prace filologiczne XX, 65-70.*

Somers, H.H. (1959). *Analyse mathématique du langage: Lois générales et mesures statistiques.* Louvain, Nauwelaerts.

Tuldava, J. (1974). O nekotorych statostilističeskich charakteristikach teksta (Some stylostatistical text characteristics). In: Lotman, Ju.M. (ed.). *Materialy Vsesojuznogo simposiuma po vtoričnym modelirujusim sistemam 1.* Tartu, University Press, 248-249.

Tuldava, J. (1977). O kvantitativnych charakteristikach bogatstva leksičeskogo sostava chudožestvennych tekstov (The quantitative characteristics of lexical richness). *Acta et Commentationes Universitatis Tartuensis 437, 159-175.*

Tuldava, J. (1980). K voprosu ob analitičeskom vyraženii svjazi meždu ob'jemom slovarja i ob'jemom teksta (On the analytical expression of the relation between size of vocabulary and size of text). *Acta et Commentationes Universitatis Tartuensis 549, 113-144.*

Williams, C.B. (1970). *Style and vocabulary: numerical studies.* London, Griffin.

Zacharova, A.V. (1967). Opyt statističeskogo issledovanija ustnoj reči rebenka (On statistical investigation of oral speech of children). In: *Issledovanije po jazyku i folkloru 2.* Novosibirsk, NGU, 16-38.

9

The Ratio of Word Forms and Lexemes in Texts

An important quantitative-typological indicator of language is the size interrelation *L/V* (or *V/L*), i.e. the ratio of the number of lexemes to that of word forms (or vice versa). To a certain extent this interrelation characterizes the morphological structure of a language from the quantitative point of view, making it possible to estimate the degree of its analyticism/ syntheticism. The problem is also of practical importance for automatic text processing. In this article we shall deal with the interrelations of *L* (number of lexemes), *V* (number of word forms), and *N* (text length) in their statics and dynamics on the basis of illustrative examples from various languages.[1]

1. Analyticism/syntheticism of languages

The higher the ratio *L/V* (i.e. the quotient of the division of the number of different lexemes by the number of different word forms in a given text), the more analytical is the language of the text, since in this case the number of different lexemes approaches the number of different word forms and, consequently, the average number of grammatical forms per word (lexeme) in the text is smaller. And on the contrary, a higher ratio *V/L* shows that in the given text the average number of word-changing forms per word is larger and the language of such a text should be regarded as more synthetic. At the same time it should be taken into account that the quantitative degree of analyticism/syntheticism is susceptible to changes in the length (size) of the text (*N*). Experience has shown that as the text is lengthened up to a certain limit, the ratio *L/V* decreases steadily, while the ratio *V/L* increases correspondingly. This tendency can best be illustrated on the basis of one and the same text (see Table 1). However, if the text is made very long, the influence of a certain regularity - the appearance of rare words - will make itself ever increasingly felt; passing over to text

[1] Here as well as later on the term 'lexeme' designates word as a whole (belonging to a certain part of speech), including all its meanings and grammatical forms, whereas a 'word form' is but one of the grammatical forms of a word. For instance, the lexeme MAN includes the word forms *man, man's, men, men's* and the meanings 'human being', 'male person', male servant', etc.

samples of large size we notice that practically all the words that make their first appearance are rare ones. This reduces the rate of the decrease in the ratio *L/V* (and the corresponding increase in the ratio *V/L*).

Table 1

Dynamics of the changes in the degree of analyticism/syntheticism on the basis of the author's text in the novel "Truth and Justice" by A.H. Tammsaare (in Estonian)

N	V	L	L/V	V/L
10,000	3636	2114	0.58	1.72
20,000	5944	3124	0.53	1.90
30,000	7503	3781	0.50	1.98
114,124 (the whole novel)	16750	7348	0.44	2.28

Comparison of texts in different languages (keeping the length of the samples equal) has shown that the ratio *L/V* (or *V/L*) clearly distinguishes more analytic languages from less analytic ones (Table 2). We can also see that the degree of analyticism/syntheticism does not depend only on the length of the text and the language in question, but to a certain extent also on its functional style (genre). For instance, the coefficient of analyticism (the ratio *L/V*) of an English text of $N = 200{,}000$ is 0.53 in the case of a newspaper text, whereas in that of a scientific-technical text it is 0.67. This allows us to conclude that the main influence on the quantitative degree of analyticism is exerted by the size of the vocabulary: in a newspaper text the size of the vocabulary is $L \approx 12{,}000$ (lexemes), whereas in a technical text of equal length the number of different lexemes in only about 7,000. It is clear that if there is such a difference between the sizes of the vocabularies, the ratio of the number of lexemes to that of word forms (*L/V*) is bound to be different as well, while its dependence on the length of the text is but indirect, i.e. only so far as the size of the vocabulary depends on the length of the given text.

2. Dynamic aspect

In view of the fact that the quantitative degree of analyticism/syntheticism is primarily determined by the size of the vocabulary of a given text, it is expedient

to bring out the analytical interdependence of the number of word forms (*V*) and the number of lexemes (*L*) whenever one of these values changes. This question is not only of theoretical but also of practical importance. Particularly, the analysis of dictionaries compiled on the basis of texts in synthetic languages may require the prognostication of the size of the vocabulary of lexemes if the size of the vocabulary of word forms is known, or vice versa.

Table 2

Degrees of analyticism/syntheticism in different languages

Language	N	V	L	L/V	V/L
Kazakh (children's literature)	98,040	23,350	10,076	0.43	2.32
Estonian (prose fiction)	99,898	30,773	14,654	0.48	2.10
Ukrainian (prose fiction)	100,000	27,570	13,954	0.51	1.98
Russian (epistolary language)	96,800	15,842	8,064	0.51	1.96
French (electronics)	100,000	8,108	4,572	0.56	1.77
English (electronics)	100,000	7,853	5,197	0.66	1.51
Russian (electronics)	200,000	21,648	6,816	0.32	3.18
Rumanian (electronics)	200,000	14,292	5,708	0.40	2.50
German (medicine)	200,000	41,041	20,367	0.50	2.02
French (prose fiction)	200,000	20,531	10,868	0.53	1.89
English (newspapers)	200,000	23,595	12,588	0.53	1.87
English (scientific-technical)	200,000	10,582	7,160	0.67	1.48

Note: The initial data were drawn from the following works: Alekseev 1975; Bektaev 1978; Grigorjeva 1981; Frequency Dictionary of contemporary Ukrainian prose fiction 1969; Jablonskaja 1976; Engwall 1974.

We proceed from the assumption that within certain limits (excluding very long texts) in the process of the generation of speech (text) there is constant interdependence between the rate of the relative growth of the vocabulary of lexemes and the rate of the relative growth of the vocabulary of word forms. This is quite a realistic assumption, being in accordance with the law of allometric growth, on the one hand, and with Menzerath-Altmann's law, on the other hand (see the article "On causal relations in language" in this issue). Mathematically the law may be expressed in the form of a differential equation

(1) $$\frac{dy/y}{dx/x} = b.$$

The formula can be rewritten as $dy/y = b\ (dx/x)$ and, integrating it, we get $\log y = A + b \log x$, which shows a linear relationship between $\log y$ and $\log x$. Taking $A = \log a$, we get the power function (allometric function of growth)

(2) $$y = ax^b$$

where a and b are parameters.

Let us take a look at the dependence of the number of lexemes on the number of word forms according to formula (2), taking $y = L$ and $x = V$, i.e.

(3) $$L = aV^b$$

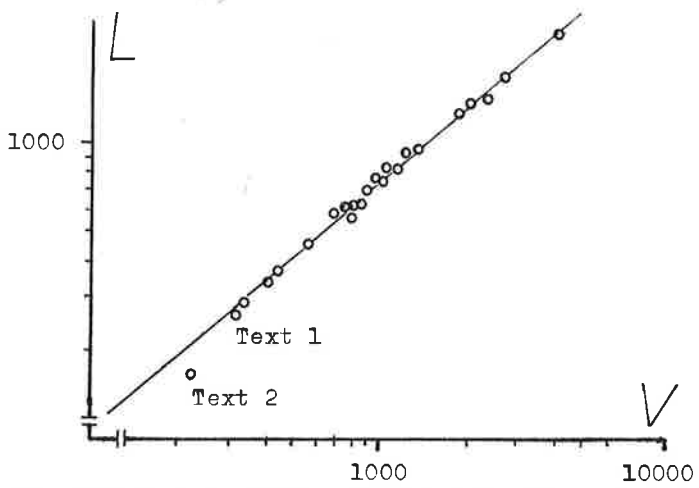

Figure 1. The relation of the number of word forms (V) to the number of lexemes (L) on the basis of 24 Russian texts of prose fiction (cf. Table 3) Bilogarithmic Scale.

on the basis of 24 texts of Russian prose fiction (Table 3). In this case the length of the text (N) ranges from 453 to 13,401, and the vocabulary of the word forms (V) from 225 to 4157. A check has shown that formula (3) yields a good fit between the empirical and the theoretical data when $a = 1.9355 \approx 1.94$ and $b = 0.8578 \approx 0.86$ (calculated using the method of least squares). The

adequacy of the formula is also confirmed by the diagram (see Fig. 1), where on the bilogarithmic scale nearly all points (denoting texts) lie on a straight line. The only exception is Text No. 2, which has the smallest vocabulary ($V = 225$ and $L = 163$) and differs from the rest of the texts also in that it is a children's story ("The three bears" by L. Tolstoj). We can conclude that the law of constant relative growth (allometric growth law) in the form of a power function is not revealed from the very beginning of the generation of a text, but approximately from the vocabulary size $V \approx 300$ (which corresponds to the length $N \approx 500$ of a Russian text of fiction).

Table 3
Length of text (N), size of vocabulary in word forms (V) and in lexemes (L); indices of analyticism/syntheticism (L/V and V/L) on the basis of Russian texts of prose fiction

No.author, work	N	V	L	L/V	V/L
1. I. Ilf, E. Petrov "Lentjaj" (The Lazybones)	453	322	278	0.86	1.16
2. L. Tolstoj "Tri medvedja" (The three bears)	508	225	163	0.72	1.38
3. E. Kononenko "Komsomolskoje serdce" (The heart of a komsomol)	550	354	296	0.84	1.20
4. A. Čechov "Posle teatra" (After theatre)	884	481	367	0.67	1.31
5. V. Inber "Bessonnica" (Sleeplessness)	915	587	452	0.77	1.30
6. L. Panteleev "Čestnoe slovo" (Word of honour)	1019	432	330	0.76	1.31
7. A. Rozen "Vremena i ljudi" (Times and people)	1289	728	587	0.79	1.26
8. V. Lidin "Doroga "žuravlej" (The route of storks)	1324	752	585	0.78	1.29
9. K. Paustovskij "Ručji, gde pleščetsja forel'" (Streams where trout are splashing)	1363	787	610	0.78	1.29
10. I. Babel "Probuždenie"(The awakening)	1500	895	706	0.79	1.27

11. M. Gorkij "O pervoj ljubvi" (On one's first romance)	1532	932	705	0.76	1.32
12. K. Paustovskij "Briz" (The breeze)	1540	801	610	0.76	1.31
13. V. Vasilevskaja "Podrugi" (Girl-friends)	1564	781	525	0.67	1.49
14. N. Koževnikova "Naša Rita" (Our Rita)	1957	986	766	0.78	1.29
15. K. Paustovskij "Sneg" (The snow)	2209	1071	751	0.70	1.43
16. V. Lidin "Drevnjaja povest'" (An ancient story)	2265	1074	784	0.73	1.37
17. V. Korolenko "Istorija moego sovremennika" (The story of a contemporary)	2336	1168	823	0.71	1.42
18. A. Kuprin "Slon" (The elephant)	2833	1281	926	0.72	1.38
19. S. Maršak "Koškin dom" (The cat's house)	3026	1347	958	0.71	1.41
20. R. Romanov "Arbatskaja skazka" (An Arbat fairy tale)	4965	2005	1324	0.66	1.51
21. A. Čechov "Dama s sobačkoj" (The lady with a dog)	5113	2052	1369	0.67	1.50
22. M. Gorkij "Starucha Izergil'" (Old woman Izergil)	6909	2454	1462	0.60	1.68
23. Yu. Kazakov "Goluboe i zelenoe" (Blue and green)	8502	2738	1698	0.62	1.61
24. P. Romanov "Pravo na žizn'" (One's right to life)	13401	4157	2412	0.58	1.72

Note: The initial data were drawn from UDAC (1974: 14-15).

3. Comparison of languages

In all probability, formula (3) is also suitable for the expression of the interdependence of the number of word forms (V) and the number of lexemes (L) in texts of many other languages. For instance, for the Estonian language the parameters of formula (3) are $a = 2.5923 \approx 2.6$ and $b = 0.8170 \approx 0.82$ on the basis of a single (individual) text of fiction (Table 1) and $a = 2.7$ and $b = 0.83$ on the basis of diverse texts of fiction (in the case of samples ranging from

5,000 to 100,000 running words; cf. Tuldava 1975). It appears that the parameters a and b remain fairly constant within the limits of one and the same language, and especially within one and the same genre. Parameter b expresses the real interdependence between V and L, thus having structural significance; e.g. if $b = 0.82$, it means that whenever V increases by 1% , L will increase by 0.82% on average. This regularity allows us to predict the number of lexemes per a certain number of word forms (and vice versa), which can be of practical importance for the solution of some tasks involving automatized text processing.

To test the goodness-of-fit between the observed and expected data we used the determination coefficient R^2 according to the following formula:

$$R^2 = 1 - \frac{\sqrt{(y_i - \hat{y}_i)^2}}{\sqrt{(y_i - \bar{y})^2}} \quad (4)$$

where y_i are observed values, $\hat{y}_i$ - expected values, and $\bar{y}$ - mean of the observed values. On the basis of Table 1 we calculated $R^2 = 0.997$ and on the basis of Table 3 (excluding Text No. 2) $R^2 = 0.996$. So the fit is actually good in both cases.

In order to calculate the counter-dependence, i.e. the dependence of V on L, we only need to paraphrase formula (2) so that we get

$$V = \alpha L^{\beta} \quad (5)$$

where $\alpha = e^{-(\ln a)/b}$ and $\beta = 1/b$.

However, this is true only in the case of a perfect fit between the observed and the expected (computed) data. As regards the data presented in Table 3 (the Russian texts), these were calculated according to formula (5):

$$\alpha = e^{-(\ln 1.9355)/0.8578} = 0.4631 \text{ and } \beta = 1/0.8578 = 1.1658.$$

In comparison, the calculation performed by means of the method of least squares has yielded the result:

$$\alpha = 0.4846 \text{ and } \beta = 1.1585.$$

Comparison of the values of parameter b in the Russian and Estonian languages has revealed that in Russian b has a higher value (here: $b \approx 0.86$ in Russian and $b \approx 0.82$-0.83 in Estonian). The fit is very good in both cases and,

in all probability, the difference between the results is statistically significant. This assumption is supported by qualitative argumentation. In terms of language structure, the higher value of b in Russian means that Russian is more analytical than the Estonian language (with an increase in the number of different word forms the number of lexemes grows faster). This is really so (suffice it to point out that Russian has five cases against the 14 in Estonian). Thus, parameter b can be regarded as a truthful indicator of the analytical nature of a language, being of structural significance and more or less independent of the length of the text (according to available data - within the range of $N = 500$ to $N = 100{,}000$). Still, it should be remembered that exact comparison can only be made of texts belonging to one and the same genre and the conditions of the experiments should be more or less equal. It is of particular importance how the notion 'lexeme' is defined (e.g. whether a lexeme includes all the meanings of a given word or whether these are regarded as separate lexemes).

4. Relation to the length of text

Lastly, let us examine the problem of the relation between the ratio L/V (or V/L) and the length of text (N). There is a correlation between them, although, as has been pointed out above, the decisive role in the formation of the degree of analyticism/syntheticism is played by the size of the vocabulary of a given text (in the case of texts of equal length, their vocabularies can vary greatly depending on the peculiarities of the functional or the individual style). We can say that on average the L/V ratio decreases as N increases (Tables 1 and 3). On the basis of the data on 24 Russian texts (Table 3) the coefficient of linear correlation $r = -0.83$, which shows a statistically significant relation between L/V and N (the critical value being $r_{0.001;\,22} = |\,0.66\,|$). A closer analysis has revealed that the relation between L/V (or V/L) and N has the same character of dependence as was noticed between V and L, i.e. dependence based on the allometric law of growth. We write

$$V/L = cN^{-d} \tag{6}$$

and

$$L/V = \gamma N^{\delta} \tag{7}$$

where c, d, γ and δ are parameters.

According to available data on the Russian language (Table 3), the values of the parameters in formula (7) are $g = 1.52$ and $d = 0.09$. Within certain limits the formula allows us to predict the numerical value of analyticism for different text lengths.

* * *

Summing up: The quantitative analysis of the relations between word forms, lexemes and text length perfomed on the basis of materials drawn from two languages (Russian and Estonian) has revealed certain regularities in the statistical organization of texts, which can be of importance for the solution of a number of tasks involving the application of quantitative methods in linguistics as well as for theoretical exploration of statistical text laws. Of course, the problem requires further investigation on the basis of more comprehensive materials in order to verify the general principles formulated here and to specify the conditions for the application of the formulas suggested by the author.

References

Alekseev, P.M. (1975). *Statističeskaja leksikografija (Statistical lexicography).* Leningrad, LGPI.

Bektaev, K.B. (1978). *Statistiko-informacionnaja tipologija tjurkskogo teksta (Statistical-informational typology of Turk texts).* Alma-Ata, Nauka.

Engwall, G. (1974). *Fréquence et distribution du vocabulaire dans un choix de romans français.* Stockholm, Skriptor.

Frequency dictionary *of contemporary Ukrainian prose fiction* (1969). (V.I. Perebejnos - ed.). Kiev, AN USSR.

Grigorjeva, A.S. (1981). *Statističeskaja struktura russkogo epistoljarnogo teksta (Statistical structure of Russian epistolary texts).* PhD thesis. Leningrad.

Jablonskaja, N.N. (1976). Častotnyj slovar' nemeckogo pod-jazyka chirurgii (Frequency dictionary of the German sublanguage of surgery). *Voprosy prikladnoj lingvistiki 6.* Dnepropetrovsk.

Tuldava, J. (1975). *Eesti keele sõnavara statistiline struktuur (The statistical structure of the Estonian vocabulary).* Tartu, TRÜ (mimeogr.).

UDAC (1974). *Uppsala University Data Center. Processing Natural Language at UDAC.* Report No. 2. Sågvall, A.-L. et al. (eds.). Uppsala.

10

Quantitative Analysis of the Phonemic System of the Estonian Language

Some aspects of the paradigmatic relations between the phonemic units of the Estonian language are examined from the quantitative point of view. The results of the investigation are compared with analogical data from some other languages.

1. The inventory

The phoneme inventory of the Estonian language contains nine vowels /a e i o u õ ä ö ü/ and seventeen consonants /p t t´ k f h j l l´ m n n´ r s s´ š v/. All these phonemes may be short or long (or overlong).[1] In addition, there are 36 diphthongs in the Estonian language, its dialects included (Piir 1985), but they are treated phonologically as vowel sequences (Viitso 1981:67).

All nine vowels contrast in stressed position, but only four of them /i e a u/ occur in unstressed position (the stress in Estonian is normally on the first syllable).

The first element of a diphthong may be any of the nine vowels, but the second element has to be one of the following five /i e a u o/, and there are combinations which are not acceptable in the literary Estonian language.

These are some of the restrictions which may play a role in quantitative investigation of the phonemic system and its functioning in the language.

The Estonian language exhibits a nearly phonemic orthography. The accepted phonemic transcription is based on graphemes used in written text. However, some differences have to be noted. In writing, long vowels are presented as two graphemes, e.g. *maa* /mā/ 'land' (a macron "-" over a phoneme means that it is long), *puu* /pū/ 'tree'. Long consonants, as a rule, are also written with the help of two graphemes, e.g. *linn* /liñ/ 'town'. But there are certain positions where a long consonant is marked with one grapheme, e.g. *linlane* /liñlane/

[1] A peculiarity of Estonian is the complicated quantity pattern traditionally described as an opposition of three distinctive quantity degrees: short, long, and overlong (cf. Lehiste 1970). In this study we do not distinguish between long and overlong degrees.

'town-dweller'. The palatalized consonants /t' l' n' s'/ are unmarked in the written text.

All stops are unvoiced, but there is a distinction between short and long stops (*lenis* and *fortis* on the phonetic level). Short stops are represented either by the graphemes *b, d, g* or by *p, t, k,* e.g. *viga* /vika/ 'mistake' (in phonetic transcription [viGa]) and *kala* /kala/ 'fish' where the short stop *k* occurs initially. Long stops are usually marked with two graphemes (*pp, tt, kk*), but in certain positions there is only one grapheme to represent a long stop; cf. *pikk* /pik/ 'long' and *piklik* /piklik/ 'oblong'.

The phonology of a language cannot be regarded as complete if it does not take into account some basic quantitative (statistical) features of the system and the functioning of its units in speech (text). Relative phonostatistical characteristics may be of special interest. For example, the ratio of the number of vowels to the number of phonemes in the system is the expression of the quantitative typological characteristic of 'vocalism' ("Vokalhaltigkeit") distinguishing a language from other languages (cf. Altmann & Lehfeldt 1973; Strauß 1980). According to our phonemic transcription the Estonian language is characterized by the vocalism ratio of 9/26 = 0.346, i.e. the vowel phonemes make up 34.6 % of the total number of phonemes in Estonian. The ratio calculated by similar experimental conditions is 8/23 = 0.348 for the cognate Finnish language. (Finnish lacks the vowel /õ/ and the palatalized consonants but has the voiced /d/ as a separate phoneme.) By contrast, the ratio of vocalism, calculated for American English (after Roberts 1965) is 8/32 = 0.250 (cf. Altmann & Lehfeldt 1973: 80).

A factor of even greater importance in phonostatistical study is the frequency of occurrence of phonemic units in text (in speech). The vocalism rate can be, for instance, measured by the well-known Krámský coefficient (Krámský 1946-1948) calculated as the ratio of the consonant phoneme proportion in the inventory (p_i) to the consonant phoneme frequency rate in the text under study (p_t): $v = p_i/p_t$. The coefficient for Estonian texts is $p_i = 17/26 = 0.654$ (or 65.4 %), $p_t = 0.545$ (or 54.5 %), and $v = 0.654/0.545 = 1.20$.

For the sake of comparison we present the data for some other languages (cf. Table 1).

The table clearly demonstrates that, judging by the Krámský coefficient, Ukrainian is highly vocalic ($v = 1.45$). According to Krámský (1946-1948) vocalism is also high in the Romance languages, the vocalism coefficient being 1.48 for Spanish and 1.58 for Italian. The vocalism coefficient is low in the Germanic languages: German - 0.85, English - 0.91 and, as our calculations

show, Swedish - 1.05 (Table 1). The three Finno-Ugric languages studied (Estonian, Finnish and Hungarian) as well as Czech occupy a middle position on this scale of vocalism ($1.10 < v < 1.30$).

Table 1
The calculation of Krámský's vocalism coefficient

Language	p_i	p_t	$v = p_i/p_t$
Swedish	18/27 = 0.667	0.636	1.05
Hungarian	25/39 = 0.641	0.583	1.10
Czech	25/38 = 0.658	0.587	1.12
Finnish	15/23 = 0.652	0.518	1.26
Ukrainian	32/38 = 0.842	0.579	1.45

(The calculations have been made on the basis of the data presented in the papers of Veenker 1982, Těšitelová 1985, Perebejnos 1969, Sigurd 1965, 1970.)

2. Phoneme frequency distribution

Our study is based on a corpus of texts of modern Estonian prose fiction (non-conversational material only) with a total of 150,000 running phonemes. There was good homogeneity of phoneme frequencies in the subsamples. In this study we measured the frequency of occurrence of phonemes in a simplified way making no distinction between the short and the long variants of a vowel or a consonant, between the palatalized and the non-palatalized variants (e.g. no distinction was made between /a/ and /ā/, /t/ and /t'/). As a result the total number of phonemes in Estonian was reduced to 22 (9 vowels and 13 consonants).

When we compare the phonemes by their frequency of occurrence in text (Table 2) three distinct groups may be distinguished: the phonemes of high frequency (9 % and more), medium frequency (between 9 and 2 %) and of low frequency (under 2 % of occurrence in text). The division into several frequency zones can also be observed on graph, cf. Fig. 1(a).

In full accord with other linguistic levels the phoneme system in its functioning manifests two tendencies, that of concentration and that of dispersion of its units. One can distinguish the nucleus, the intermediate part and the peri-

phery. The five most frequent phonemes in Estonian texts /a t e i s/ account for 53.6 % of the text, the eight least frequent phonemes - only 7.4 %.

The phenomenon of concentration and dispersion is well-known in lexical statistics where the statistical distribution of the units may be expressed analytically by the so-called Zipf's law in the form of a *power function*. The question arises as to whether the phonemic level with its limited inventory (unlike the lexical level with a very large number of units) is also subject to Zipf's law. This problem may be approached in different ways and we shall try to examine it in more detail.

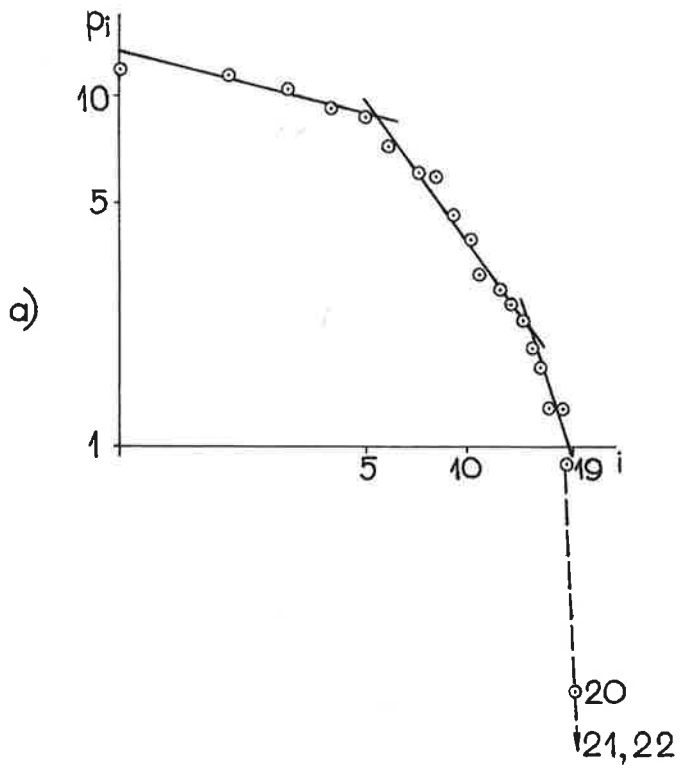

Figure 1a. Phoneme frequency p_i(%) as power function of rank i

We start with checking whether Zipf's law holds with regard to the rank distribution of phoneme frequencies under study. A graphical way of checking is as follows. Write the usual Zipf function as

$$p_i = Ai^{-b} \quad (1)$$

where p_i denotes the relative frequency of a phoneme with rank i; A and b are constants (parameters). Taking logarithms, we have

$$\ln p_i = \ln A - b \ln i$$

which means that $ln\ p_i$ should plot against $ln\ i$ as a straight line with $-b$ as its slope. As can be seen from Fig. 1(a), the empirical points (data from Table 2) do not form a straight line, but show deviations at the beginning and at the end of the rank distribution of phoneme frequencies. One would suggest that such a distribution was hybrid or composite, displaying a division into several groups (nuclear, intermediate, and peripheral, as stated above).

Table 2
Phoneme frequency groups in Estonian

I	II	III
/a/ 12.2	/k/ 7.3 /l/ 6.2	/j/ 1.9 /h/ 1.7
/t/ 11.9	/u/ 6.0 /n/ 4.6	/ä/ 1.3 /õ/ 1.3
/e/ 11.0	/m/ 4.0 /o/ 3.1	/ü/ 0.9 /ö/ 0.2
/i/ 9.5	/r/ 2.9 /p/ 2.6	/f/ 0.05 /š/ 0.05
/s/ 9.0	/v/ 2.3	
53.6 (%)	39.0 (%)	7.4 (%)

If we wanted to describe the distribution with the help of one single function, then we could use the following methods:

First, it would be possible to modify the Zipf distribution by adding Mandelbrot's correction factor (parameter B) to formula (1):

$$p_i = A(i + B)^{-b}. \tag{2}$$

In linguistics it is widely accepted that the ranked frequencies of phonemes (and letters) are distributed according to this function, called Zipf-Mandelbrot's law (cf., e.g., Zörnig & Altmann 1983, 1984; cf. also Krylov 1982). As we know, the correction factor B in formula (2) helps to straighten the distribution at its beginning where there is deviation from the straight line in the bilogarithmic coordinate system. The corresponding differential equation has the form

$$\frac{dp_i/p_i}{di/i} = \frac{d \ln p_i}{d \ln i} = \frac{b}{1 + B/i}$$

(with $b < 0$) which by integration gives formula (2). The ratio between the relative growth rates is no longer constant as it was in the case of Zipf's function (1) but will be modified by the expression $(1 + B/i)$. It means that at the beginning of the rank distribution (with small i values) the influence of the modification may be rather strong, but with increasing i values the ratio B/i will be nearing zero and the relation between the growth rates will practically become constant again.

The application to our material of Zipf-Mandelbrot's law according to formula (2) is somewhat complicated. In order to approximate the function to the observed frequencies adequately - as can be seen from Fig. 1 (a) - we have to consider both the initial and the medium parts of the distribution to be "deviations" from the straight line. Consequently, the starting point for the linearization should be the lowest segment of the distribution. Our first task would be to calculate the value of the correction factor B. This can be made graphically and by means of iterations. A simple procedure might be to calculate the B value using the following formula (cf. Haitun 1983: 162):

$$B = \frac{n - 1}{(p_1/p_n)^a - 1} - 1 \tag{3}$$

where n denotes the size of the inventory (here: the number of phonemes in the system; $n = 22$), p_1 - the frequency of the phoneme with rank $i = 1$, p_n - the frequency of the last phoneme in the rank distribution. Parameter $a = 1/b$ where

b represents the slope in the bilogarithmic co-ordinate system. The approximate value of a can be established graphically, see Fig. 1 (a). We can calculate the value of a approximately by taking two points at the lowest end of the graph, for instance, the points with the ranks 19 and 21 and the corresponding frequencies $p_{19} = 0.9$ and $p_{21} = 0.05$. The calculation yields the result:

$$a = \frac{\ln 21 - \ln 19}{|\ln 0.05 - \ln 0.9|} \approx 0.03.$$

According to formula (3) the tentative value of B will be

$$B = \frac{22 - 1}{(12.2/0.05)^{0.03} - 1} - 1 = 116.$$

By means of the method of least squares we then calculate the values of the parameters A and b on the linearized model

$$\ln p_i = \ln A - b \ln(i + B).$$

In the course of iterative correction of the fitting we ascertained that the approximation improved slightly with the increasing value of B. A satisfactory approximation is attained with $B = 120$ when $\ln A = 86.4078$ and $b = -17.4721$. Taking $A = e^{86.4078} = (3.36)10^{37}$, we get

$$\hat{p}_i = (3.36)10^{37}(i + 120)^{-17.47}.$$

The observed and the computed values of p_i are presented in Table 3 (2nd and 3rd columns). The deviation, calculated as the sum of squares of differences (residuals), i.e. $\sum d_i^2$, equals 8.01. The determination coefficient (R), calculated according to the formula

$$R = \frac{\sum_i d_i^2}{\sum_i (p_i - \bar{p})^2} \tag{4}$$

where $\bar{p}$ denotes the mean value of the frequencies (here: $\bar{p} = 4.545$ and $\Sigma_i(p_i - \bar{p})^2 = 333.85$), equals 0.976. The analysis of residuals is of importance. We see

(Table 3) that the standardized residuals (d_i/s_d) do not exceed 2.0, except for p_1 (the highest frequency in the rank order of frequencies). (It is generally accepted that a significant difference is indicated by a standardized residual with a magnitude in excess of 2.0. We mark it with an asterisk in our tables.) Thus, on the whole, the fit may be considered quite satisfactory.

Table 3
Rank frequency distribution of phonemes: fitting to power functions

i	p_i	$\hat{p}_i$	$d_i = p_i - \hat{p}_i$	d_i/s_i	$\hat{p}_i$	$d_i = p_i - \hat{p}_i$	d_i/s_i
1	12.2	13.8	-1.6	-2.5*	11.9	+0.3	+0.5
2	11.9	11.9	0	0	13.5	-1.6	-2.8*
3	11.0	10.4	+0.6	+0.9	12.1	-1.1	-2.0*
4	9.5	9.0	+0.5	+0.8	10.2	-0.7	-1.2
5	9.0	7.8	+1.2	+1.9	8.6	+0.4	+0.7
6	7.3	6.8	+0.5	+0.8	7.2	+0.1	+0.2
7	6.2	5.9	+0.3	+0.5	6.0	+0.2	+0.4
8	6.0	5.2	+0.8	+1.2	5.1	+0.9	+1.6
9	4.6	4.5	+0.1	+0.2	4.4	+0.2	+0.4
10	4.0	4.0	0	0	3.8	+0.2	+0.4
11	3.1	3.4	-0.3	+0.5	3.2	-0.1	-0.2
12	2.9	3.0	-0.1	+0.2	2.8	+0.1	+0.2
13	2.6	2.6	0	0	2.5	+0.1	+0.2
14	2.3	2.3	0	0	2.2	+0.1	+0.2
15	1.9	2.0	-0.1	-0.2	1.9	0	0
16	1.7	1.8	-0.1	-0.2	1.7	0	0
17	1.3	1.6	-0.3	-0.5	1.5	-0.1	-0.4
18	1.3	1.4	-0.1	-0.2	1.3	0	0
19	0.9	1.2	-0.3	-0.5	1.2	-0.3	-0.6
20	0.2	1.1	-0.9	-1.4	1.1	-0.9	-1.6
21	0.05	0.9	-0.85	-1.3	1.0	-0.95	-1.7
22	0.05	0.8	-0.75	-1.2	0.9	-0.85	-1.5
Σd_i			-1.4			-2.4	
Σd^2_i			8.01			6.43	
s_d			0.6447			0.5693	
Formulas		$\hat{p}_i = (3.36)10^{37}(i + 120)^{-17.47}$			$\hat{p}_i = 11.88i^{0.4843 - 0.4277 \ln i}$		

The only disadvantage of using Zipf-Mandelbrot's formula for the approximation of phoneme frequency distributions seems to be the circumstance that by taking the lowest frequencies as the starting-point, the values of the parameters (especially A and B) will become disproportionately large in comparison with the measured values of the frequencies. In our case, another discord is the slight imbalance revealed by the sum of residuals ($\Sigma d_i = -1.4$, instead of expected 0) and the concentration of plus signs at the one end and minus signs at the other. But, as we have seen, the deviations are negligibly small and may be discarded as insignificant.

Another way of looking at the question is the conception of "non-linear interpretation of Zipf's law" (cf. Piotrowski 1975; Alekseev 1978). It is stated that in the case of limited inventory (such as a phoneme system) the rank distribution of the entities forms a curve rather than a straight line in the coordinate system with logarithmic axes.

Alekseev proposes a modification of Zipf's law in the form of

$$p_i = Ai^{b - c \ln i} \tag{5}$$

where A, b and c are parameters. By taking logarithms we get

$$\ln p_i = \ln A + b \ln i - c \ln^2 i,$$

i.e., practically, the equation of a parabola of the type

$$Y = a + bX - cX^2.$$

Indeed, the graph in Fig. 1 (a) resembles a curve and we can try to approximate the rank distribution of phoneme frequencies to the parabola. Calculating the values of the parameters by means of the method of least squares (for the method see, e.g., Förster & Rönz 1979, chap. 5.1), we get $ln\ A = 2.4751$, $b = 0.4843$, and $c = -0.4277$. Taking $A = e^{2.4751} = 11.8829 \approx 11.9$, we get the result:

$$\hat{p}_i = 11.9i^{0.4843 - 0.4277 \ln i}.$$

The fitting of the modified Zipf distribution (5) to the empirical data is presented in Table 3 (6th column). The results of this table indicate that, formally, the fitting is not bad ($\Sigma d_i^2 = 6.43$, i.e. slightly better than in the case of Zipf-

Mandelbrot's law) and the standardized residuals do not exceed 2.0 except for the second frequency (p_2). But there is a serious shortcoming with regard to the process of approximation: the computed values for p_2 and p_3 (frequencies for the ranks $i = 2$ and $i = 3$) exceed in magnitude the value of p_1 (the highest frequency), which contradicts the postulate of frequencies ranked in descending order. Analogous results have been received when investigating phoneme or letter distributions in other languages (cf. Alekseev 1978: 64, Fig. 5). We have to conclude that the "non-linear approach" to Zipf's law according to formula (5) cannot be considered a correctly specified model for the rank frequency distribution of phonemes.

A more promising undertaking is to try to approximate the rank frequency distribution of phonemes to the exponential law (cf., e.g. Herdan 1966; Sigurd 1968; Tuldava 1975)[2]. We shall use the formula

$$p_i = Ae^{-bi} \tag{6}$$

where A and b are constants, and e the basis for natural logarithms.

The exponential law is said to be closely connected with the theory of random processes and the formula for phoneme distribution is derived under the assumption of the occurrence of phonemes being a "pure chance event" (cf. Herdan 1966: 130). It means (in our case) monotonic decreasing of frequencies with constant decreasing rate, in differential form:

$$\frac{dp_i/di}{p_i} = -b.$$

A model of this kind can be expected "in case of phonemes as in case of the so-called grammatical words, which are usually considered less dependent on the contents of text, and, consequently, on the individual choice of lexical units" (Králík 1976: 227).

In linearized form (by taking logarithms) the exponential formula (6) can be expressed as follows:

[2] Proceeding from some theoretical assumptions, Yu. Orlov (1976) derives the formula $p_i = p_1 e^{-(i-1)(p_1 - p_n)}$ where p_1 denotes the frequency of rank $i = 1$ and p_n - the lowest frequency in the system with n symbols (letters or phonemes). This formula gives a rather rough approximation to the empirical data.

$$\ln p_i = \ln A - bi.$$

Consequently, we expect a linear relation between rank (i) and logarithm of frequency ($ln\ p_i$). Calculating the parameters by means of the method of least squares, we get

$$\ln p_i = 2.7473 - 0.1348i$$

and according to formula (6):

$$\hat{p} = 15.6e^{-0.1348i}.$$

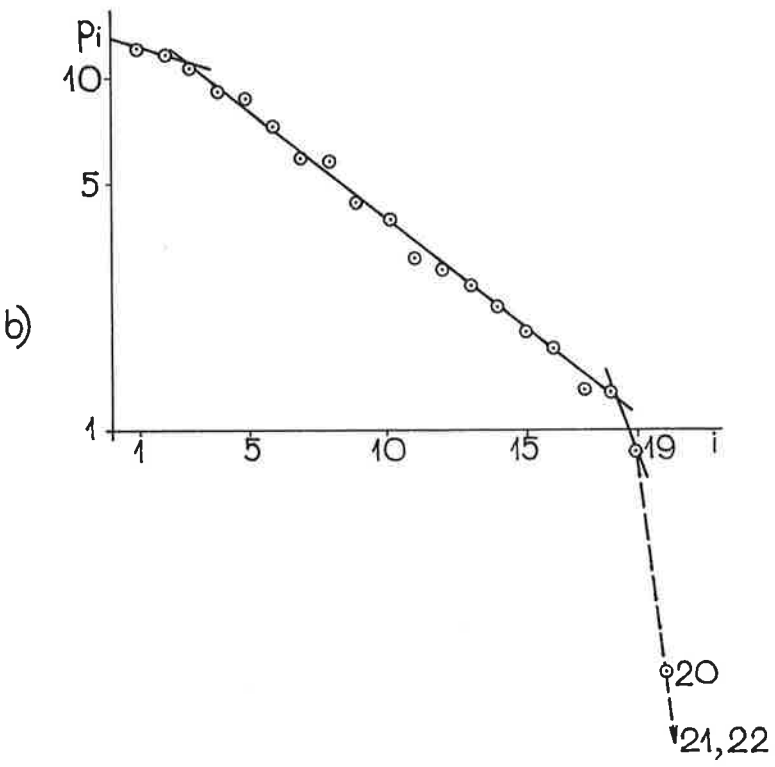

Figure 1b. Phoneme frequency p_i(%) as exponential function of rank i.

There is satisfactory agreement between the observed and the computed values (see Table 4, 2nd and 3rd columns), and when our data are plotted on semi-logarithmic graph paper (with logarithmic y-axis), see Fig. 1(b), they, on the whole, conform with the linearized equation which gives a straight line of negative slope.

However, as can be seen from Fig. 1 (b), there are still deviations from the linear relation at the beginning and at the end of the distribution. The frequency distribution of phonemes is not an event of "pure chance", as was assumed when using the exponential formula in its simplest form. We have already mentioned some restrictions in the usage of Estonian phonemes (restrictions in the use of certain phonemes in stressed and unstressed syllables and in the formation of phoneme sequences - diphthongs, see Section 1). There may be more dependences in the usage of phonemes. So we are forced to modify the initial formula (6) in order to get better results, or to leave the exponential law and turn to another law.

One possible modification of the exponential law is analogous to Alekseev's modification of Zipf's law, i.e. considering the graph on Fig. 1 (b) to be a parabola. This would give us the equation (on semi-logarithmic graph paper):

$$\ln p_i = \ln A - bi - ci^2$$

and, consequently,

$$p_i = Ae^{-bi - ci^2} \tag{7}$$

where A, b and c are parameters.

The results of Table 4 (6th column) indicate that the fit is quite good with the residual sum of squares only 3.62, which is the best result up to now. But concrete analysis of residuals shows considerable imbalance in their distribution ($\Sigma d_i = -5.5$), so the hypothesis of "parabolic-exponential" distribution of phoneme frequencies is not fully corroborated on our material.

Finally, we shall try to approximate the frequency distribution of phonemes with the help of a logarithmic function which is closely related to the exponential function, being its "opposite". In this case we can use semi-logarithmic graph paper with logarithmic scale on the horizontal (x-) axis. Figure 1 (c) shows the relation between $ln\ i$ and p_i which seems to be perfect at the end of the distribution while there is deviation only at the beginning of the distribution.

Table 4
Rank frequency distribution of phonemes: fitting to exponential functions

i	p_i	$\hat{p}_i$	$d_i = p_i - \hat{p}_i$	d_i/S_d	$\hat{p}_i$	$d_i = p_i - \hat{p}_i$	d_i/S_d
1	12.2	13.6	-1.4	-2.4	13.0	-0.8	-2.3*
2	11.9	11.9	0	0	11.9	0	0
3	11.0	10.4	+0.6	+1.0	10.8	+0.2	+0.6
4	9.5	9.1	+0.4	+0.7	9.7	-0.2	-0.6
5	9.0	8.0	+1.0	+1.9	8.7	+0.3	+0.9
6	7.3	6.9	+0.4	+0.7	7.7	-0.4	-1.2
7	6.2	6.1	+0.1	+0.2	6.8	-0.6	-1.8
8	6.0	5.3	+0.7	+1.2	6.0	0	0
9	4.6	4.6	0	0	5.2	-0.6	-1.8
10	4.0	4.1	-0.1	-0.2	4.5	-0.5	-1.5
11	3.1	3.5	-0.4	-0.7	3.9	-0.8	-2.3*
12	2.9	3.1	-0.2	-0.3	3.3	-0.4	-1.2
13	2.6	2.7	-0.1	-0.2	2.8	-0.2	-0.6
14	2.3	2.4	-0.1	-0.2	2.4	-0.1	-0.3
15	1.9	2.1	-0.2	-0.3	2.0	-0.1	-0.3
16	1.7	1.8	-0.1	-0.2	1.6	+0.1	+0.3
17	1.3	1.6	-0.3	-0.5	1.4	-0.1	-0.3
18	1.3	1.4	-0.1	-0.2	1.1	+0.2	+0.6
19	0.9	1.2	-0.3	-0.5	0.9	0	0
20	0.2	1.1	-0.9	-1.4	0.7	-0.5	-1.5
21	0.05	0.9	-0.85	-1.5	0.6	-0.55	-1.6
22	0.05	0.8	-0.75	-1.3	0.5	-0.45	-1.3
Σd_i			-2.4			-5.5	
Σd_i^2			6.75			3.62	
s_d			0.5834			0.3429	
Formulas		$\hat{p}_i = 15.6\ e^{-0.1348\ i}$			$\hat{p}_i = 14.5e^{-0.07976i - 0.003434i^2}$		

The formula for the logarithmic function is

$$p_i = A - b \ln i \tag{8}$$

where A and b are parameters. Calculation yields

$$\hat{p}_i = 14.9 - 4.7315 \ln i$$

and the fitting seems to be good except for the first frequency p_1 (see Table 5). The fact that there is good fitting at the end of the distribution (in contrast with all other distributions examined before) and deviation only at the beginning of the distribution suggests the modification of formula (8) with the help of an adjustment factor analogous to Mandelbrot's correction factor of Zipf's law. We write

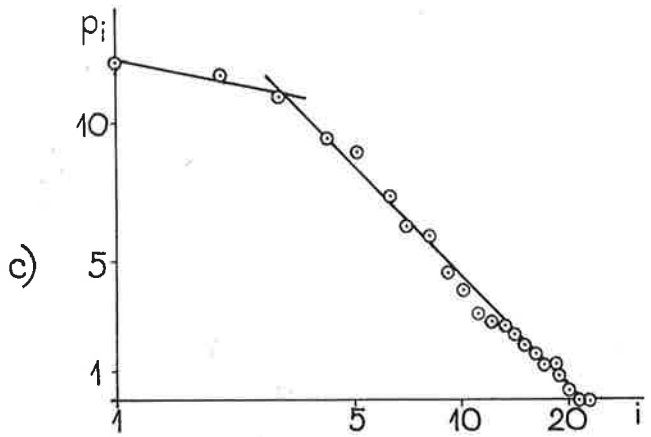

Figure 1c. Phoneme frequency p_i (%) as logarithmic function of rank i

(9) $$p_i = A - b \ln (i + B)$$

where B is the adjustment factor. Iterative calculation by means of the method of least squares yields the result:

$$\hat{p}_i = 25.14 - 7.8041 \ln(i + 4).$$

As can be seen from Table 5 the fit is actually very good with $\Sigma d_i^2 = 5.27$ and Σd_i near to zero. None of the standardized residuals exceeds the critical value of 2.0.

Thus, the analysis has shown that, on our material, all the functions examined (power, exponential, and logarithmic functions) can be used to approximate

the rank frequency distribution of phonemes with satisfactory accuracy if correction (adjustment) factors are added to the simple forms of the functions. From the formal point of view, the logarithmic model is to be preferred, taking into account the fact that in this case the deviation from the theoretical line occurs only at one end of the distribution (in the region of the highest frequencies) which can be easily corrected by adding an adjustment factor analogous to Mandelbrot's parameter B.

Table 5
Rank frequency distribution of phonemes: fitting to logarithmic functions

i	p_i	$\hat{p}_i$	$d_i = p_i - \hat{p}_i$	d_d/S_d	$\hat{p}_i$	$d_i = p_i - \hat{p}_i$	d_d/S_d
1	12.2	14.9	-2.7	-2.9*	12.6	-0.4	-0.8
2	11.9	11.6	+0.3	+0.3	11.2	+0.7	+1.3
3	11.0	9.7	+1.3	+1.4	10.0	+0.1	+1.9
4	9.5	8.3	+1.2	+1.3	8.9	+0.6	+1.1
5	9.0	7.3	+1.7	+1.8	8.0	+1.0	+1.9
6	7.3	6.4	+0.9	+1.0	7.2	+0.1	+0.2
7	6.2	5.7	+0.5	+0.5	6.4	-0.2	-0.4
8	6.0	5.1	+0.9	+1.0	5.8	+0.2	+0.4
9	4.6	4.5	+0.1	+0.1	5.1	-0.5	-1.0
10	4.0	4.0	0	0	4.5	-0.5	-1.0
11	3.1	3.6	-0.5	-0.5	4.0	-0.9	-1.7
12	2.9	3.1	-0.2	-0.2	3.5	-0.6	-1.1
13	2.6	2.8	-0.2	-0.2	3.0	-0.4	-0.8
14	2.3	2.4	-0.1	-0.1	2.6	-0.3	-0.6
15	1.9	2.1	-0.2	-0.2	2.2	-0.3	-0.6
16	1.7	1.8	-0.1	-0.1	1.8	-0.1	-0.2
17	1.3	1.5	-0.2	-0.2	1.4	-0.1	-0.2
18	1.3	1.2	+0.1	+0.1	1.0	+0.3	+0.6
19	0.9	1.0	-0.1	-0.1	0.7	+0.2	+0.4
20	0.2	0.7	-0.5	-0.5	0.3	-0.1	-0.2
21	0.05	0.5	-0.45	-0.49	0.02	+0.03	+0.06
22	0.05	0.3	-0.25	-0.27	0	+0.05	+0.09
Σd_i			+1.50			-0.22	
Σd^2_i			16.25			5.27	
S_d			0.9206			0.5261	
Formulas		$\hat{p}_i = 14.9 - 4.7315\ ln\ i$			$\hat{p}_i = 25.14 - 7.8041\ ln(i + 4)$		

The analysis of the residuals also favours the logarithmic model. The logarithmic model can be submitted to content interpretation according to which phoneme frequency (p_i) decreases proportionately with the logarithm of rank (i) in accordance with the law of "adapted restraint" (Nalimov & Mulčenko 1969: 41-42), characterizing the functioning of systems with a limited number of elements (objects).[3]

3. Comparative analysis

The comparative analysis of the distribution of certain phonemes in texts of different languages reveals both differences and resemblances (see Table 6).[4] In doing this one should remember that no similar-sounding phonemes in even closely related languages can be considered identical. There are eight phonemes of relatively identical articulatory and acoustic qualities among the ten most frequent phonemes of Finnish and Estonian. They are /a e i t s k l n/. And yet the phoneme arrangement in the rank presentation is considerably different. The relatively high frequency rate of /n/ in Finnish can be easily noted. The /n/ frequency values for Finnish and Estonian are 8.9 and 4.6 % respectively. The difference can be accounted for by the frequent final /n/ phoneme in Finnish words. The Estonian language in its evolution has lost the /n/ ending, present in the related languages, for example Finnish *jalan* - genitive case and *jalkaan* - illative case of the word *jalka* - 'leg' in comparison with the corresponding Estonian word forms *jala* and *jalga* (from the initial form *jalg*). The high frequency of the /i/ vowel in Finnish can be partly accounted for by the high frequency of diphthongs with the /i/ element, such as *ei, oi, äi, ui, yi, ie,* as the Finnish and Estonian diphthongs are treated as vowel combinations in this

[3] Investigating the rank frequency distribution of letters in small samples, A.Mackay (1965) proposes the logarithmic function in the form: $p_i/p_1 = 1 - (\ln i)/(\ln (n+1))$, where p_1 is the frequency with rank $i = 1$ and n the number of symbols in the system, demonstrating good fit on the material of Japanese and English texts and the so-called Phaistos Disc inscription.

[4] To unify the experimental conditions for the data in Table 6, the frequencies are concerned with the canonical phoneme forms, i.e. in the present case the forms include frequencies of both short and long forms, fortis and lenis forms summed up (a + ā, t + t′, etc.).

paper.[5]

Finnish and Estonian are similar in the high concentration of the most frequent phonemes. So the ten most frequent phonemes cover 81.7 % of Estonian text and 77.5 % of Finnish texts (see Table 6). High concentration of the most frequent phonemes could be noted in Spanish (80.4 % of coverage), whereas in the other languages under study the coverage by the ten most frequent phonemes ranged between 60 and 70 %. It is worth mentioning that in all ten languages the most frequent phoneme was a vowel (usually /a/, /e/ or /i/).

Another way to compare the functioning of phoneme systems in various languages is to measure the entropy of the systems. The well-known Shannon formula is used to measure the entropy of a system:

Table 6
Ten most frequent phonemes in ten different languages
(language, phonemes, percentages of frequency, coverage)

Estonian

a	t	e	i	s	k	l	u	n	m	Σ^{10}(%)
12.3	11.9	11.0	9.5	9.0	7.3	6.2	6.0	4.6	4.0	81.7

Finnish (Veenker 1982: 346, Table 53, Variant B)

i	a	e	n	t	s	k	l	o	ä	
11.7	10.9	9.5	8.9	8.0	7.4	5.8	5.7	4.9	4.7	77.5

Hungarian (Veenker 1982: 313)

e(ɛ)	t	a(ǎ)	k	l	n	é(e)	r	m	o	
12.8	10.0	9.4	6.7	6.2	5.0	4.7	4.3	3.9	3.0	66.0

Udmurt (Veenker 1981: 203)

į	o	e	i	s	n	k	a	u	z	
10.5	8.2	6.5	6.4	5.6	5.5	5.1	5.1	5.0	4.6	62.5

Lithuanian (Svecevičius 1966)

a	i	s	k	o	t	u	n	r	m	
10.9	9.3	8.0	6.9	6.7	6.6	5.5	5.3	3.9	3.4	66.5

Russian (Andreev 1967: 227)

a	e	i	t	n	r	s	v	j	m	
17.2	9.0	7.3	6.2	6.2	5.1	5.0	4.6	4.2	3.2	68.0

[5] If the Finnish diphthongs are considered independent phonemes, the rank order of the ten most frequent Finnish phonemes will be as follows: a i t n e s l k ä o (Veenker 1982: 322).

Ukrainian (Perebejnos 1969)

a	o	n	i	i	e	t	v	r	u	
10.4	9.4	6.4	6.3	6.1	5.5	5.3	5.3	4.7	4.4	63.8

Czech (Těšitelová 1985)

e	i	a	o	t	s	n	l	k	v	
11.0	10.1	9.1	6.7	4.8	4.7	4.6	4.3	4.1	3.8	63.2

Polish (Segal 1972: 193)

e	a	o	i	n	t	u	r	m	p	
10.7	9.6	8.8	8.4	6.7	4.4	3.4	3.3	3.1	3.0	61.4

Spanish (Guirao & García Jurado: 1987)

e	a	o	s	n	i	r	t	k	d	
15.0	13.3	10.8	9.4	7.1	6.6	5.4	4.5	4.3	4.0	80.4

(10) $$H = -\sum_{i=1}^{k} p_i \log_2 p_i$$

where H is entropy, p_i - probability (relative frequency), k - number of units, log_2 - logarithm to the base 2.

The entropy of the simplified system of Estonian phonemes ($k = 22$) is $H =$ 3.9063. In terms of information theory the entropy value means that on the average there is 3.9063 bits of information per phoneme in Estonian texts. In its formal sense entropy is a measure of equal distribution of probability levels; the higher the H value, the more equally are distributed the phonemes in the text. A text will have maximal entropy if the frequencies of all the phonemes in the text are equal. Maximal entropy (H_0) is calculated by the formula

(11) $$H_0 = \log_2 k$$

where k is the number of phonemes. The maximal entropy value is needed for calculating the relative entropy value which measures the phoneme distribution in contrastive studies of languages with different numbers of phonemes. The relative entropy value (H_{rel}) is calculated as a ratio of the existing entropy (H) to the maximal possible entropy by the given number of phonemes:

(12) $$H_{rel} = \frac{H}{H_0}.$$

We shall illustrate the case with the relative entropy calculations of the Estonian language in comparison with other languages (partly on the data published by Zörnig & Altmann 1984; see Table 7).[6]

Table 7
Entropy of the phonemic systems of different languages

Language	k	H	H_0	H_{rel}
Estonian	22	3.9063	4.4594	0.8760
Finnish	23	3.9584	4.5235	0.8751
Hungarian	39	4.5618	5.2854	0.9631
Russian	41	4.8257	5.3575	0.9007
Ukrainian	38	4.6293	5.2479	0.8821
Czech	35	4.7006	5.1292	0.9164
French	35	4.1018	5.1292	0.8000
Italian	31	4.2512	4.9541	0.8581
English	39	4.7098	5.2853	0.8911
Swedish	45	4.8406	5.4918	0.8814

The table demonstrates that cognate languages have similar phoneme distribution and relative entropy values. So with the Finno-Ugric languages the values of relative entropy are within the range 0.86 - 0.88, for the Slavic languages it is 0.88 - 0.92, for the Romance languages - 0.80 - 0.86, and for the Germanic languages - 0.88 - 0.89. These data should still be considered tentative, as the source material was taken from studies with different approaches to the composition of the phonemic system. More reliable results are to be found in contrastive studies of different genres of a language, where the experimental conditions have been unified; see for example the paper by T. Zsilka (1974).

4. Phoneme classes

Statistical studies of phoneme classes, in particular the frequency of occurrence of certain phoneme classes in text (in speech) play an important role in the typological study of the phonological systems of languages. Coefficients of relationships between vowels and consonants in the system are supplemented by

[6] For an attempt to calculate the theoretical entropy value as a function of the number of phonemes in a system, see Zörnig & Altmann 1984; cf. Zörnig & Rothe 1992.

textual indices in phonostatistical study. Besides the Krámský coefficient (see section 1) the so-called consonant coefficient is usually calculated (cf., e.g., Tambovcev 1986). The coefficient is expressed as a ratio of the consonant frequency to the vowel frequency in the text (C:V). In Estonian texts the coefficient is 54.5:45.5 = 1.20, or in other words the consonants are 20 % more frequent than the vowels.

Vowel phonemes can be divided into subgroups according to the articulation site (horizontal) and the degree of height to which the tongue is raised (vertical). Table 8 presents the frequencies of the subgroups in Estonian text (unr - unrounded, ro - rounded).[7]

As can be seen from the Table, the Estonian front and back vowels are distributed equally in text (the front vowels constituting 50.3 % of all the vowels). In Finnish and Hungarian the front vowels clearly predominate (58 and 56 % respectively). The phenomenon is most probably caused by vowel harmony in those languages (there is no vowel harmony in Estonian). The percentage of front vowels is also rather high in Italian and Spanish (48 and 45 % respectively), while in Russian text the front vowels make up only 39 %, in Udmurt text 30 % and in Sanskrit only about 20 %. This suggests that the distribution of the frequencies of the vowel subgroups in different languages could be of typologically distinguishing value.

More detailed analysis of frequency characteristics of the vowel system in different languages reveals other typological peculiarities. So by comparing three Finno-Ugric languages, one can see a significant difference in the distribution of vowels by vertical articulation between Estonian and Finnish on the one hand and Hungarian on the other (see Table 9). The very high percentage of low vowels in Hungarian (59.6 %), due to the high frequency of occurrence of the vowels /ɛ/ and /å/, differentiates it clearly from Estonian (29.7 %) and Finnish (32.2 %).

[7] The vowel /õ/ (in Finno-Ugric phonetic transcription [ɐ]) is a high-back (or high-central) vowel which sounds very like the first part of the English diphthong /ou/ as in ´loan´ (however, the English sound differs from the Estonian one by a lower tongue position; cf. Kostabi 1993). Estonian /ä/ is like English /æ/ in ´hand´. The remaining vowels sound like the corresponding vowels in German.

Table 8
The vowel system: frequencies in text

Articulation site → / Rise of the tongue ↓	Front		Back		Total (%)
	unr	ro	unr	ro	
High	i 20.8	ü 2.0	õ 2.9	u 13.2	38.9
Mid	e 24.2	ö 0.4	- -	o 6.8	31.4
Low	ä 2.9	- -	a 26.8	- -	29.7
Total (%)	47.9	2.4	29.7	20.0	100.0

The ratio of long vowels to short vowels (diphthongs considered vowel clusters) in Estonian text is 8:92, in Finnish text 11:89 and in Hungarian text 26:74 (%). It is obvious that the low vowel percentage and correspondingly high consonant percentage in Hungarian (see Table 1) is compensated to a great degree by the high percentage of long vowels. High percentage of long vowels was found to characterize Czech (ratio 22:78) and Latvian text (ratio 27:73).

Table 9
Frequencies of vowel subgroups in Estonian, Finnish and Hungarian text

Vowels	Estonian	Finnish	Hungarian
High	38.9	37.0	13.0
Mid	31.4	30.8	27.4
Low	29.7	32.2	59.6
Rounded	22.4	23.6	45.5
Unrounded	77.6	76.4	54.5

Table 10 presents the classification of Estonian consonant phonemes according to their articulatory manner and place (pal. - palatalization, non-pal. - non-palatalization).

The alveodentals with the exception of /r/ fall into two subgroups: palatalized and non-palatalized consonants. It has been ascertained that except in the case of automatic palatalization before /i/ and /j/ the palatalized consonants /t´ s´ n´ l´/ are not frequent in Estonian text as they cover only 0.15 % of all the phoneme usage (Hint 1978: 113). There are no voiced stop consonants in the Estonian language.

Table 10
The consonant system: frequencies in Estonian text (%)

Place	Labial	Alveodental		Palatal	Velar	Total
		Non-pal.	Pal			
Stops	p	t	t'	-	k	
	4.8	21.8		-	13.4	40.0
Fricatives	f	s	s'	š	h	
	0.1	16.5		0.1	3.1	19.8
Nasals	m	n	n'	-	-	
	7.3	8.5		-	-	15.8
Laterals	-	l	l'	-	-	
	-	11.4		-	-	11.4
Trills	-	r		-	-	
	-	5.3		-	-	5.3
Semivowels	v	-		j	-	
	4.2	-		3.5	-	7.7
Total (%)	16.4	63.5		3.6	16.5	100.0

It is not easy to determine the subgroups of long and short consonants in Estonian. Preliminarily the ratio 17:83 (%) between long and short consonants in Estonian text can be suggested.

The semivowels /v/ and /j/ are considered to be fricatives by some scholars (e.g. Eek 1987) bringing the percentage of fricatives in text to 27.5 % (cf. Table 10). The corresponding distribution of consonants in Finnish text is as follows: stops - 33.3 %, fricatives - 28.1 %, nasals - 23.7 % (cf. 15.8 % in Estonian text), laterals - 11.0 %, trills - 3.9 %. The composition of the system of consonants in Hungarian text is: stops - 41.4 %, fricatives - 24.3 %, nasals - 15.2 %, laterals - 10.6 %, trills - 7.4 % and affricates - 1.1 %.

Table 11
Distribution of phoneme sub-class frequencies in Estonian text

Consonants	Vowels 45.5 Sonorants 21.9 Obstruents 32.6	Resonants
54.5 %	100 %	67.4 %

As a whole, the phoneme distribution in Estonian text can be presented in Table 11 (Sonorants = Nasals + Laterals + Trills + Semivowels and Obstruents = Stops + Fricatives).

It is possible to compare different languages as to their degree of consonant frequency and resonant frequency (resonants are the sum total of vowels and sonorants which may be considered to express some kind of "resounding" in speech).[8] Table 12 shows that the highest level of resonant frequency is in Finnish text (73.3 %). The lowest level of resonant frequency is manifested by the Uzbek language (below 60 %). The other languages under study have their resonant frequencies ranging between 60 and 70 %.

To illustrate the point we shall present the data about the Estonian, Finnish and Hungarian phoneme systems in a graphical manner (Figure 2). It is demonstrated that the distribution of phoneme classes in text greatly differs in the phoneme systems of Estonian, Finnish and Hungarian, all three being Finno-Ugric languages.

Quantitative analysis of positional and distributive relationships within the phoneme system of a language is a special field of study. Some data concerning the Estonian language can be found in the paper by J. Tuldava (1980) and a combinatory analysis of the Estonian phonemes as to their syllabic structure and position in M. Hint's paper (1988).

[8] We have used the materials published by N.D. Andreev (1967), G. Herdan (1966), V. Perebejnos (1969), M. Těšitelová (1985) and W. Veenker (1982).

Table 12
Distribution of phoneme subclass frequencies in % in texts of different languages: consonant frequencies (Obstruents + Sonorants) and resonant frequencies (Vowels + Sonorants)

Language	Vowels	Sonorants	Obstruents	Consonant Freq.	Resonant Freq.
Estonian	45.4	21.9	32.6	54.5	67.4
Finnish	48.2	25.1	26.7	51.8	73.3
Hungarian	41.7	23.4	34.9	58.3	65.1
Lithuanian	43.7	20.1	36.2	56.3	63.8
Russian	42.0	24.0	34.0	58.0	66.0
Ukrainian	42.1	26.8	31.1	57.9	68.9
Czech	41.3	24.3	34.4	58.7	65.6
Spanish	48.5	20.2	31.2	51.5	68.1
Uzbek	32.0	26.0	42.0	68.0	58.0

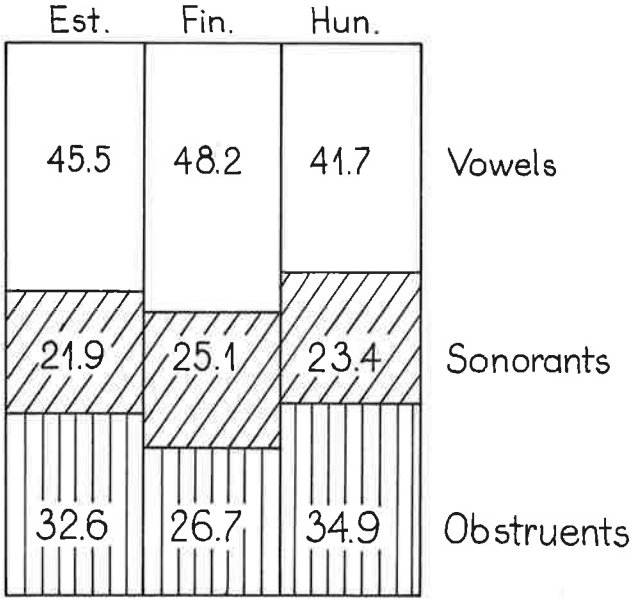

Figure 2. Distribution of phoneme sub-class frequencies in % in Estonian, Finnish and Hungarian text.

References

Alekseev, P.M. (1978). O nelinejnych formulirovkach zakona Zipfa (On non-linear formulations of Zipf's law). *Voprosy kibernetiki 41, 53-65.*

Altmann, G. & Lehfeldt, W. (1973). *Allgemeine Sprachtypologie. Prinzipien und Meßverfahren.* München, Fink.

Andreev, N.D. (1967). *Statistiko-kombinatornye metody v teoretičeskom i prikladnom jazykovedenii* (Statistical combinatory methods in theoretical and applied linguistics). Leningrad, Nauka.

Eek, A. (1987). Kvantiteet ja rõhk eesti keeles (Quantity and accent in the Estonian language). *Keel ja Kirjandus 3, 153-160.*

Förster, E. & Rönz, B. (1979). *Methoden der Korrelations- und Regressionsanalyse.* Berlin, Die Wirtschaft.

Guirao, M. & García Jurado, M.A. (1990). Frequency of occurrence of phonemes in American Spanish. *Revue québécoise de linguistique 19:2, 135-150.*

Haitun,S.D. (1983). *Naukometrija* (Scientometrics). Moscow, Nauka.

Hint, M. (1978). *Häälikutest sõnadeni* (From sounds to words). Tallinn, Valgus.

Hint, M. (1988) *Eesti ilukirjanduskeele statistiline fonotaktika* (Statistical phonotactics of the language of Estonian fiction). Tallinn, Pedagogical Institute Press.

Kostabi, L. (1993). A *contrastive analysis of English and Estonian monophthongs* (Master's thesis). Tartu (mimeogr.).

Králík, J. (1976). An application of exponential distribution law in quantitative linguistics. *Prague Studies in Mathematical Linguistics 5, 223-233.*

Lehiste, I. (1970). *Suprasegmentals.* Cambridge, MIT Press.

Mackay, A. (1965). On the type-fount of the Phaistos Disc. *Statistical Methods in Linguistics 4, 15-25.*

Nalimov, V.V. & Mulčenko, Z.M. (1969). *Naukometrija* (Scientometrics). Moscow, Nauka.

Orlov, Yu. K. (1976). Obobščennyj zakon Zipfa-Mandelbrota i častotnye struktury informacionnych edinic različnych urovnej. (The generalized Zipf-Mandelbrot law and the frequency structures of informational units on various levels). In: Andryuščenko, V.M. et al. (eds.) *Vyčislitel'naja lingvistika.* Moscow, Nauka, 179-202.

Perebejnos, V.I. (1969). IV Part in*: Sučasna ukrainska mova. Fonetika.* (Contemporary Ukrainian language. Phonetics). Kiev, Naukova dumka.

Piir, H. (1985). Acoustics of the Estonian diphthongs. In: Remmel, M. (ed.). *Estonian Papers in Phonetics 1982-1983.* Tallinn, Keele ja Kirjanduse Instituut, 5-96.

Piotrowski, R.G. (1975). *Tekst - mašina - čelovek* (Text - Computer - Man). Leningrad, Nauka. (German translation: Text - Computer - Mensch. Bochum, Brockmeyer, 1984.)

Roberts, A.H. (1965). *A statistical linguistic analysis of American English. The Hague*, Mouton.

Segal, D.M. (1972). *Osnovy fonologičeskoj statistiki* (Fundamentals of phonological statistics). Moscow, Nauka.

Sigurd, B. (1968). Rank-frequency distribution for phonemes. *Phonetica 18, 1-15.*

Sigurd, B. (1970). *Språkstruktur* (The structure of language). Stockholm, Wahlström & Widstrand.

Strauß, U. (1980). *Struktur und Leistung der Vokalsysteme.* Bochum, Brockmeyer.

Svecevičius, B.I. (1966). K voprosu o častote vstrečaemosti fonem v litovskoj pis'mennoj reči (Phoneme frequencies in the Lithuanian writing). In: *Materialy kollokviuma* (Theses of the seminar of the laboratory of experimental phonetics and psychology of speech, Part 2). Vilnius, Pedagogical Institute Press, 19-22.

Tambovcev, Yu. A. (1986). *Konsonantnyj koefficient v jazykach raznych semej* (Consonant coefficient in languages belonging to different families). Thesis for the academic degree of candidate of philological sciences. Leningrad (mimeogr.).

Těšitelová, M. et al. (1985). *Kvantitativní charakteristiky současné češtiny* (Quantitative characteristics of the present-day Czech language). Praque, Academia.

Tuldava, J. (1975). *Eesti keele sõnavara statistiline struktuur* (Statistical structure of the Estonian lexis). Tartu (mimeogr.).

Tuldava, J. (1980). Eesti keele sõnavara foneetilis-grafeemilised mõõted (Phonetic-graphemic measures of the Estonian lexis). *Acta et Commentationes Universitatis Tartuensis 518, 51-100.*

Veenker, W. (1982). Konfrontierende Darstellung zur phonologischen Statistik der ungarischen und finnischen Schriftsprache. *Különlenyomat a nyelvtudományi közlemények 84. Budapest, 305-348.*

Viitso, T.-R. (1981). *Läänemeresoome fonoloogia küsimusi* (Problems of Baltic-Finnic phonology). Tallinn, Keele ja Kirjanduse Instituut.

Zörnig, P. & Altmann, G. (1983). The repeat rate of phoneme frequencies and the Zipf-Mandelbrot law. *Glottometrika 5, 205-211.*

Zörnig, P. & Altmann, G. (1984). The entropy of phoneme frequencies and the Zipf-Mandelbrot law. *Glottometrika 6, 41-47.*

Zörnig, P. & Rothe, U. (1992). Confidence limits for the entropies of phoneme

frequencies. *Glottometrika 13, 186-195.*

Zsilka, T. (1974). *Stilistika és statisztika* (Stylistics and statistics). Budapest, Akadémiai Kiadó.