QUANTITATIVE LINGUISTICS

Vol. 56

Luděk Hřebíček

Text Levels

Language Constructs, Constituents and the Menzerath-Altmann Law

QUANTITATIVE LINGUISTICS

Volume 56

Luděk Hřebíček

Text Levels

Language Constructs, Constituents and the Menzerath-Altmann Law

WVT Wissenschaftlicher Verlag Trier

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Hřebíček, Luděk :
Text levels : language constructs, constituents and the Menzerath-Altmann law /
Luděk Hřebíček. -
Trier : WVT Wissenschaftlicher Verlag Trier, 1995
(Quantitative linguistics ; Vol. 56)
ISBN 3-88476-179-X
NE: GT

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

ISBN 3-88476-179-X
ISSN 0179-3616

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

Preface

The text-linguistic experiments described in the following chapters are based on generally formulated models. These models are nothing more than attempts at certain interpretations of the Menzerath-Altmann law with the purpose of demonstrating the deepness and range of applicability of this law. These models are applied to a set of Turkish texts. Certain general properties of text are thus tested as well as their validity in Turkish. This work originated in the Oriental Institute of the Academy of Sciences (Prague).

A group of quantitative linguists working at several German universities for many years very intensively supported the author in his theoretical attempts. The author's greatest thanks are to Gabriel Altmann from the Ruhr-Universität Bochum. It was Altmann who newly formulated the aims and foundations of quantitative linguistics. His devotion to science as well as his friendship helped to surmount certain obstacles in the author's work and life. This help included supplying me with literature, reading manuscripts, providing advice and consultation, etc. His criticism helped to change substantially the manuscript of the present work. The studies contained in this volume join the endeavour of the international group developing the ideas of synergetic linguistics. This investigation is coordinated by the universities of Bochum (G. Altmann) and Trier (R. Köhler). The author is indebted to R. Köhler for reading the manuscript and valuable corrections of errors and inadequacies.
The author expresses great thanks to Sheila Embleton, an eminent specialist in quantitative linguistics, who wasted time with making corrections of my errors and improper formulations. If there is anything correct on the following pages then it is thanks to the above mentioned personalities. Mistakes and errors are mine. During the years 1991-1993 the studies presented were supported by the Grant Agency of the Czech Academy of Sciences. In 1992 they were also supported by Deutscher Akademischer Austauschdienst which enabled the author a short-term attachment at Ruhr-Universität Bochum.

The author is indebted to all these institutions and personalities for their great help and kindness.

L.H.

Contents

Main Problems and Assertions

In order to connect together the individual studies contained in this book, their main ideas, assertions and problems are presented below in an abbreviated form. It can be said that their basic issue is the application of the Menzerath-Altmann (MA) law to text together with the presentation of certain consequences of this law. The main points of these studies can be formulated as follows:

Quantitative linguistics can supplement the descriptions and structural analyses of the great authors working in the field of text linguistics (S. v. Coseriu, T. v. Dijk, W. Kintch, B. Palek, P. Sgall and others) with theorems enabling the testing of data and opening a more direct way to explanations. The present work is an attempt at putting forward such an extension.

Universality of the MA law. The Menzerath-Altmann law (*'The longer a language construct the shorter its components (constituents).'* see Altmann 1980) can also be applied to the supra-sentence levels of text. This law is to a certain degree universal and can serve as a yardstick for an arbitrary language level. It can also describe a gradual arrangement of levels when considering more than two levels.

Semantic structure of a text. One of the possible units belonging to the (both sentence and supra-sentence) semantic structures contained in an arbitrary text is one called *sentence aggregate* which is also defined on the basis of the MA law. An aggregate includes all sentences of a given text in which a certain lexical unit (semantically interpreted or not) occurs.

Synergy (= cooperation of many subsystems of a system, see H. Haken 1978, 334) in language systems. This idea, formulated by R. Köhler in application to language (see, for example, R. Köhler 1986), is compatible with the principle contained in the MA law. This fact stresses the stochastic character of the language subsystems cooperating in all communication processes in natural languages. The MA law seems to be the basic principle of self-regulation in languages.

Text and its producer/receiver. Text taken in the commonly accepted meaning of the term as a language construct separable from a communicator is in fact to a critical degree a vague phenomenon requiring some semantic interpretation. It is senseless to suppose text without a human being who is in contact with the text's *texture* and interpreting it.

The semantic system is a phenomenon contained in each individual human mind. This hypothesis stands against the usual understanding of the concept of semantic system as an abstract phenomenon embedded in natural languages. Language then changes into an abstract phenomenon incapable of being analyzed by an empirical science. Such a position denies the communicative functioning of languages. Any natural language, however, also communicates together with the transmitted information some information concerning the language formation.

Text constructs and word associations are two parallel consequences of a certain orderliness proper to each individual semantic system.

Text is a construct of aggregates. Sentences function as constituents of the semantic structures called *sentence aggregates*. It can be assumed that the same structures function as constituents of the higher construct which is the entire text.

The structure of language levels can be interpreted from the viewpoints of the mathematical theory of communication and fractal theory.

The subsystem formed by a construct and its constituents is a system regulated by coefficient b of the MA law (i.e. Altmann's b) as the subsystem's dimension. The system of all language levels can be characterized as a mutual affine transformation of the sets corresponding to the levels. The number of levels in each language for the present cannot be

stated definitely; between each two levels standing in the relation of a construct and its constituents a new level can be found with the help of the MA law. The ability of a natural language to carry information is based, among other things, on this property to be decomposed into parts, and again parts of parts.

Text is an increasing entity. When one or several sentences are added to the already existing part of a text by its producer, a modified or new text occurs, to which all its preceding parts appear to be its constituents in the sense of the MA law.

We tried to test the assertions presented above as well as some others in several statistical experiments. It is, however, evident that all these ideas are not and cannot be conclusive and that they need further investigation and precision.

1. Introductory Notes

'Similia similibus: we gain knowledge of the world through that which we ourselves are, and recognizing the world we discover ourselves.'
K. Čapek (1934)

'Science cannot concern itself with what and how things are but with what they **are for us**, how we (want to) see them.'
G. Altmann (1993, 4)

The present work is directed to the linguistic investigation of text. This phenomenon is understood here as a segment of the universe of discourse. It is analyzed and interpreted under certain standard conditions of the cognitive approach. This approach is often labelled as science. Stress is laid on the synergetic treatment of the structures observed. Further, we seek consequences of the Menzerath-Altmann (MA) law that are relevant for this phenomenon, i.e. for text; the MA law is connected with the notion of level, and thus text levels are at the centre of our interest.

At first, this aim requires the characterization of the three primitive notions mentioned above:

- text,
- the standard methodological requirements, i.e. methods of theoretical linguistics,
- synergetic treatment of language.

1.1 TEXT

What does one mean in saying 'text'? There is no generally valid invention concerning the comprehension of this concept. Nevertheless, everybody understands it and this term is apparently used by linguists as well as by laymen in the same sense. Therefore we do not

need its definition but what we really need are certain restrictions applied to this phenomenon for the purpose of the exclusion of incoherent and discontinuous discourses, for example, lists of names, subscripts to sets of pictures, tables or graphs, etc. Text means here and below a continuous construct in a natural language. We do not suppose broken or defective texts. If a part of a larger text is analyzed, then this part is always unending. Text is a linguistic phenomenon attributed by continuity.

It is generally accepted that text is a complex phenomenon. In order to understand complicated phenomena, human beings usually divide them into parts and only then try to apprehend them gnoseologically; parts seem to be less complicated than the original whole. Sometimes this idea appears to be true. Any part of a whole can be better observed. However, the deed of partition converts itself into a problem of cognition. Where are the cuts to be placed if they are not to hit some significant segments of the phenomenon analyzed? Which idea or principle represents the basis of this analytical act? If these questions are not answered, something substantial may remain apart from cognition in the acquired phenomenon.

The history of philology has lasted for thousands of years. During this long period text has been still on the periphery of scientific interest. The reason is evident: text is a very complicated phenomenon and therefore only its parts or constituents are at the centre of attention in linguistics. Phones and phonemes, words and their formants, syntagms and at least sentences as the highest units of the philological concern secured the occupation and attention for multitudinous investigators. And language constructs of a higher level, i.e. parts of texts, as well as text itself have remained an item wrapped in mystery.

One circumstance is interesting in this context: The lower units like phonemes, morphemes, syllables and clauses were explained both from the viewpoint of their constituents and from the viewpoint of the relevant higher units. The sentence quietly continued to be the ultimate unit handled by philologists. Sometimes the notion of context occurred in an appending of non-theoretical explanations. Sentence was treated as a syntactic unit, i.e. from its inner properties, and not as a result of text segmentation. Text was then a phenomenon constituted of sentences and sentences were nothing but some corals on a thread. Of course, branches like stylistics, textology, poetics or literary criticism operate with the notion of text; however, they are not able to treat text as a language unit.

During the last decades a new subfield of linguistics has been founded called *text linguistics* or *text science*. This indicates that today the problems of text are at the centre of interest. One remarkable concept was introduced into this branch of linguistics, namely, 'text cohesion', see especially M.A.K. Halliday & R. Hasan (1976). This concept seems to be the first and most substantial attempt of linguistics at surpassing the limits of the sentence. This notion indicates that there must exist some means keeping sentences together and thus forming higher units. The following step was the formulation of the problem which can be called *text as a unit,* see, e.g., Hřebíček (1985). The notion of text cohesion represents the same conception treating text as a sequence of sentences pasted together - this time mainly with the help of text references as instruments of cohesion. Nevertheless, similar to sentences which also are not only sequences of words tied together with the help of grammatical instruments, text is a more complicated structure; text is a system of a specific kind.

There are methodological reasons restricting the range of linguistic interest to sentences. The majority of philologists (with only a few exceptions) understood their aim to be a description of a certain language. Consequently, their aim was not the scientific explanation of the object of their interest as is generally required as a result of each scientific activity. This is quite understandable; scientists dealing with languages have as their main task writing textbooks, language grammars and dictionaries as well as solutions of orthographic problems and formulation of linguistic norms, codification of the language expressions, formulation of hypotheses concerning language history, etc. All these and other similar tasks are of applicational character. They involve many difficult problems which cannot be solved otherwise than with the help of a deep intuitive knowledge of the respective language and anticipation of its structure on the basis of semantic intuition. Scholars able to solve such problems and then to compile works concerning languages are worthy of full respect.

Such aims and tasks, however, lead to procedures including certain formalizations that can be characterized as the formulation of rules - rules for the creation of everything. Language appears to be a complicated instrument and philology, or linguistics, supplies it with directions for use. Our century improved this task by demanding automatization and the generation of language constructs in an automatic way.

One may ask the following question: What is in fact applied in those applicational operations? Where is the theory which is (consciously or unconsciously) applied in the practical usage of

a language? The branch consisting of rules instead of theoretical expressions seems to be nothing but a technology based on concepts and classifications having scholars' subjective choice as its background.

In section 6.1 of his famous work 'Syntactic Structures', N. Chomsky (1957) suggested that each grammar of a language (literally: of language L) represents a language theory. Thus linguistic theory was reduced to grammar. Chomsky declared that theory is the formulation of general laws in terms of hypothetical constructs, such as 'mass' or 'electron' are in physics; these constructs are allegedly based on a finite number of observations. Theory should be able to state the relations among the observed phenomena and to predict new phenomena. Consequently, the most important gain expected from theory is the selection of a grammar from a set of possible grammars.

Chomsky's followers decided to accept this position in relation to scientific theory. For example, J. J. Katz & P.M. Postal (1964), without the cautiousness shown by the head of the school, declared that the description of a natural language in the form of a system of rules is a scientific theory from which phonological, syntactic and semantic rules can be deduced. This alleged linguistic theory is, according to the opinion of these authors, approved if the rules of any correct description have the form prescribed by this theory. This is apparently circular reasoning: 'theory' equals 'rules', rules prescribe what is correct and then theory is also correct.

Nonetheless, these ideas propagated rapidly among linguists. No wonder, since they contain a clear explication of what theory is; and at the same time a lucid and easily comprehensible prototype of all correct linguistic theories is offered. For this reason a great proportion of the works executed in text linguisticss have the form of a set of rules for generating correct texts. This endeavour, however, could not reach its goal.

1.2 METHODS APPLICABLE IN THEORETICAL LINGUISTICS

Everything that should be mentioned in connection with this topic is already contained in the works of Gabriel Altmann and Reinhard Köhler. The methodological principles formulated for the first time by these linguists are followed in the present work. Let us outline in a shortened form the main ideas formulated by Altmann (1993).

The most important question for each methodology is: What is meant by the concept of science? This question cannot be answered only with regard to the experiences of linguistics as they occurred in the formulations of the generative school. The cognitive activity called 'science' is broader, and all sciences as well as the level of development reached by them must be considered. The reason is obvious: human cognition cannot be postulated as an item segregated into pieces having little or nothing in common with one another. No science can ignore fundamental results and proceedings leading to these results in other sciences. Everyone in science should look around themselves and take interest in the meaning of the notion of science in the other branches.

Altmann frequently refers to the works of many specialists in general methodology and philosophy of science, but mainly to the works by Mario Bunge (1967, 1983, and other works). This enables him to formulate quite anew the tasks and aims of theoretical linguistics. These principles are applied in Altmann's own linguistic works, where also his general methodological formulations are presented; see, for example, Altmann (1987, 1988, and others).

The objects investigated by a science are freely chosen. Their forms are apprehended in concepts which are purely qualitative and thus expressed in words of natural languages. They can also be quantitative, which has many advantages. Both kinds of concepts do not depend on the nature of the objects; quantities and qualities are only properties of our concepts used for the purpose of bringing order into reality. Besides objects, the approaches (aspect, aim, problem, method) represent important constituents of sciences. The most significant property, however, which differentiates science from, let us say, novel-writing or essayistics is theory.

An arbitrary set of expressions concerning some topic cannot be supposed to be a scientific theory. Theory exhibits quite precisely specified structure: it includes concepts, conventions and - *hypotheses.* Concepts represent a necessary but not a sufficient condition for a theory. 'Convention' represents a summarizing concept for definitions, operations, rules, etc. Scientific hypotheses are the *sine qua non* of a science. With reference to Bunge (1967, 354-361), Altmann writes:

> 'Only syntactically well-founded, semantically meaningful general statements that are empirically testable, not including observational concepts, stating

something about invariances and going beyond our present knowledge should be considered as hypotheses. If a hypothesis is derived from assumptions (axioms) or from a theory, if it is corroborated by an empirical test and if it can be connected with other similar statements (systematized), then we can call it a law.'

Laws can also be characterized as statements concerning observable phenomena and mechanisms generating them. All sciences always strive for the formulation of laws. Laws systematize our knowledge about the investigated objects.

It is needless to stress that the majority of linguistic investigations are not in agreement with these requirements. Concepts, definitions, classifications and operational conventions (for example, in the form of substitutional rules) represent the greatest portion of the theoretical knowledge about languages. It is quite natural when this knowledge is used for different applicational goals. Theory, however, cannot be presented in this incomplete form. Theory cannot miss a refutable hypothesis. This requisite was formulated by the famous philosopher Karl R. Popper. In his epistemological system the idea of refutation is put as an equivalent of testing. What is refutable, is testable, and *vice versa*. Existential hypotheses such as 'there is life on Mars' need not always be refutable; however, this property is also limited in time. This means that none of the empirical laws is formulated for eternity. On the contrary, scientific laws are only valid until somebody succeeds in their rejection through testing.

This is essential for the difference between mathematical assertions concerning abstract structures and the laws of empirical sciences which always concern some observable parts of reality. Mathematical propositions once proved are unremittingly valid because the principles according to which their validity is examined are constant. On the other hand, the validity of the laws in empirical sciences is always considered in relation to the objects of reality to which they refer. It is generally accepted that reality and our view of it is far from being constant. Let us point to the psychological theory formulated by Jean Piaget, especially his rejection of the opinion that language produces thinking with ready-made structures. He emphasizes the collective educative impact of language (see H.G. Furth 1969, 121).

It should be noted here that language evidently is an observable object with the general attributes indicated above proper to reality. Let it be stressed that language taken in its

synchronic aspect represents an open system. This system is in close connection with the human brain and through it with many systems of reality.

Some linguistic schools treat language as an abstract phenomenon. Such an approach deserves respect, but it must be emphasized that the object of investigation of such linguists is something different from the phenomenon analyzed by philologists and linguists. These scholars introduce abstract concepts and with their help they construct abstract systems. These systems have the character of items described in mathematics; they are exactly defined and algebraically operable. This kind of linguistics describes languages in the manner usual for the description of artificial languages (e.g. the language describing logic). They can formulate assertions like the one concerning the sum of angles in a triangle: this sum equals 180° only when we move in Euclidean space; when we pass to another geometry, for example, to that of Lobachevsky or Riemann, this truth changes into untruth. This or that geometrical system as well as other abstract systems can be used for application in different sciences; its choice must be sustained empirically by testing. Each empirical theory (refutable scientific theory) may contain such an abstract system. Naturally, nothing can be said against the geometrical systems of Euclid, Lobachevsky or Riemann in general; however, only a certain theoretical context, i.e. an application in an empirical theory, can decide which one of such abstract systems is correctly applied in relation to a given problem, to that part of reality which became the object of investigation. To formulate it briefly, abstractivists (in linguistics and in other branches) construct an *a priori* space and try to identify it with some real space without testing their mutual concordance. Their proofs concern only those items which they introduced into their considerations through their own definitions. Language, as well as everything in reality, is an infinite phenomenon; consequently, the set of applicable definitions and rules is also infinite. Each language evidently operates in many abstract spaces and a principle of selecting one of them must be based on a scientific (= refutable) theory. Each abstraction (in the form of definitions or rules, etc.) when applied to reality requires an empirical testing. Abstractions cannot be tested other than in a context of empirical hypotheses. When they are "proved" with the help of axioms on which the abstract system (e.g., of generative rules and transformations) is based, then the internal cohesion of the abstract system itself is possibly tested, but not its agreement with observations.

All these methodological requirements define science as a sort of literary category. Similarly to, e.g., a sonnet which scarcely ever can have more or less than fourteen lines and a certain

scheme of rhymes, scientific theory has its quite clearly defined structure: its core is represented by a testable hypothesis and all other procedures are related to this core, or are clearly derived from it. Scientific theory cannot be supposed to be an arbitrary sequence of definitions, classifications, rules or abstract assertions.

1.3 SYNERGETIC TREATMENT OF LANGUAGE

Synergetics is a discipline founded by the physicist Hermann Haken (1973, 1978). This theory describes the behaviour of the systems composed of many subsystems. These subsystems are often well organized while the total system remains unclear, and its manner of organization can hardly be deciphered. Nevertheless, the entire system behaves in a way which can create order. H. Haken writes:

> 'Thus the question arises, who are the mysterious demons who tell the subsystems in which way to behave so to create order, or, in a more scientific language, which are the principles by which order is created' (Haken 1973, 9).

The founder of the discipline suggests that in spite of the completely different nature of subsystems, their behaviour is regulated by a few very general principles which offer an explanation of the similarity in the conduct of such complicated systems. Synergetics is characterized as a theory of cooperative phenomena in multi-component systems. The pilot principle of the functioning of such systems is self-regulation and self-organization. One of the fundamental ideas of synergetics is the idea of seeking the *order parameter*. This concept is coined with the purpose of two functions:

> 'It describes order because it is zero in the disordered state and assumes a maximum value in the completely ordered state. (...) In a more modern language one would say the order parameter gives *instructions* to the subsystems. Order parameters need not necessarily be fictitious quantities' (Haken 1973, 10-11).

The general formulations concerning synergetics by the founder of this branch are closely connected with the problems of physics. It can be supposed that more complicated systems, as is the case of natural languages, are arranged by more than one parameter of order, and

their organization can be represented by a set or system of such parameters. This is the predicament of natural systems functioning as a means of communication, the best one of them - language - being permeated by complicated structures. This, however, is not only the case of natural languages but also of the matter forming all living creatures.

The concept of 'instruction' mentioned by Haken involves the idea of communication. To give instruction means to give impulses to a movement of the respective system from one of its states to another one. Such an instruction can also concern the creation of a new element which will enter the set of the elements forming the respective system.

Each beginner in theoretical linguistics is usually instructed with the help of the postulate asserting that each natural language is a system consisting of relatively independent subsystems. These subsystems are: phonology, morphology, syntax, semantics, and perhaps also the lexical and stylistic levels. This position is common to almost all classic linguistic schools, if they operate with the notion of system at all.

Modern linguistic schools treat language subsystems as mutually separated items, see, e.g., the generative and transformational treatment of phonology and semantics as interpretations of the main string of abstract rules and transformations. 'Interpretation' means, however, nothing more than construction of a new string of phonological or semantic rules which are also based on some arbitrarily selected abstract notions.

One exception is contained in the idea of text cohesion mentioned above, though the ties forming the cohesive relations are usually described similarly to the description of the phonological, morphological, etc. phenomena. Text ties are thus also described as members of the relatively independent subsystems. Our ability to describe language as a system of subsystems is limited by our inability to characterize the mutual relations of the subsystems forming from them one whole.

The methodological position appearing to be the closest one to the character of the complicated language system is the one operating with the notion of function. This concept, however, has not very much in common with the mathematical term "function". In mathematics, function means something like a defined dependence, while in linguistics the same term is understood as a synonym of "purpose" or "role" (e.g.: 'The function of the

definite article is to indicate phenomena already known or mentioned in the preceding discourse,' or something of that sort).

The application of the theoretical principles of synergetics, as they were elaborated in physics (by H. Haken) and in chemistry (by I. Prigogine), in linguistic theory appears to be incentive. This idea was applied in linguistics for the first time by Reinhard Köhler (1986) and developed in his further works. Köhler writes (on p. 34 and 154) about synergetics:

> 'Ihre grundlegenden Prinzipien sind so allgemein, daß die sprachliche Selbstregulierung lückenlos in ihren Beschreibungs- und Erklärungsrahmen paßt. (...) Andererseits bietet der dargestellte Ansatz - insbesondere dank des fundamentalen Axioms der Selbstregulation - die Möglichkeit, sämtliche sprachwissenschaftlichen Forschungsaspekte in einem Modell zu vereinheitlichen.'

As was stressed by Köhler, the synergetic approach enables linguistic theory to exploit tools elaborated in the neighbouring disciplines and in the modern natural sciences.

When linguists during their investigations move from one level to another one, from a language subsystem to another, the image acquired in these investigations substantially changes. When a linguist decides to view language as a system, the task is to seek something remaining unchanged in this respect. It can be said that the aim of linguists is to seek invariants in the system of subsystems.

1.4 TERMINOLOGICAL CONVENTIONS

The reader's attention should be drawn in advance to several terminological conventions applied in the following chapters which may cause misunderstandings.

If it is not exceptionally stated otherwise, the term 'sentence' is used below in the sense "the result of a certain segmentation of text." Text is segmented into sentences and/or clauses. 'Sentence' then means "sentence or clause", i.e. segments syntactically based on finite or infinite verb forms. There is a certain degree of similarity in the syntactic functioning of these two units. The reasons for this approximation are discussed in sections 3.5 and 4.5.

Consequently, where the term 'neighbouring levels' is used, it does not imply that the entire staircase of levels, i.e. all the steps formed by language subsystems, is known. This mainly concerns different syntactic levels, be they recognized or potential ones, not only those formed by sentences and clauses. This problem differs from language to language and it remains far from being solved. Let us mention, for example, the lack of clarity in understanding the word level in Chinese. The general solution of language levels evidently consists in application of the MA law as a criterion. The individual language items to which this law is applicable must always be discovered and tested.

Misunderstandings may occur in connection with the usage of any term and its meaning. This also concerns the term 'dimension.'

1.5 DIMENSION AND LANGUAGE

Opponents of quantitative linguistics often argue using the idea that in language there do not exist such quantities, variables and constants, as they are in physics and other natural sciences. According to these suppositions language is a dimensionless phenomenon containing sets of discontinuous elements. For this reason the opponents refuse not only explanatory metaphors using geometrical imagination but also quantitative arguments and testing.

Each kind of geometrical theory, be it classical Euclidean, non-Euclidean, or Mandelbrot's fractal geometry, is based on the concepts defined in set theory. While in dictionaries the word 'dimension' is usually explained as the number of independent coordinates, the same need not be valid for a mathematical context.

It was stressed by Benoit B. Mandelbrot (1982, 14 ff.) in his work *The Fractal Geometry of Nature* that in mathematics it had already been accepted that one cannot be satisfied with defining dimension as a number of coordinates. He stresses that while Euclid is limited to sets for which all the useful dimensions coincide, and thus they are *dimensionally concordant* sets, Mandelbrot's ideas and his fractal theory concern the sets which are *dimensionally discordant.* Doubtlessly such sets can also be found in languages. In the quoted work (on p. 16) Mandelbrot writes:

'...this Essay often invokes *effective dimension*, a notion that *should not* be defined precisely. (...) In other words, effective dimension inevitably has a subjective basis. It is a matter of approximation and therefore of degree of resolution.'

His theory is dedicated to seeking invariants under certain transformations of scale. One can scarcely argue against the supposition that one of the important scales in languages is represented by language levels. We will try to indicate that the theory based on the MA law offers a sufficiently general theorem for giving explanations of this very scaling. This does not mean that everything in languages and in semantics is explicable with the help of this theory. However, the MA theory
has a large set of consequences and as such it deserves linguists' attention.

2. The Menzerath-Altmann Law on Lower Levels

In this chapter, after a recapitulation of the basic properties of the Menzerath-Altmann (MA) law, several possibilities of its application in text linguistics are discussed. Our main intention is to indicate later that the principle which is valid for lower levels (i.e. such as the level of phonemes, morphemes, syllables) is also valid for text and its higher structural parts.

2.1. DERIVATION OF THE LAW

The Menzerath-Altmann law was derived as a continuous function; see Altmann (1980). This derivation represents a standard approach to the relation of language constructs and their constituents. Let us briefly recapitulate the steps leading to the discovery and derivation of this law with reference to the basic linguistic works in which it was derived and applied.

Altmann started with the assertion first formulated by P. Menzerath in 1928 and also later in his book *Die Architektonik des deutschen Wortschatzes* (1954, p. 100). It must be stressed, however, that the importance of his idea was not recognized by linguists until 1980; Altmann's merit consists in discovering the great possibilities for language knowledge and theoretical linguistics hidden in this idea. Menzerath writes:

'The relative number of sounds in the syllable decreases with the increasing number of syllables in word, or said in a different way: the more syllables occur in a word the (relatively) shorter the syllable is.'

The observational concepts such as sound, syllable, word, etc., were removed from this

formula by Altmann, enabling him to reformulate it in the following general form:

'The longer a language construct the shorter its components (constituents).'

Thus the realm of validity of the law has been widened from syllables and their constituents to other possible language constructs.

Altmann starts with the differential equation (see Altmann & Schwibbe 1989, 6):

$$\frac{y'}{y} = \frac{b}{x}, \qquad (2.1)$$

wher y = mean length of constituent,
x = length of construct,
b = constant.

Equation (2.1) can be interpreted in two ways:

- the increase of the length of constituent (y') is proportional to the length of constituent (y);
- the increase of constituent (y') is inversely proportional to the length of the respective construct (x).

The variables in (2.1) are separated so that the following solution can be written directly

$$\ln y = b \ln x + c.$$

With $A = e^c$ the formula of the MA law is obtained:

$$y = A x^b \qquad (2.2)$$

This basic formula was analyzed and tested on various observed data obtained from different languages for the constructs of different levels; see the literature quoted above. With regard to those results, the law can be taken as strongly confirmed. Altmann used two techniques for estimation of its parameters: the method of least squares and an iterative method. Furthermore, a technique based on the analysis of variances of the variables was presented

by Michael H. Schwibbe (see Altmann & Schwibbe 1989, 26-39).

The relation between constructs and their constituents offers a criterion for distinguishing levels in languages. Level can be understood as a synonym of language subsystem. Level is, in fact, a consequence of the MA law. This means that from this law many other structures and their properties stem. This was convincingly demonstrated by Köhler (1986), who analyzed the structure of vocabulary in connection with the amount of the total vocabulary of a language and in connection with the number of phonemes, word length, word frequency and other variables; see also Köhler (1982) and (1989). Another sphere of application of this law was demonstrated by A. Fenk & G. Fenk-Oczlon (1993). These authors observe phonemes and syllables in words and sentences in different languages together with the interpretation of dynamics of these entities.

2.2. EXEMPLIFICATION OF THE LAW

Let us exemplify the MA distribution on data obtained from Text 7 of the Corpus of Turkish texts (their list is presented below); see TABLE 2.1. The procedure for obtaining the values of both the parameters and the expected values rests upon the method of least-squares; see Altmann (1980). The coefficient of determination D is, in fact, the squared correlation coefficient r. It represents the ratio of the variation explained by the MA law to the total variation. When there is zero explained variation, the total variation is all unexplained and the ratio D is zero. On the contrary, if there is zero unexplained variation, the ratio equals one. Consequently, this coefficient exhibits the degree of the explanative ability of the law in relative numbers (percentages).

It must be stressed, however, that the functioning of the MA law is observable only when the distribution of the data, concerning the two variables involved, is stable. This means that the text analyzed must be sufficiently long, otherwise the functioning of the MA law cannot emerge from the variation of data. For example, for the data in Table 2.1 concerning higher word lengths, there are few records and thus these numbers are unreliable.

TABLE 2.1: Word length and syllable/morpheme length

Text 7 (see Corpus of Turkish texts below)

Syllables	z	p	y	Y
1	62	149	2.40	2.42
2	142	671	2.36	2.34
3	176	1211	2.29	2.29
4	114	1039	2.28	2.25
5	51	563	2.21	2.23
6	9	118	2.19	2.21
7	4	70	2.50	
8	1	19	2.38	

A = 2.4213; b = -0.0520; D = 0.9307.

Morphemes	z	p	y	Y
1	180	752	4.18	4.34
2	160	1099	3.43	3.23
3	126	1069	2.83	2.72
4	70	636	2.27	2.41
5	15	164	2.19	2.19
6	5	75	2.50	
7	2	26	1.86	
10	1	19	1.90	

A = 4.3403; b = -0.4246; D = 0.9670.

s = word length in number of syllables
z = number of words having length s (or m);
m = word length in number of morphemes;
p = sum of phonemes occurring in the respective syllables (or morphemes);
y = observed mean syllable (or morpheme) length (in number of phonemes);
Y = expected syllable (or morpheme) length (in number of phonemes).

A and b are parameters of the MA law obtained with the help of the method of least squares; the expected values Y are computed according to (2.2).

This problem is connected with the question of sampling. One must carefully weigh the question of the possible corruption of the relation between constructs and their constituents by the act of sampling. It will be indicated later that the MA law has something in common with the choice of language units made by producers of language constructs. In Table 2.1 we do not present the results of the Wilcoxon test, which is used in the subsequent tables only for testing the concordance between the observed and expected distributions of means y and Y. Here the concordance is evident.

The data observed concerning the length of syllables and morphemes in different languages and a demonstration of their agreement with the MA law can be found in Altmann & Schwibbe (1989, 48-59).

2.3. MEASUREMENT CONTAINED IN THE MA LAW

Formula (2.2), when cautiously analyzed, indicates that underneath the variables involved certain hidden relations are present.

Let us suppose a construct with length $x = 1$; according to (2.2) the value $y_1 = A$ is obtained. One of the parameters of the Menzerathian function is defined by, so to speak, a unit of the measurement. This is evident if the formula is rewritten with y_1 instead of A and unfolded as a progression in the following way:

$$\begin{aligned} y_1 &= y_1 \\ y_2 &= y_1\, 2^b \\ y_3 &= y_1\, 3^b \\ &\dots \\ y_x &= y_1\, x^b \end{aligned} \qquad (2.3)$$

The structure of each element of this row is the same as if, for example, the length of a line is expressed as $L = 3$ m. Instead of meters, here is the mean length of constituents for the construct of length $x = 1$. Consequently, each level has its own measure in the number of units defined as $A = y_1$. Possibly this is a sort of *effective dimension*; see the quotation from Mandelbrot in Section 1.5.

This enables us to try to derive the formula (2.2) in a new way. With respect to (2.3) and to the distributions of the observed data corresponding to the MA law, we can formulate the following hypotheses:

1. y_x is proportional to y_1, i.e. $y_x \sim y_1$.
2. y_x is inversely proportional to x, i.e. $y_x \sim 1/x$.

The two hypotheses can be combined together, supplemented by a proportionality coefficient, say, b and the whole expressed in a logarithmic transformation:

$$\log y_x = -b \log x + \log y_1 \qquad (2.4)$$

This is another expression for the general term of the progression (2.3); it is also identical with (2.2), when $A = y_1$. Evidently, the two hypotheses presented above, combined in (2.4), are counterparts to the general formulation of the MA law. It must be emphasized, however, that speaking about units of measurement (i.e. A or y_1) we use nothing but a figurative expression. It is generally accepted that each unit of an arbitrary measurement is proportional to the size of the measured object. A text producer who implements a certain distribution of the relevant language elements declares the "units of measurement" for the abstract space into which he/she situates the language construct produced. Thus, together with these constructs, their dimension is exhibited. This is the deeper sense of the statistical properties of the universe of discourse.

The hypotheses presented above can be abridged to an assertion saying that

the mutual relation of y_x and y_1 is proportional to x.

This follows directly from (2.4). Then the formula

$$\frac{y_x}{y_1} = x^b$$

can obtain an inverted form

$$\frac{y_1}{y_x} = \frac{1}{x^b} \qquad (2.5)$$

which will be proved below to be important in connection with the question of distribution

corresponding to the data of the MA law. Thus, if the relation of the two variables is put in the form y_1/y_x, it is inversely proportional to x. The variable y is a function of x which is monotonously increasing with positive b, and monotonously decreasing when $b < 0$. The relation of constructs and their constituents exposes the latter one of the two possibilities. This seems to be one of the fundamental properties of languages.

In connection with (2.4) another interpretation of the parameters can be assumed. In a logarithmic transformation the expression *log* y_1 (i.e. *log* A) becomes a quantity newly defining the initial point of the logarithmic system of coordinates in which the relation of constructs and constituents is described: the initial point is shifted by the distance *log* A along the coordinate *log* y. Then the dimension of this system is expressed as b having the sense of the "measure" of the same system.

It can be said that units of different levels have their own "measures". Two different relations of levels, say, α and β, and γ and δ, are mutually comparable only if their Menzerathian coefficients A are equal. It is needless to stress that "measures" have different meanings, for example, the number of syllables having the length one phoneme on the one hand, and the number of sentences/clauses having the length one word on the other hand. Each special theory must surmount these problems. Speaking about the unit of measurement in connection with A, we mean the measure for a given distribution derived from the same distribution and stated as the mean length of constituents related to constructs having length $x = 1$.

Altmann stresses that the meaning of the word 'length' used in the formulation of the MA law should be understood in a wider range of its possible metaphorical unfolding. We can also speak about 'complexity' (Altmann & Schwibbe 1989, 5):

> *'Je größer ein sprachlicher Konstrukt, desto kleiner seine Konstituenten.*
> "Größer" bedeutet eigentlich "komplexer", d.h., aus mehr Entitäten bestehend, "kleiner" bedeutet "einfacher", d.h., aus weniger Entitäten bestehend. Entität ist nicht unbedingt ein materieller Teil, es kann auch eine Funktion, Relation, Bedeutung usw. sein.'

This is another aspect of the universality of the MA law. As for the other constant of the law, see Section 3.7 where Altmann's parameter *b* is discussed in connection with the data concerning aggregates.

3. Text Aggregates

3.1. LEVELS AND SUBSYSTEMS

It is obvious that the MA law represents the only objective criterion for discovering and establishing any level in languages. Level represents a sort of classification or typology. Units which are constructs and units which are constituents in relation to certain constructs form two typological classes. The approach to levels through the MA law represents a strong empirical sustenance of classification.

Level is sometimes also called 'subsystem'. This often occurs in the usual characterizations of languages which are so often referred to as systems of *relatively independent* subsystems. The subsystems are partly independent and partly not; this is the specific meaning of the relativity mentioned.

The indicated expression frequently occurring in linguistic descriptions, however, means rather that we are not able to show the degree of dependence and independence of the language subsystems. Nonetheless, the MA law offers one solution to this problem. How should one treat language levels? Is the number of levels limited, so that there exists nothing more than the generally known levels (phonemes, morphemes, syllables, words, syntagmas, etc.)? Or can we expect that a new level will emerge, e.g., between phonemes and morphemes?

Let us stress a certain general methodological position which cannot be avoided in the following argumentation. Each object of empirical investigation can be assumed to be infinitely complicated as one moves in its treatment into more and more finer details.

Everything which we notice in the structure of reality depends on our ability to presuppose, and then to grasp it theoretically.[1] Language seems not to be different in this respect from an arbitrary sort of reality. In language, too, with advancing knowledge we are able to assume and observe structures which were not earlier considered. One of these advancements in language knowledge is the MA law with the precisely defined concepts of construct and constituent.

Thus it is correct and incontestable to presume that the set of language levels is unlimited. This set is open when we move down to lower levels, as well as up to higher levels. Such a higher level is, for example, represented by a unit called text, as we will try to prove in what follows; and then it is natural to assume the existence of a further level between the level of sentence and that of text.

Such a level was found, and its units are called 'aggregates',[2] i.e. language constructs found in texts and defined with the help of the MA law. Their constituents are sentences of a text together with the occurrence of a common lexical element. More details are presented in Hřebíček (1989, 1992), where their existence was tested on several Turkish and Old Ottoman texts. Aggregates were also observed and their distribution analyzed in German texts by C. Schwarz (1992). Below in this volume we offer a new tested hypothesis concerning the distribution of aggregates.

In what follows we present a short recapitulation of the basic definitions and their consequences.

[1] As was stated by G. Altmann, these ideas are formulated in a concise form by S.N. Salthe (1985).

[2] The term 'aggregate' seems to be not quite convenient and, consequently, provisional. We do not know yet whether the elements of an aggregate form an organized and arranged unit or whether aggregate is only a set of inconsistent constituents. This problem deserves further investigation. In any case, the term 'aggregate' is used here as a merely distinguishing and non-defined sign; it is not meant as an antithesis to 'system.'

3.2. DEFINITION OF AGGREGATES

'Aggregate' or 'sentence aggregate'[3] is a set of sentences of a text in all of which a certain word (lexical unit), or semantically interpreted word, occurs. By 'semantic interpretation' we denote the following two indications:

- of the subset of lexical units occurring in a text and having an identical meaning;
- of the subset of mutually differing lexical units.

Thus a new vocabulary of the text is elaborated (in the text producer's or the recipient's mind), possibly more or less differing from the standard lexicons of the given language or from semantic interpretations of (another, or even of the same) text made by other language users. (Semantic interpretation of a text is discussed in Section 3.6.) The following hypothesis concerning aggregates can be formulated:

> The longer an aggregate (in number of sentences) the shorter the mean length of its sentences (in number of words).

This assertion is a modification of the verbal formulation of the MA law. The existence of aggregates fills in the structural hiatus between the level of sentences and the whole text. It seems to be true that during the whole history of language knowledge, no unit was found and tested in the structural space between these two levels. Obviously, text always was non-explicitly understood as a linearized conglomerate of sentences. We can pose the question as to which the mutual relations of sentences are. The only exception in this sense represents the eminent idea of *text cohesion* applied in the last decades; see especially Halliday & Hasan (1976). This idea can be transformed into a testable (refutable) form; see Hřebíček (1985, and also 1989, 1992).

It can be said that it is an indisputable duty of linguistics to seek new structures in gaps where structural elements are missing. The intuition of language users cannot satisfy the task

[3] In Hřebíček (1992) the term 'aggregation' was used instead of 'aggregate' as in the present work. The terms 'vehicle aggregation' and 'sign aggregation' represent a terminological curiosity; here we speak about *(non-interpreted) aggregates* and *(interpreted) aggregates* or about *sentence aggregates* instead of *aggregates.* The provisional nature of all these terms is mentioned in the footnote on the preceding page.

of scientific explanation. The assertion concerning the possibility of gaps between language levels should also be seriously investigated. And with reference to the general methodological guess presented above about the unlimited character of the objects studied by linguistics, we dare to presume that aggregates will appear not to be the only language level in the space between sentences and text.

The relevance of the structure called 'aggregate' should not be a wonder for philologists. A number of highly important texts such as the Bible, Koran, Shāh-nāme by Ferdousī, plays by Shakespeare, etc., are supplemented by text concordances. Each concordance contains either all, or only some, important lexical units of the respective text together with indication of the location of each unit in the text, so that the environments, e.g. sentences, in which a unit occurs can easily be found. It is obvious that a complete text concordance is, in fact, certain evidence of aggregates. Consequently, aggregate is not a novelty for language specialists. On the basis of the MA law, however, we can treat it as a unit of a language level and develop a refutable theory of aggregates.

Sometimes it is difficult to explain the functioning of aggregates in a language system as such a functioning is usually understood. In aggregates the units of different levels, i.e. of (interpreted or non-interpreted) lexical units and sentences, are connected together with semantic ties. The MA law is then the instrument of their explanation. Typologically differing units, such as words and sentences, are connected together with systematic ties, which should be disclosed in a scientific way when an explanation of a language system is sought. The MA law appeared to be an instrument facilitating this.

Obviously, the number of aggregates of a text equals the number of lexical units. It was already indicated that this number is, in fact, a variable dependent on the semantic interpretation of lexical units by language users. This variable v has two extremes: a text can be assumed, in which all word units are interpreted as one and the same unit; on the other hand, each of its lexical units has a frequency not greater than one. The observed values of v always are far from both these extreme values, when the supposed text is sufficiently large. This fact guarantees a high variability to this quantity.

The approval by linguists of the idea of aggregates as units represents a complication, apparently for the reason that aggregates are not ordered together in texts in a way similar to

phonemes, morphemes, syllables, etc. Aggregates represent a permeating structure - they are not characterized by the simplicity of linear arrangement. We can show, however, that not all the conceded levels are linearly ordered; for example, in different languages syntactic relations, that of sentence subject and its predicate in a longer sentence, permeate other syntactic structures and these structures are not always coordinated one to another, but mutually embedded. As we ascend to the higher language levels, the relation 'to consist of' becomes more complicated and remote from its common understanding. It is not difficult to conceive the way of coordination of semantic units, whatever they can be, as permeating structures. And in the case of aggregates we are in contact with the semantic subsystem (= level) of a text. This property will be evident later in connection with aggregates and word associations. We suppose that in aggregates the semantic system hidden under the levels of language expression surprisingly breaks out. Linguistics, when seeking language structures, needs not only to investigate the sequence of units, but also to bracket the linear form of language expressions, being nothing but a surface evident to anybody at first sight.

On a short English text borrowed from Popper (1963, 4), we will demonstrate what sort of units sentence aggregates are. This text is continuous and represents a closed section of Popper's philosophical work *Conjectures and Refutations.* The sequence of its sentences and clauses with their ranks are quoted below, the embedded clauses are ranked after their respective sentence and their position in the sentence is marked by /.../. Definite and indefinite articles and prepositions are not classified as separate units with lexical meanings. In the following presentation, after each sentence/clause its length in number of words is indicated in brackets.

AN ANALYZED ENGLISH TEXT

Quotation from Popper (1963, 4).

1 The problem /.../ may perhaps be described as an aspect of the old quarrel between the British and the Continental schools of philosophy -(15)
2 which I wish to examine afresh in this lecture, (7)
3 and which I hope (4)
4 not only to examine (3)
5 but to solve (2)
6 - the quarrel between the classical empiricism of Bacon, Locke, Berkeley, Hume, and Mill, and the classical rationalism of Descartes, Spinoza, and Leibniz. (19)
7 In this quarrel the British school insisted (5)
8 that the ultimate source of all knowledge was observation, (7)
9 while the Continental school insisted (4)
10 that it was the intellectual intuition of clear and distinct ideas. (10)
11 Most of these issues are still very much alive. (8)
12 Not only has empiricism, /.../ conquered the United States, (5)
13 still the ruling doctrine in England, (4)
14 but it is now widely accepted even on the European Continent as the true theory of *scientific* knowledge. (14)
15 Cartesian intellectualism, alas, has been only too often distorted into one or another of the various forms of modern irrationalism. (15)
16 In this lecture I shall try to show of the two schools of empiricism and rationalism (10)
17 that their differences are much smaller than their similarities, (9)
18 and that both are mistaken. (5)
19 I hold (2)
20 that they are mistaken (4)
21 although I am myself an empiricist and a rationalist of sorts. (8)
22 But I believe (3)
23 that, though observation and reason have each an important role to play, (10)
24 these roles hardly resemble those (5)
25 which their classical defenders attributed to them. (6)

26 More especially, I shall try to show (5)
27 that neither observation nor reason can be described as a source of knowledge, in the sense (12)
28 in which they have been claimed to be sources of knowledge, down to the present day. (10)

AGGREGATES

Larger (non-interpreted) aggregates observed in the quoted text are presented together with the lexical units on which each aggregate is based. The appended numbers are ranks of sentences. The numbers in brackets indicate the length of the respective aggregates (in number of sentences).

to be:	1, 8, 10, 11, 14, 15, 17, 18, 20, 21, 27, 28 (12)
and:	1, 3, 6, 10, 16, 18, 21, 23 (8)
I:	2, 3, 16, 19, 21, 22, 26 (7)
that (conj.):	8, 10, 17, 18, 20, 23, 27 (7)
this:	2, 7, 11, 16, 24 (5)
knowledge:	8, 14, 27, 28 (4)
school:	1, 7, 9, 16 (4)
which:	2, 3, 25, 28 (4)
as:	1, 14, 27 (3)
but:	5, 14, 22 (3)
empiricism:	6, 12, 16 (3)
observation:	8, 23, 27 (3)
only:	4, 12, 15 (3)
quarrel:	1, 6, 7 (3)
source:	8, 27, 28 (3)
they:	20, 25, 28 (3)

In addition, 21 aggregates of the length (2) and 86 aggregates of the length (1) were observed.

The data characterizing aggregates which occur in Popper's text are contained in the

following table:

x	z	n	y	Y
1	86	928	10.79	11.41
2	21	408	9.71	9.63
3	8	234	9.75	8.72
4	3	104	8.67	8.13
5	1	35	7.00	7.70
7	2	96	6.86	7.09
8	1	80	10.00	
12	1	116	9.67	

A = 11.4095; b = -0.2445; D = 0.8294.

Wilcoxon T = 10 > $T_{0.05}(6)$ = 0.

We can say that about 82% of the variation observed on *y* is explained by the MA law. The distribution of *y* is not significantly different from that of *Y*, as is testified by the Wilcoxon test. It is evident that in this text by Popper the MA law is valid. The estimate of parameters is based on the observed values of *x* equalling 1 to 7 inclusive. The extreme values for *x* = 8 and *x* = 12 were not taken into account as they represent the conjunction *and* and the copula *to be* which differ in the character of their meanings and function from the other units. When the lexical units are interpreted together with all text references, the results are more convincing.

3.3. AGGREGATES OBSERVED IN TURKISH TEXTS

The observed data for aggregates in several Turkish texts are presented in TABLES 3.1.1-3.1.10. In these Tables, the following variables are presented:

- length of aggregates *x* (in number of sentences),
- number of aggregates *z* having length *x*,
- total length *n* of *z* aggregates (in number of words),
- distribution of the observed mean sentence length *y* (in number of words) and
- parallel expected mean values *Y*.

TABLE 3.1.1-10 : Distribution of sentence aggregates in Turkish texts

3.1.1 *Text 1*

x	z	v	y	Y
1	273	4005	14.67	14.90
2	83	2435	14.67	14.44
3	39	1697	14.50	14.17
4	14	834	15.05	13.98
5	17	1066	12.54	13.84
6	5	427	14.23	13.67
7	7	806	16.45	13.63
8	4	324	10.13	13.55
9	2	232	12.89	13.47
10	2	225	11.25	13.41
12	1	182	15.17	13.30
13	1	222	17.08	13.25
17	2	426	12.53	13.08
19	1	234	12.32	13.02
Σ	451	-	-	-

A = 14.9028; b = - 0.0459

Wilcoxon T = 52.5 > $T_{0.05}(14)$ = 21. The distributions of the observed *y* and expected *Y* do not significantly differ. *z* is the number of aggregates having length *x*, *n* is the total length of *z* aggregates in number of words.

3.1.2 *Text 2*

x	z	n	y	Y
1	364	3795	10.43	10.63
2	116	2204	9.50	10.07
3	55	1680	10.18	9.76
4	29	1212	10.45	9.55
5	12	610	10.17	9.38
6	10	476	7.93	9.25
7	10	723	10.33	9.14
8	1	57	6.33	9.05
9	4	359	9.97	8.97
10	4	313	7.83	8.89
11	3	315	9.55	8.83
12	1	133	11.08	8.77
13	2	219	8.42	8.71
15	1	152	10.13	8.62
18	2	351	9.75	8.50
22	2	272	6.18	8.37
Σ	616	-	-	-

A = 10.6266; b = - 0.0773

Wilcoxon T = 58 > $T_{0.05}(16)$ = 30. The distributions of *y* and *Y* do not significantly differ.

3.1.3 *Text 3*

x	z	n	y	Y
1	192	1239	6.45	6.96
2	69	933	6.76	6.67
3	31	599	6.44	6.50
4	14	341	6.09	6.39
5	13	340	6.07	6.30
6	7	289	6.88	6.23
7	3	145	6.90	6.17
8	6	320	6.67	6.12
9	1	57	6.33	6.07
10	2	135	6.75	6.03
11	2	126	5.73	6.00
15	1	70	4.67	5.88
31	1	176	5.68	5.62
32	1	174	5.44	5.61
Σ	343	-	-	-

A = 6.9647; b = - 0.0625.

Wilcoxon T = 35 > $T_{0.05}(14)$ = 21. There is no significant difference between the distribution of *y* and *Y*.

3.1.4 *Text 4*

x	z	n	y	Y
1	150	763	5.09	6.35
2	44	434	4.93	5.44
3	16	264	5.50	4.98
4	13	274	5.27	4.67
5	8	169	4.23	4.44
6	5	174	5.80	4.26
7	3	91	4.33	4.12
8	4	133	4.16	4.00
9	2	91	5.06	3.90
10	1	42	4.20	3.81
11	2	83	3.77	3.73
12	2	68	2.83	3.66
14	1	56	4.00	3.53
15	1	42	2.80	3.48
18	1	53	2.94	3.34
19	2	110	2.89	3.30
23	1	72	3.13	3.16
Σ	256	-	-	-

A = 6.3522; b = - 0.2224.

Wilcoxon T = 60 > $T_{0.05}(17)$ = 35. There is no significant difference between the distributions of *y* and *Y*.

3.1.5 *Text 5*

x	z	n	y	Y
1	260	6999	26.92	26.60
2	76	3469	22.82	24.75
3	41	2629	21.37	23.73
4	20	2014	25.18	23.03
5	12	1898	31.63	22.50
6	7	909	21.64	22.07
7	5	657	18.77	21.72
8	2	286	17.88	21.42
9	4	769	21.36	21.16
10	2	466	23.30	20.93
11	1	248	22.55	20.72
12	1	253	21.08	20.53
13	3	812	20.82	20.36
14	1	268	19.14	20.21
15	2	562	18.73	20.06
18	1	331	18.39	19.68
19	3	1075	18.86	19.57
21	1	398	18.95	19.37
28	1	554	19.79	18.80
29	1	565	19.48	18.73
Σ	444	-	-	-

A = 26.6048; b = - 0.1042.

Wilcoxon T = 96 > $T_{0.05}(20)$ = 52. There is no significant difference between the distributions of *y* and *Y*.

3.1.6 *Text 6*

x	z	n	y	Y
1	323	3998	12.38	12.45
2	99	2428	12.26	12.20
3	44	1504	11.39	12.06
4	17	788	11.59	11.97
5	16	861	10.76	11.89
6	12	852	11.83	11.83
7	7	611	12.47	11.78
8	6	599	12.48	11.73
9	6	667	12.35	11.69
10	4	476	11.90	11.66
11	2	315	14.32	11.63
13	1	150	11.54	11.57
14	1	186	13.29	11.55
15	1	157	10.47	11.53
17	1	166	9.76	11.48
19	1	218	11.47	11.45
21	1	234	11.14	11.42
46	1	481	10.46	11.16
49	1	508	10.37	11.14
54	1	654	12.11	11.11
Σ	545	-	-	-

A = 12.4459; b = - 0.0284.

Wilcoxon T = 91 > $T_{0.05}(19)$ = 46. There is no significant difference between the distributions of *y* and *Y*.

3.1.7 *Text 7*

x	z	n	y	Y
1	171	1826	10.68	12.38
2	57	1238	10.86	11.31
3	28	1008	12.00	10.73
4	11	566	12.86	10.34
5	5	257	10.28	10.04
6	5	321	10.70	9.81
8	3	249	10.38	9.45
9	1	92	10.22	9.30
11	1	85	7.73	9.07
13	2	153	5.88	8.87
30	1	272	9.07	7.96
Σ	285	-	-	-

A = 12.3777; b = - 0.1299.

Wilcoxon T = 30 > $T_{0.05}(11)$ = 30. There is no significant difference between the distributions of *y* and *Y*.

3.1.8 *Text 8*

x	z	n	y	Y
1	148	2738	18.50	20.09
2	40	1485	18.56	19.77
3	17	967	18.96	19.59
4	11	908	20.64	19.46
5	1	135	27.00	19.36
6	3	314	17.44	19.28
7	2	281	20.07	19.21
8	4	670	20.94	19.15
9	4	665	18.47	19.10
10	1	189	18.90	19.06
14	1	220	15.71	18.91
15	1	262	17.47	18.88
16	1	320	20.00	18.85
Σ	234	-	-	-

A = 20.0903; b = - 0.0230.

Wilcoxon T = 38 > $T_{0.05}(13)$ = 17. There is no significant difference between the distributions of *y* and *Y*.

3.1.9 *Text 9*

x	z	n	y	Y
1	190	2980	15.68	15.72
2	39	1210	15.51	15.68
3	15	618	13.73	15.66
4	11	760	17.27	15.65
5	7	517	14.77	15.63
6	1	128	21.33	15.62
7	2	232	16.57	15.61
8	1	101	12.63	15.61
9	1	142	15.78	15.60
10	1	152	15.20	15.59
11	1	174	15.82	15.59
12	1	164	13.67	15.58
14	1	235	16.79	15.58
Σ	271	-	-	-

A = 15.7224; b = - 0.0035.

Wilcoxon T = 44 > $T_{0.05}(13)$ = 17. There is no significant difference between the distributions of *y* and *Y*.

For x = {1 - 5}: A = 15.4709; b = - 0.0083.

3.1.10 *Text 10*

x	z	n	y	Y
1	287	4091	14.25	13.94
2	65	1840	14.15	13.73
3	32	1208	12.58	13.61
4	20	1062	13.28	13.52
5	14	884	12.63	13.45
6	9	805	14.91	13.40
7	4	421	15.04	
8	6	782	16.29	
9	3	412	15.26	
10	2	307	15.35	
12	1	144	12.00	
14	1	204	14.57	
19	1	272	14.32	
23	1	363	15.78	
33	1	478	14.48	
Σ	447	-	-	-

A = 13.9377; b = - 0.0220.

Wilcoxon T = 10 > $T_{0.05}(6)$ = 0. There is no significant difference between the distributions of *y* and *Y*.

For each text the appropriate values of parameters *A* and *b* are introduced. The testing criterion *T* of the Wilcoxon test is simultaneously introduced with its results. The Wilcoxon test was applied here for comparison of the distributions of *y* and *Y*. This test is currently used for testing two mutually dependent samples ordered in pairs; see R. Reisenauer (1970, 103-107). This test of paired data can also be used in approvals of instruments through the paired comparison of the observed data with those obtained from the respective patterning instrument, see V. Kubánková & J. Hendl (1986). The latter type of data in our testing is represented by *Y*. This nonparametric test is based on differences $d_i = y_i - Y_i$ and the evaluation of their ranks.[4] In each text of our Corpus the applications of the Wilcoxon test result in insignificant differences between the distributions of *y* and *Y*.

3.4. THE STOCHASTIC CHARACTER OF LANGUAGE CONSTRUCTS

In the Tables presented above, the mean sentence-length distributions corresponding to the observed values of *x* are displayed in order to demonstrate that, regardless of the non-stabilized course of the variable *y* occurring at the higher values of *x* in texts which are not too long, the MA law holds. The fluctuations of the variable *y* in the texts which are not very large do not affect the validity of the law.

Text 10, analyzed in Table 3.1.10, represents an exception; its parameter *b*, when the total distribution of mean sentence length is tested, is not negative, as is required by the MA law. But when the extreme values of this distribution are excluded and only the values of *y* corresponding to $1 \leq x \leq 6$ are tested, as is demonstrated in Table 3.1.10, we see that the tendency predicted by the MA law is present in the distribution of the mean sentence length in aggregates of this text.

It must be stressed that the phenomenon described, like all language phenomena, is of a stochastic character. No scientific and theoretical approach to language can ignore this probability aspect.

[4] Here and elsewhere in the present work the Wilcoxon *T* is used for testing the agreement of the two respective distributions, i.e. of the values of the observed and expected variable. On the other hand, when the coefficient of determination *D* is used, its sense is different: it indicates the degree to which a pertinent principle or law explains the variation of the observed values. When the coefficient of determination is applied to the values of *y* and *Y* in Tables 3.1.1-10, the result is not interesting, as the worst result equals *D* = 0.9991.

What is understood under stochasticity of natural languages? Language, especially when supposed together with the semantic systems carried by human individuals, is lacking in certainty as it cannot be predicted in which state the entire language system will be found at a given moment. Occurrences of units as words, sentences, texts are in fact the results of a choice which is influenced by an immense number of impacts. Each act of selection can be considered as an experiment which cannot always be repeated with the same predicted outcome. This is true especially when we take into account the evolution of the language system over time. Certain properties of language constructs can be expected with a greater probability than others; however, nothing can be predicted with certitude. We cannot know positively which of the possible language constructs will be produced by a given speaker (producer) at a given time interval. This is partly due to the complexity of the language systems and partly to the non-identical circumstances in which any text is produced. Languages are universal means of communication, they function in all imaginable situations in which the human race and its individuals occur. This creates complicated but always relevant conditions for a language formation.

The other reason for taking the probabilistic character of language systems into account consists in a fact that represents a universal epistemological principle: The entire reality discovered with the help of sciences is determined by the laws of probability. This has already been recognized by all actual sciences including quantitative linguistics.

Let us return to aggregates. They were also observed in German texts; see C. Schwarz (1992). In her work, this author stated that from a large set of the theoretical distributions, deduced beforehand, of length x of aggregates, the nearest one to the observed z is the Waring-Herdan distribution. This problem will be touched upon here again in connection with the distribution of word associations.

3.5. OBTAINING DATA FROM TURKISH TEXTS

Each presentation of data comprises a measurement expressed in frequencies. For this reason a more detailed specification of the units observed is unavoidable.

As for the aggregates in the Turkish texts, at first sentences were observed. It is usual to comprehend a sentence as that part of text which contains predication. This, however, is not

quite clear. In the case of finite verbs there are no problems, and a sentence can be delimited without doubts. In certain styles, especially in the scientific functional style, it is usual to formulate uncommonly long Turkish sentences with the help of different infinitival verbal forms. The finite verb then comes at very long intervals. In this case it is evident that also in clauses with infinitive, short infinitive, verbal noun *-DIk*, or gerundial forms at their heads some sort of predication is present. This is the problem of the classification of clauses and sentences.

In the analyses of Turkish texts presented above we maintain the latter possibility, i.e. we also classify as sentences the constructions headed by infinitival forms. Thus smoother distributions of data are obtained without extraordinary fluctuations. This of course is not a rational argument. With several texts we made an experiment resulting as in the case which is presented in TABLE 3.2. The same text as in Table 3.1.6 (i.e. Text 6) was analyzed with "larger sentences", which means with sentences strictly defined by verbs in finite forms. The results indicate that the kind of predication, meaning the classification of sentences either according to verb in finite or in infinitival forms, can be viewed as irrelevant from the standpoint of the MA law applied to aggregates. Where the infinitival forms of verb constituted a combined verb with the finite form which followed them immediately, they were classified as one sentence headed by a finite verb supplemented by a verbal attribute. Unfortunately, our argumentation based on Table 3.2 is purely empirical and thus far from being conclusive. This is the way in which the length of aggregates in number of sentences has been actually counted in the present work.

Sentence aggregates are constructs combining the environments of word units together into one item. Each environment headed by a verb, regardless of whether it contained a finite or infinitival form, functions as a constituent of these constructs in Turkish. The label 'sentence' assigned to these units is not beyond dispute, but the problem of which label should be used is not substantial. See also Section 1.4.

Words - or better, word forms - are often counted by quantitative linguists as continuous constructs occurring between two spaces in written texts. This is more comfortable, though not always very convenient, for statistical operations. In the analyses presented in different chapters of the present work, *word* or *word form* was considered to be the part of text between two spaces plus the indefinite article *bir*, plus postpositions which are not combined

with its head-word by the Genitive Construction, and plus the enclitics *da, de, ta,* and *te*. These units were not counted as individual words. For example, *adam gibi* ("like a man") was counted as one word and *ev(in) önünde* (in front of the house") as two words. The expression *bir adam* ("a man") was counted as one word, and *bir adam* ("one man") was counted as two words. Combined numerals were divided into their constituent numerals; thus, e.g. 1961 has been counted as five words (= *bin dokuz yüz altmış bir).*

The selection of Turkish texts for our corpus was partly designed and partly made at random. Texts were selected for the purpose of analyses with certain limitations, namely, no text should contain formulas, tables, figures and other elements corrupting its continuity. Otherwise, they were selected in a way that can be supposed to be random. It must be stressed that the texts analyzed here do not represent a statistical sample in the strict sense of the word. The MA law should be valid (and really *is* valid) for each arbitrarily selected text. It must be also stressed that in similar statistical analyses an entire text cannot be substituted by some number of sentences sampled at random or by randomly selected pages. Our interest is concerned with ties mutually joining together different language constituents on different levels. The Turkish texts analyzed are not elements of a random statistical sample, and they do not represent some specificity of Turkish texts in comparison with texts in other languages. Each text of each language should corroborate the generality of the MA law.

TABLE 3.2: Aggregates based on sentences headed by finite verbs only

Text 6

x	z	n	y	Y
1	331	8048	24.31	24.36
2	95	4500	23.68	23.38
3	44	2988	22.64	22.83
4	18	1710	23.75	22.44
5	16	1457	18.21	22.15
6	13	1640	21.03	21.91
7	4	675	24.11	21.71
8	6	1093	22.77	21.54
9	5	1017	22.60	21.39
10	4	888	22.20	21.26
11	1	251	22.82	21.14
14	2	619	22.11	20.84
16	1	297	18.56	20.68
18	1	356	19.77	20.53
19	1	337	17.84	20.47
36	1	659	18.31	19.71
37	2	1677	22.66	19.68
Σ	545	-	-	-

A = 24.3604; b = - 0.0591.

Wilcoxon T = 68 > $T_{0.05}(17)$ = 35. There is no significant difference between the distributions of *y* and *Y*. See also Table 3.1.6.

3.6. INTERPRETATION OF LEXICAL UNITS

The semantic interpretation of lexical units of a text means a certain transformation of the list of its lexical units. In this operation different lexical units are understood as having identical lexical meaning; on the other hand, in the same interpretation two or more occurrences of the same lexical unit can be interpreted as occurrences of different units. The solution is decided by each individual interpreter, producer or recipient of a text. For example, the producer or recipient of a text may interpret the units "to come" and "to enter" at certain locations of an interpreted text as semantically identical; then the number of lexical units v is one unit lower. On the other hand, the language user may qualify "to come" in 'John came' and 'winter came' as two different units; and the value of v is then one higher. Two interpreters may not agree with each other in their interpretations of a given text.

It is evident that there is no authority enabling one to make an interpretive decision which could be treated as absolutely true. This is not in contradiction with the work of lexicographers who indicate the possible identities in meanings of lexical units. Both the language and situational context represent more delicate devices making decisions concerning semantic identities between lexical units in one text. Lexicographers suppose all real and potential occurrences of a unit in real and potential contexts. Their decisions resemble a juridical law for users of a language. Their work is very useful because it helps to stabilize a given language as a communication means, but in contrast to juridical laws linguistic prescriptions are not strictly obligatory. The activity of language users is based on quite a different model of functioning and the problems of human individuals in different life situations cannot be conjectured in advance by any juridical law. On the other hand, language is always ready to do it.

This is not far from the way of thinking incorporated in the works cited by P. Menzerath, who understands the principles of the MA law as having some psychological background. P. Menzerath (1954, 100) writes:

> 'Es tritt eine **'Sparsamkeitsregel'** in Erscheinung, die sich psychologisch auf eine Ganzheitsregel dieser Art gründet: **je größer das Ganze, um so kleiner die Teile!** Diese Regel (...) wird aus der Tatsache verständlich, daß das Ganze jeweils 'übersehbar' bleiben muß.' (Quoted from Fenk & Fenk-Oczlon 1993,11.)

Language laws, however, formulated as stochastic expressions, enclose this psychological background. It must be stressed that the psychology of language users and the laws sought and uncovered in language and tested on texts are not objects standing in mutual contradiction. Semantic interpretation, as described above, is a process observable in language constructs. This can be documented by the following analysis:

From a poem by the famous Old Ottoman mystical poet *Yunus Emre*, the data concerning the observed aggregates, presented in TABLE 3.3, were obtained. The positive value of b indicates that in this short text the MA law is neglected. This is not acceptable. When the same poem is interpreted and its aggregates analyzed again, the data presented in TABLE 3.4 are obtained. The analysis based on the interpreted units results in b having a negative value, as is predicted by the MA law.

TABLE 3.3: Aggregates considered *without* a semantic interpretation of lexical units observed in a poem by Yunus Emre

x	z	n	y	Y
1	114	409	3.59	3.70
2	21	154	3.66	3.81
3	4	46	3.83	3.88
4	2	31	3.88	3.92
5	2	47	4.70	3.96
15	1	60	4.00	4.15
17	1	68	4.00	4.17
Σ	145	-	-	-

A = 3.697; b = 0.0427.
Wilcoxon T =2 7 > $T_{0.05}(7)$ = 2. There is no significant difference between the distributions of y and Y.

Text: A poem by Yunus Emre, cf. LXXXI in: Abdüllâh Gölpınarlı (ed.), *Yunus Emre, Risâlat al-Nushiyya ve Dîvân,* Garan, Istanbul 1965, p. 81.

TABLE 3.4: Aggregates based on the semantically interpreted lexical units in a poem by Yunus Emre (see Tab. 3.3)

x	z	n	y	Y
1	111	395	3.56	3.73
2	25	165	3.30	3.62
3	3	38	4.22	3.56
4	1	20	4.00	3.51
5	1	16	3.20	3.48
17	1	49	2.88	3.30
47	1	157	3.34	3.16
Σ	143	-	-	-

A = 3.73; b = - 0.04298
Wilcoxon T = 13 > $T_{0.05}(7) = 2$. There is no significant difference between the distributions of *y* and *Y*.

Text: as in Tab. 3.3

This indicates that

- the law and its consequences are also hidden in shorter texts, though it is difficult to reveal and test them;
- the non-interpreted text appears to be a result of linguistic abstraction hiding in itself certain effects of the real semantic structure of text.

What is meant by 'linguistic abstraction'? Communicative activity in a natural language represents in fact a continuous flow of signs (including gaps between signs also functioning as signs). We are not competent to say how the respective parts of the human brain act in producing and receiving such a flow of language signs serving for communication. Activities associated with writing are based on a particular selection or classification of units which always is an act of abstraction. This concerns the units on all evident levels (phonemes and letters, words and phrases). Philology and linguistics systematized this classifying endeavour. Words and lexical units are entities produced by such classifications. Each word bears something that once was called 'semantic field'. This is an indication of the generally

accepted conviction that lexical meaning is not a precisely located point in the semantic system, but something vague that is attributed by indefinite limits. With the help of semantic interpretation, understood as an act of text analysis occurring whenever text is produced and received, this vague semantic field is indicated without any abstraction. If linguistics wants to make theoretical progress, it should seek ways to remove simplifications brought about by too brusque abstractions.

Another consequence of semantic interpretation concerns the status of text. This entity is traditionally treated as a firmly fixed item. Textological, juridical and literary convention always tends to fixed texts (in critical editions, in texts of laws supplied by commentaries) with the intention of obtaining a petrified expression and unified understanding. The same is typical for sciences. Linguistic theory requires however another approach, enabling us to build up explanatory theories. Semantic systems proper to language users are activated in the production and reception of texts; they appear to be parts or peripheries of the system represented by the human mind on the one hand, and of the language system communicated from mind to mind by each text on the other hand.

Language constructs produced by this periphery in the form of text are evidently an output of the system. Text, however, when separated from its producer or receiver, loses something substantial and this lost property is nothing other than its interpretation. Two receivers of the same (non-interpreted) text construct need not receive the same interpreted text. A great part of mankind's intellectual activity is directed to discovering the way to force language users to speak (write) and understand texts uniformly. This endeavour inevitably meets with failure but it cannot be left aside. The character of the communicated reality and also the character of the communication means (which is the semantic system having its biological basis in brains) cannot allow it. These considerations can be summarized in the following assertion:

> Text is a language construct beginning and ending the interval of its existence only in connection with the act of semantic interpretation by a language user (text producer/receiver).

This assertion rather resembles a philosophical position in relation to text. It is valid, however, in any theoretical linguistic context which tries to find the scientific laws of language. The applicational branches of linguistics (e.g. philology) will continue to state the rules of correct

language expressions etc., because this is necessary for making the communication means more stable. For certain purposes of theoretical linguistics this assertion may also appear to be redundant if the consequences of radical abstractions do not spoil theoretical statements based on a refutable and tested theory. When, however, the problems comprising meanings and semantic systems are solved, then one must take into account the necessity of the real processes which are called 'semantic interpretation' here. This is testified too by the analysis of aggregates and applications of the principle formulated by the MA law.

3.7. ALTMANN'S PARAMETER *b*

The function *y(x)* defined by the MA law implies the relation of two different units:

- the units identical with constituents in which the length of the constructs is measured;
- the units in which the length of constituents is expressed.

In the case of aggregates, constituents are sentences and the length of aggregates is given in number of sentences; the length of sentences is measured in number of words.

Suppose the case where the number of sentences in an aggregate is expressed by the same value as the number of words in its mean sentence. Aggregate is a construct (in relation to sentences) and also sentence is a construct (in relation to words). The following supposition is equivalent to the delivering of observational concepts preceding the general formulation of the MA law, see formula (2.1) in Section 2.1. And now we suppose that both these quantities are theoretical constructs with $x = c$ and $y = c$, where

$$c = A\,c^{b}, \qquad x = y = c. \tag{3.1}$$

and

$$c = A^{\frac{1}{1-b}}. \tag{3.2}$$

The observed values of *c* computed according to (3.2) from our Corpus of texts are presented in TABLE 3.5. When they are compared with the means E(y) and E(Y), it becomes evident that *c* actually equals these means. The means E(y) and E(Y) are calculated from the values of *y* and *Y* presented in Tables 3.1.1 -10. Thus it can be asserted that these means which are

equal to one another can be treated as functions of *A* with *b* as the coefficient of proportionality. This is evident from the logarithmic transformation of (3.2):

$$\log c = \frac{1}{1-b} \log A \tag{3.3}$$

Let us remember that the value of parameter *A* equals the value of *y* corresponding to $x = 1$, so that $A = y_1$. The mean of *y* can be expressed as follows:

$$E(y) = \frac{1}{r}(y_1 + y_2 + \ldots + y_r) = \frac{1}{r}\sum_i y_i$$

With respect to the MA law, the sum of y_i can be written:

$$\sum_i y_i = Ax_1^b + Ax_2^b + \ldots + Ax_r^b = A\sum_i x_i^b \tag{3.4}$$

With respect to the observed equality of *c* and E(y), we can write:

$$A^{\frac{1}{1-b}} = E(y) \tag{3.5}$$

There is a direct proportionality relation between the logarithm of the mean length of the constituents of aggregates in Turkish texts and log *A*. The expression 1/(1 - b) functions as a coefficient of proportionality of this relation.

Let us recapitulate this discussion of parameters and means: The curious condition $x = y = c$ contained in (3.1) is approved by the properties of the mean constituent in the light of the MA law. Suppose a volume (= a basket) containing a certain number of units (= apples) and the same volume containing another number of other units (= pears). Under certain circumstances the number of apples by which the volume is characterized equals the mean size of pears that got into the same volume. What are these circumstances? If the volume is an "apple-basket" called language construct, these circumstances are described by the MA equation. Then we can ask what the properties of the parameters contained in this equation are. Among others, they are described by (3.5): a certain relation of these parameters can be explained as the mean value of the constituent. As for Turkish, all these relations are proved by observations resulting in a high similarity between *c* and *E(y)*, which represents in fact equality of these quantities.

TABLE 3.5: The constant *c* in texts analyzed in Tables 3.1.1-10

Text	A	b	c	E(y)	E(Y)
1	14.90	-0.045929946	13.23	13.82	13.67
2	10.63	-0.077346014	8.97	9.26	9.16
3	6.96	-0.062488108	6.21	6.20	6.18
4	6.35	-0.222382345	4.54	4.17	4.13
5	26.60	-0.104227684	19.52	21.43	21.30
6	12.45	-0.028384608	11.61	11.72	11.67
7	12.38	-0.129869846	9.27	10.06	9.93
8	20.09	-0.022968391	18.78	19.44	19.29
9	15.72	-0.003534831	15.57	15.57	15.62
10	13.94	-0.021983934	13.17	13.63	13.6

TABLE 3.6: The observed values of b

Corpus 1	Corpus 2
-0.04593	-0.08997
-0.07735	-0.04503
-0.06249	-0.05359
-0.22238	-0.09530
-0.10423	-0.08452
-0.02838	-0.04276
-0.12987	-0.04298
-0.02297	-0.18124
-0.00353	-0.07369
-0.02198	-0.08038
	-0.08625
E(*b*): -0.07191	-0.07187

Another way of characterizing Altmann's *b* in Turkish texts can be obtained by observation of its values. These values were observed in two corpora of Turkish texts. One is analyzed in the present volume (Corpus 1; see the list at the end of the book); the other one was analyzed in Hřebíček (1992, 70-72; Corpus 2). The observed values of *b* are shown in table 3.6. Regardless of the variation of these values, their means are almost equal.

3.8 AN ESTIMATE OF *b* FROM THE OBSERVED TEXT VARIABLES

If we try to find a linguistic sense for this parameter, i.e. if we try to characterize it by concepts close to expressions occurring in linguistic descriptions, our reasoning can proceed in the following way:

From the formulation of the MA law in (2.2), the following sequence can be derived:

$$y_1 = A\ 1^b$$
$$y_2 = A\ 2^b$$
$$y_3 = A\ 3^b$$
$$\dots$$
$$y_r = A\ r^b$$

Let us suppose a text having *k* sentences. Then the ultimate possible member of this sequence is:

$$y_k = A\ k^b \tag{3.6}$$

The value of y_k is the mean sentence length in an aggregate containing all *k* sentences of the text, i.e. the whole text. This value is at the same time the mean sentence length of the whole text. When the whole text has the length *n* (in number of word forms) and *k* (in number of sentences), the mean sentence length is *n/k*. From (3.9) it follows that

$$\frac{n}{k} = A\ k^b \tag{3.7}$$

and

(In this formula the base *e* of the natural logarithms is used and thus *ln* is written instead of *log*. The reason will be evident from the interpretation of the formula which soon follows.)

$$b = \frac{\ln n - \ln A}{\ln k} - 1. \qquad (3.8)$$

A set of observed values of b was obtained from a sample of Turkish texts. These values were applied to the respective aggregates. It turned out that formula (3.8) can be utilized as an estimate of this parameter of the MA law for aggregates; see also Hřebíček (1992, 77-91). This is quite important as this procedure enables us to estimate b from the basic text characteristics n and k. Of course, A must also be known, as b and A are always firmly linked together; it was indicated that A nails down the initial point of the new (non-explicitly defined) coordinates in which the system actually is defined. As far as aggregates are concerned, A symbolizes the number of *hapax legomena* of the text supposed .

3.9. THE ENTROPY OF AGGREGATES

Formula (3.8) can also be interpreted as an expression with a certain communicative sense. To present this idea let us proceed in the following way:

Any text at the moment of its semantic interpretation (by producer or recipient) can be in an auxiliary way treated as an isolated system. At a given moment the language user and text can be assumed to be a unity without any connection to the environmental reality. The level of organization of this isolated system can be characterized with the help of the concept of entropy. If the assumption of isolation is correct, the entropy S of this system can be expressed by Boltzmann's famous formula.

Each lexical unit of a text defines one aggregate of the same text. With the approximation mentioned in connection with the definition of aggregates above, it can be supposed that each occurrence of this lexical unit represents one sentence in this aggregate, i.e. one of its constituents. The entire number of word occurrences is n. The entire number of word occurrences thus represents n possible states of the respective text. The entropy of the system rendered as an isolated system of aggregates and their constituents can be expressed as

$$S = K_B \ln n,$$

where K_B = Boltzmann's universal constant.

Note the possibility that in the universe of the semantic relations another constant is also valid. An supposition analogous to those concerning n states of a text can be stated in connection with the system of k sentences of the same text. Thus we obtain:

$$S = K_B \ln k$$

As both these expressions concern the same text, their entropies must be equal, as the level of organization of both is compacted together by the relations of the respective levels. Therefore we can risk writing an equation in which Boltzmann's universal constant on the both sides is reduced. On the contrary, this equation can be completed with a coefficient of proportionality, say b, and we obtain:

$$b \ln k = \ln n \qquad (3.9)$$

This equation requires two corrections taking into account certain specific properties of text constructs:

- n different states of the system include the grammatical relations which are already included in the assumption concerning the states of k sentences because sentences are grammatical items. Therefore we must take the value $ln\ k$ from $ln\ n$;
- the value of A denotes the number of aggregates corresponding to the constructs with $x = 1$. Formally, it has been derived by Altmann as the differential constant $e^C = A$; this constant moves the curve of the distribution upwards by a distance equalling the value of A. Such aggregates which have their constituents of length $x = 1$ are always present in the system; this means that they are subsumed under n states, but not under k states. Therefore the number of states $ln\ k$ must also be increased by the value $ln\ A$.

Thus from (3.9) the following relation is obtained:

$$b \ln k + \ln A = \ln n - \ln k$$

from which

$$b = \frac{\ln n - \ln A}{\ln k} - 1,$$

And this is formula (3.8) which was derived from other assumptions.

Naturally, the way in which language structures are organized into systems always provides communication properties. Let the term 'communication' be understood correctly. Communication in the codes carried by natural languages is often comprehended as being interpretable in new texts. These new texts can be further interpreted by some new texts, and so on to infinity. This is an act of semantic interpretation but not some explanatory description of information contained in a text. This approach is sufficient outside science, in everyday life. At the same time, the traditional approach to semantics based on semantic interpretations cannot explain the relationship between the way in which the language system is organized and the semantic specificity of this system. Exactly this connection between the way of organization and the ability of a system to bear information represents the essence of the mathematical theory of communication. In linguistics, this theory deserves greater attention. It is not an empty formalism staying in contrast to the intuitive evaluation of semantic constructs. This theory *is* semantically relevant as it is also testified by the interpretation presented above of formula (3.8) in terms of communication. The system of text aggregates is formed from units semantically pertinent and semantically interpreted in text production and reception. Semantics appears to be a way in which certain units are mutually organized in texts.

Let us illustrate briefly how different values of b alter the shape of the curve y, i.e. the MA curve. In FIGURE 3.1 different values of b are applied to $x = \{1 - 10\}$ with fixed value of $A = 15$. This Figure indicates that all curves y are symmetrical with respect to the point [(x = 1); A]. The curves with negative b descend from this point down and asymptotically approach the coordinate x. The curves with positive b ascend without limit. The higher the absolute value of b, the closer the curve y is to the line parallel to axis y at the distance $x = 1$.

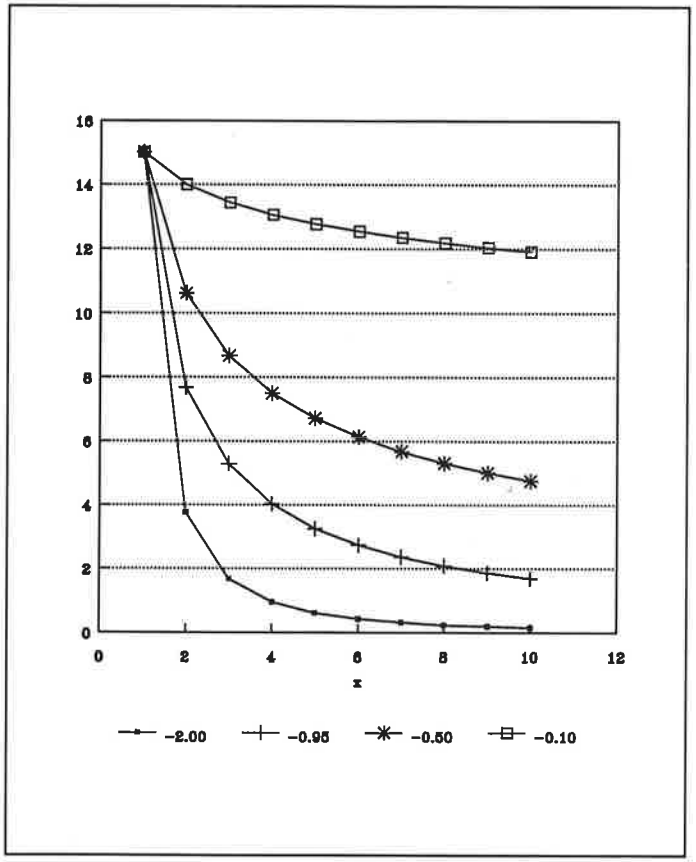

FIGURE 3.1: The shape of the MA curves with
b = {-2.00; -0.95; -0.50; -0.01}

3.10. THE MA LAW AS A RESULT OF CHOICE

The system of aggregates is a result of complicated processes, each process being a specific choice. We mention choice as another point of view from which this system can be characterized. We may pose the question of how many possibilities are at hand for a language user when producing or interpreting a text; how large is the space in which the language user can move freely without breaking the MA law, and how large is the opposite space where this law is inevitably offended.

Let us start with the assumption of an arbitrary text for which a language user needs a number of lexical units $1 \leq u \leq v$. Each of these units is equipped with frequency f_u. When the supposed text is sufficiently large, it happens to be obvious that the frequencies of individual lexical units can be arranged into a sequence in the approximate shape as in FIGURE 3.2 (the values of f_u, or simply f, are arbitrarily selected in this Figure).

Further, let us imagine that the language user takes the units from a certain lexical column u of Figure 3.2 and places its items (= word occurrences) into different sentences. Let us remember our approximate assumption that into each sentence no more than one item of a unit u is placed. A certain unit u is thus placed in a set of sentences forming an aggregate with the mean sentence length

$$y_u = \frac{n_u}{f_u}, \qquad (3.10)$$

where n_u = the total length of an aggregate in number of words,

f_u = the frequency of lexical unit u.

Now let us suppose two arbitrary lexical units with frequencies $f_u = \{w_1, w_2\}$ fulfilling the condition $w_1 > w_2$. Their aggregates are of total length n_1 and n_2 in number of words. For $y_1 = n_1/w_1$ and $y_2 = n_2/w_2$ three different cases can occur:

1. $n_1 = n_2$; then $y_1 < y_2$;
2. $n_1 < n_2$; then $y_1 < y_2$;
3. $n_1 > n_2$; then $y_1 \; ? \; y_2$.

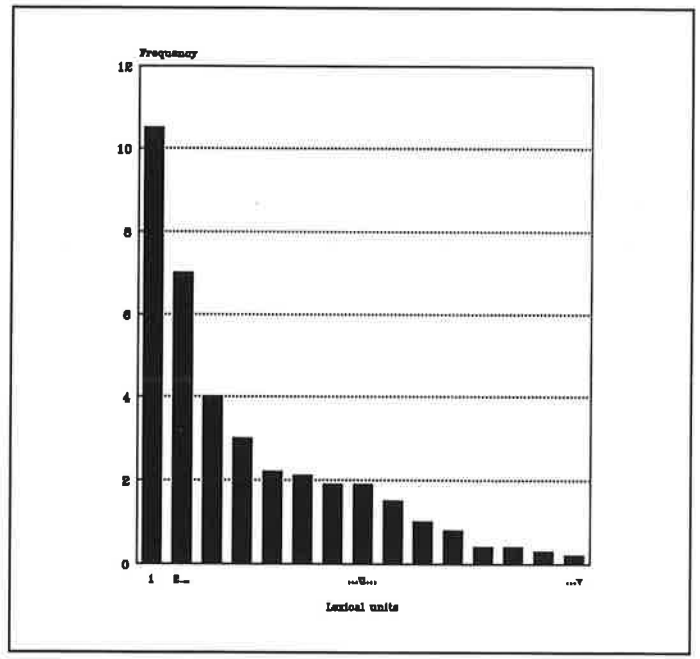

FIGURE 3.2: Frequencies of v lexical units in a text

In the first two cases the MA law holds, as is indicated by the mean sentence y_1 and y_2 of the respective aggregates. As for the third case, it is impossible to state some general conditions for the validity of the MA law in the frame of the variables supposed. In this case the laws of probability are involved. If we survey the total situation of aggregates, it is immediately evident that the space for the unquestionable validity of the law (cases 1. and 2. above) is larger than that of the opposite case. Also a part of the third case belongs to the space of its validity, though we cannot say in advance how large that part is. This is, however, specified by the law, by its parameters. Thus we can understand why the law functions in the way in which it functions.

4. Semantics and Text

In the preceding chapter text was comprehended as a unity of a language construct and the mind of that language user who is in contact with this construct. Here we present a new variation to the problem presented above. Thanks to the user's act of semantic interpretation, the semantic system (= a part of the user's mind) represents a certain source of items forming text. Text then is a language phenomenon which may differ from user to user, though the construct is the same. This occurs also in the case when two or more users are in contact with one and the same texture. In order to make clear the difference between the (non-interpreted) text construct and (interpreted) text let us introduce for the former the term *texture*. It is defined as follows:

> Texture is a non-interpreted text construct which after some semantic interpretation changes into text. Each text production and reception is accompanied with interpretation. Texture thus represents a conceptual abstraction from text.

Arguments for these ideas are presented in Section 3.6. Now we try to present a discussion of the shape of semantic systems, i.e. of the items hidden in the minds of language users. Our interpretation of this shape is, of course, nothing but a consequence of the linguistic treatment of the issue.

4.1 LINGUISTICS AND THE HUMAN BRAIN

As far as the treatment of semantics is concerned, the position of linguistics is approximatey as follows: The semantic system is usually declared to be a part of the language structure, one

of the language levels. Thus the semantic system appears to be quite an obscure phenomenon. Meanings are intuitively attributed to language expressions and thus they are - through this operation of abstraction - identified with language constructs. Consequently, meanings are taken as being intruded into language entities. Meanings are sought somewhere in the part of the language carriers of meanings lying somewhere behind. Words and morphemes are usually taken as the simplest units of meaning as well as grammatical formants. Grammar is often understood as a sort of combinatorics formulated on these carriers of meanings. Grammar indicates the permitted combinations as well as those which are forbidden or compulsory under certain circumstances. Permissions, prohibitions and instructions are based on semantic intuition. The attempt of descriptivist schools at exclusion of semantics from the set of preliminary assumptions in linguistics indicates that the approach based on semantics is perceived as an inconvenient way for a scientific analysis. Descriptivism has formalized the approach to the semantics by introducing the *informant.* Thus, in the comprehension of descriptivism, texture is always interpreted by an informant.

Meanings are often used as a criterion for differing between correct and incorrect language patterns. Language then contains meanings and meaning is the basis for the choice of the language units and their arrangement. This treatment is absolutely correct and useful. All sciences as well as common cognition combine language meanings with language expressions in one unity. The modelling imagination presented in the present work differs from the generally accepted conception only in the location of the semantic system; i.e. of the reality designated by the concept of semantic system. The semantic system is a component of the human mind and this mind cannot be taken as an abstract phenomenon before it has been accepted that it is proper to human individuals. Then if text really is a semantically interpreted entity (and everybody knows that this is a fact), the text's carrier (i.e. texture) inevitably forms a unity with that part of the individual mind which is in contact with it. Meanings are not observable without their carriers of some kind.

Meanings stretch a space the geometry of which should be envisaged in a more complex manner. If the basic semantic units are outlined as points in a space, the everyday experience of language users makes us know that this semantic space is a galaxy which is supplied with an incessant motion.

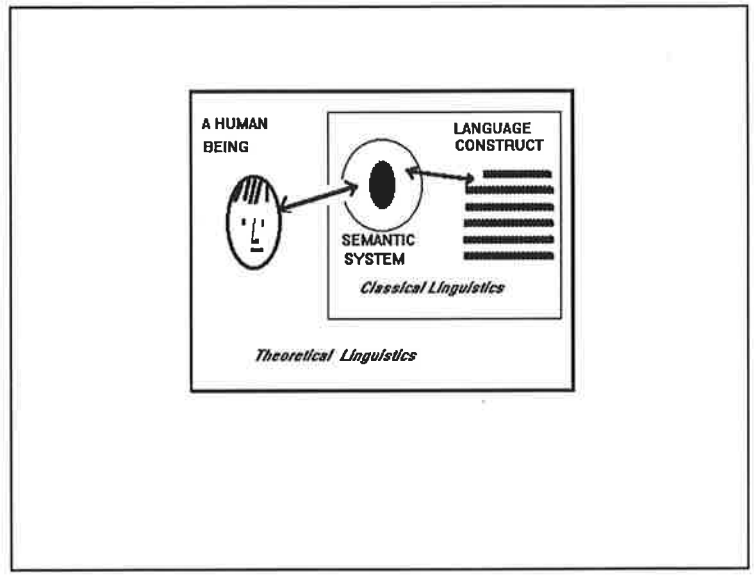

FIGURE 4.1: A classical and advanced linguistic theory

Its entities are in motion one around another. This motion is complicated and irregular. An arbitrary point in the semantic mind may change its position under the influence of all sorts of communicative acts. One type of these acts is language communication. The combinatorics of these points is many times more complex than that known from grammar and than that testified by the combinatorics of the abstract lexical meanings.

This conception includes three entities and their interrelations:

1. a human being (this time taken as a biological object);
2. a system of meanings;
3. a language code supposed as a pure output of a system of rules producing all possible sentences including the meaningless ones and all possible textures.

FIGURE 4.1 indicates what is investigated by classical linguistics and what must be taken into account when one wants to build up an advanced linguistic theory. Semantic system,

however, is a very complicated formation, and its structure cannot be gathered similarly to some syntactic typology in a satisfactory way.

It is evident that classical linguistics, which is so often limited by its normative and pedagogical aims, understands a human being as an abstract entity located somewhere out of its concern. This foreshadows the possible future enlargement of interests in theoretical linguistics. And we are assured that our abstract concept of the semantic system also will appear as an abstraction limiting the future development of theoretical conceptions.

The range of the usual explanations concerning the semantic system in linguistics is represented by units having mutual relations simplified through lexical and grammatical abstractions. 'Abstraction' here means the usage of the concepts introduced into the system supposed by scholars preoccupied with the prescriptions for language users. Let it be stressed that this approach to meanings is also useful as it is close to some common understanding of meanings. However, the conception presented here is not in contradiction to these classical approaches. It only tries to broaden the limits based on too early abstractions.

On the other hand, if the semantic system is taken as a structure existing in individual human minds, then biological phenomena are involved in this linguistic treatment. A philosophical background to this approach was formulated by Bunge (1983, 7):

> 'Hence biology is, or ought to be, interested in cognition, and all the cognitive sciences are, or ought to be, based on biology.'

The same author stresses that every cognitive act is a process in some nervous system and that cognition is a function of the nervous system. There is no doubt that the usage of language is in a way combined with cognition. The consequence of this approach is the requirement not to exclude in advance the possible relations between biological systems and language. Communication of meanings in language structures may be encroached upon by biological structures. Future linguistic explanations cannot bypass the obvious fact that the physiology of the brain is involved in language problems. In the following sections certain arguments for this position are discussed.

Regardless of the location of individual and abstract semantic systems, the problem remains of how to model it on the basis of our knowledge based on language structures. To describe words as points in a space is a trivial means of description, but what about with such higher units as texts? What is some geometrical analogy of their shape?

4.2 THE SEMANTIC SYSTEM AND ITS ARRANGEMENT

In linguistic methodology procedures can be found which are characterized by a transfer of successful approaches from level to level, from one linguistic discipline to another one. Structural phonology may serve as an example. The success of its analogies on higher levels is completely understandable because this approach is evidently founded on the self-regulative principle of language systems.

The success of structural phonology consists of relating phones to meanings by taking their word context into consideration. This operation results in obtaining a set of the basic phonological units called 'phonemes'. The same principle (= taking a larger context into account and applying the semantic criterion) is applied to morphemes, words and syntactic units. It seems to us, however, that at this point linguistic structuralism cannot be further developed. The sentence as a syntactic unit represents a limit for its development. For higher units, including text, this approach cannot be applied.

The only exception to the integration of text as a subject of a theoretical reflection is literary criticism, sometimes also called 'literary science'. This discipline doubtlessly fulfils the generally accepted non-explicit requirements imposed on cognition but certainly not the requirements placed on scientific theory. To state it as Popper does, its theories are not refutable. The demand formulated by the illustrious specialist in literary theory Jan Mukařovský (as referred to by his followers), to make out of literary criticism a real literary science, cannot be fulfilled without developing text linguistics as a theoretical science.

For both schools of linguistic structuralism (those of Copenhagen and of Prague), semantics remained an axiomatic starting point, a criterion serving for establishing the sets of basic units on different levels. Thus semantics is treated by the associates of these schools in the same way as in classical linguistics. However, if semantics is a part of the language system and if our desire is to study the properties of this system, then two steps are expected to be taken:

- changing the epistemological principles;
- changing the modelling imagination of linguists.

The epistemological principles were mentioned in the introductory chapter; the authorities are for us Karl R. Popper and Mario Bunge.[5] The second requirement concerns the approach applied in the present work: each language user has his/her own semantic system which functions in a specific way for each language communicative act.

What is common to many individual semantic systems was implanted and accepted through communicative acts to a specific biological structure. These common parts of each semantic system are accepted and fixed in the same or similar shape by a group or community. The other parts appear to be less fixed: from the language viewpoint they belong to the sphere of idiolect. A sharp boundary cannot be put between these parts of the system. This quality is a source of dynamics proper to languages. Any new communication can change or reinforce what was fixed and the mutable (individual, less fixed) units can become firmly settled.

In which way can we obtain information pertaining to the pattern of a (individual) semantic system? Linguistics needs to rely on everything that can be found as the output of the language communication system. The most important output probably is text. Naturally, people communicate in many other ways utilizing carriers of meaning other than language. Each of these means of communication can become a source of data about the producer's/receiver's semantic system. This holds, of course, only in the case when a practicable theory is at hand.

[5]The criticism of certain of Popper's ideas by M. Bunge cannot represent the reason for taking the quotation of these two epistemologists side by side as a contradictory expression. The principle of refutability of a scientific theory formulated by Popper is an important criterion for each science. After its application it becomes clear that a greater piece of responsibility for the catastrophes met by European civilization in the 20th century belongs not to real sciences, but to those disciplines which do not fulfil Popper's criterion. When these two philosophers are attentively read, it becomes evident that the principle of refutation is compatible with the precise analysis of sciences and their content made by M. Bunge. As a matter of fact, refutability is only one aspect among those representing criteria for science by both these philosophers.

On the other hand, the stress laid by M. Bunge on materialism seems to be surplus in any epistemology supposing human beings to be parts of the totality studied by all the sciences subordinated to the criterion of refutability. The dichotomy of "materialism" and "idealism" appears to be improper for sciences. Nevertheless, this is a philosophical problem resulting out of our competence.

The transmission of information represents for a receiver a stimulation to an interpretative activity. Any reception of information means some change of its arrangement for a part of an individual semantic system . And this very change of the organization of the system of meanings represents an act of communication (= transmission of information). This is the way in which communication is defined in the respective mathematical theory.

Thus in the receiver's semantic system a new organization of units and their relations occurs. The texture produced bears imprints of the instant shape of arrangement conjectured on the producer's semantic system. 'Instant' means here "approximately instant", occurring at a not very large time interval during which a text is produced *in continuo*.

The task of linguistics is to seek methods enabling us to glance at the respective semantic system and make inferences about its arrangement. It can be assumed that in each individual mind there is one semantic system which operates with different means of communication, one of them being natural language. We formulate the following hypothesis:

> Insights into the semantic system through different means of communication should present the same general picture of its arrangement.

For the meantime, when this hypothesis is not yet approved, the assumption about the existence of one semantic system run by each individual can be accepted. This semantic system is doubtlessly structured in some way and it cannot be excluded that certain of its parts are totally separated from the other parts, e.g. language semantics from art-design semantics, etc. Modern psychology with its concepts of consciousness and subconsciousness indicates that this separation is quite probable. It can be supposed that in dreams the semantic system is active, sometimes with language partcipation (= talking in sleep). These presumably are possibilities for future investigations in the field of psycholinguistics and its contributions to the theory of meanings.

We suppose that the most general property of the semantic system proved on texts is the structuring based on sentence aggregates. This property is expressed by the MA law. Let us seek support for the assumptions concerning the semantic systems, a way in which the same or a similar picture of this system is obtained as in the case of aggregates.

4.3 WORD ASSOCIATIONS AND AGGREGATES

Our attempt to find such support again depends on the results obtained by Altmann (1992), this time during his systematic investigation directed to the distribution of word associations. Word associations are a method used in psychology for entering into an individual human mind. Altmann stresses the linguistic viewpoints of this problem and emphasizes two of its properties:

1. Each word can provide a different number of associations; those accepted by a language community are characterized by high frequencies of occurrence. Those characterized by low frequencies consist of idiolectal, individual associations. These latter associations are supposed to be relevant for psychiatry.

2. Word associations are connected with diversifications of words and their investigation is relevant for synergetic linguistics concerned with self-regulation of the language systems.

As Altmann indicates, the distributions used to fit the data - namely Yule, Borel, Height's zeta and the logarithmic distribution (see W.J. Horvath 1963, F.A. Haight 1966) - are not adequate in the majority of cases.

The ideas of G.K. Zipf were further developed by P. Alekseev (1978) and his formula was applied by V.A. Dolinskij (1988) to word associations. A different and completely original derivation of this theoretical distribution was presented by R. Hammerl (1991). The formula of this distribution is:

$$f_x = f_1 \, x^{-(a+b \ln x)}, \qquad x = 1,2,... \tag{4.1}$$

where

x = the number of word associations,

f_x = the number of cases with x associated words,

a, b = coefficients.

Formula (4.1) has the status of a theoretical expression inspired by the Zipf distribution and empirically proved by comparison with the observed data. This formula was carefully investigated by Altmann together with the other theoretical distributions mentioned above.

Altmann redefined it as a probability distribution in the following way:

$$P_x = \begin{cases} \alpha, & x = 1 \\ \dfrac{(1-\alpha)\, x^{-(a+b \ln x)}}{T}, & x = 2,3,...,n \end{cases} \tag{4.1a}$$

where

$$T = \sum_{i=2}^{n} i^{-(a+b \ln i)}$$

and a, b, α are parameters.

After a statistical analysis Altmann comes to the conclusion that the fitting which uses (4.1) is excellent in all analyzed cases of word associations where the other theoretical distributions failed. Now the task of linguistics is to find the reason why it is so.

TABLE 4.1: Observed data of word associations, the distribution of "high" (Source: Palermo & Jenkins 1964; Altmann 1992)

x	f_x	x	f_x	x	f_x	x	f_x	x	f_x
1	129	8	4	15	2	22	1	29	1
2	16	9	3	16	2	23	1	30	1
3	14	10	3	17	2	24	1	31	1
4	12	11	3	18	2	25	1	32	1
5	6	12	2	19	2	26	1	33	1
6	5	13	2	20	2	27	1	34	1
7	4	14	2	21	1	28	1	35	1

In TABLE 4.1 the data quoted by Altmann are presented. The original source is Palermo & Jenkins (1964). The data are quoted here for the purpose of demonstrating the course of the distribution; see also FIGURE 4.2, where the respective curve is outlined.

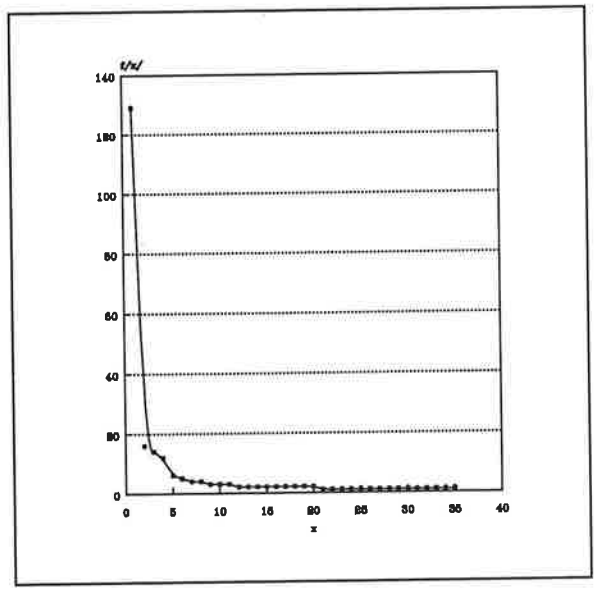

FIGURE 4.2: The curve fitting of data from Table 4.1
Source: Palermo & Jenkins 1964; Altmann 1992

After a careful inspection of several distributions of the observed word associations, it becomes evident that they are similar in two aspects:

- in each of them there is an indirect proportionality between the values of f_x on the one hand and of x on the other hand;
- the values of f_x are related to the respective f_1 as directly proportional.

These two characteristics can be written as:

$$f_x \sim \frac{1}{x}$$

$$f_x \sim f_1$$

Both these assumptions can be combined and formulated as the following hypothesis:

The mutual relation of f_x to f_1 is inversely proportional to x.

$$\frac{f_x}{f_1} \sim \frac{1}{x}$$

Thus our considerations can start with the expression:

$$\frac{f_1}{f_x} \sim x.$$

It can be rewritten in the logarithmic transformation as the following equation:

$$\ln f_1 - \ln f_x = \ln c \ \ln x, \qquad (4.2)$$

where *ln c* is a coefficient of proportionality that is also presented in a logarithmic form. From (4.2) it follows that

$$\left(\frac{f_1}{f_x}\right)^{\frac{1}{\ln x}} = c. \qquad (4.3)$$

Let us recall formula (2.5), which can be rewritten in the form:

$$\left(\frac{y_1}{y_x}\right)^{\frac{1}{\ln x}} = e^{-b},$$

where the structure of the left-hand side is analogous to (4.3), and x is the length of the constituent, in the terms of the MA law. Expression (4.3) is evidently inadequate, since the relation of the three variables on the left-hand side of (4.3) can hardly equal any constant. Another solution must be sought, and its source should be of a linguistic nature.

The ability to associate words to a given word unit can intuitively be connected with the ability of the person tested to produce a text on that given issue which contains the associated words. For words (= lexical units) 'to occur in a text' means "to be structured into sentence aggregates". In short, it can be assumed that the associated "text" is structured into aggregates. Thus we can formulate the second assumption about word associations:

> The associated words represent a (potential) text structure having the properties of aggregates.

This means, however, that word associations have the properties formulated by the MA law in the sense in which this law has been applied to aggregates. This justifies the substitution of c in (4.3) with the expression of the MA structure, i.e. with the right-hand side of (2.2). Thus we obtain:

$$\left(\frac{f_1}{f_x}\right)^{\frac{1}{\ln x}} = A\,x^b \qquad (4.4)$$

Hence it directly follows that:

$$f_x = f_1\, x^{-(a+b \ln x)}$$

which is formula (4.1) with $A = e^a$.

Summarizing this procedure, we can say that the above interpretation of (4.1) found the MA structure in it. The entire structure expressed in this formula can now be explained by the two hypotheses used in the supplemental derivation of this formula presented above. This derivation indicates that besides a normal text a set of associated words appears also to be a "text" similar (in a certain sense) to that structure occurring in each text having the usual form. Both these texts are obviously operated on with one and the same text producer's equipment, i.e. with the semantic structure of the mind.

We can treat this phenomenon as if we were looking into the producer's mind through two different windows: text and "text". And now we can, with certain grounds, assert that in these two insights we see the same - the structure described by the MA law. We are looking at something that belongs to the properties of the system of meanings. This can be vouched for, because word associations are language constructs stripped of almost all connections with higher language structures down to the semantic marrow. We are inclined to see in these facts a confirmation of the theory of aggregates. We suppose that the semantic system's outputs concealed in texts and in word associations are phenomena observable when the MA law is applied to them.

4.4 DISTRIBUTION OF AGGREGATES

The function expressed by formula (4.1) appeared to be a basis for the occurrence of the special sort of aggregates contained in word associations. It is quite natural to expect that the

same effect will be observed in aggregates of ordinary texts. Our interest concerns the degree of correspondence between the distribution z and the curve (4.1) when applied to data obtained from our text Corpus. This means that the values of frequencies z were substituted for x in (4.1). The respective values computed with the help of (4.1) are the expected values Z. The observed data were presented in Tables 3.1.1 - 10 in columns denoted z. The results of the curve fitting of these data are presented in TABLE 4.2. The course of the analyzed function together with the respective observed values are demonstrated in FIGURE 4.3 and FIGURE 4.4; these two curves are based on the values of z taken from Text 1 and Text 9. Their similarity with the curve for word associations in Figure 4.2 is evident. For each text in Table 4.2, the respective coefficient of determination (Coef. det. D) is presented, as well as the respective parameters a and b. The values of the coefficient of determination testify to the theory comprised in formula (4.1) and the respective values of Z explain more that 99 per cent of the variation in this variable.

Thus coefficient b has both positive and negative values and all that can be said at this moment about its properties is that it is very close to zero. Nevertheless, the influence of the MA law is obvious. We can pose the question of whether this influence comes from the semantic system and its supposed structure which is in agreement with the expression Ax^b.

TABLE 4.2: The distribution of aggregates in Texts 1-10
(see the z values in Tables 3.1.1 - 10)

Text 1

x	z	Z
1	273	273.01
2	83	83.11
3	39	36.97
4	14	19.76
5	17	11.81
6	5	7.61
7	7	5.18
8	4	3.67
9	2	2.69
10	2	2.03
12	1	1.22
13	1	0.98
17	2	0.44
19	1	0.32

Coef. det. D = 0.9989
a = 1.5377
b = 0.2570

Text 2

x	z	Z
1	364	363.94
2	116	117.01
3	55	52.45
4	29	27.89
5	12	16.50
6	10	10.50
7	10	7.05
8	1	4.93
9	4	3.57
10	4	2.65
11	3	2.01
12	1	1.56
13	2	1.23
15	1	0.79
18	2	0.44
22	2	0.23

Coef. det. D = 0.9995
a = 1.4213
b = 0.3113

Text 3

x	z	Z
1	192	192.08
2	69	68.19
3	31	31.67
4	14	17.10
5	13	10.17
6	7	6.48
7	3	4.34
8	6	3.03
9	1	2.18
10	2	1.61
11	2	1.22
15	1	0.47
31	1	0.04
32	1	0.03

Coef. det. D = 0.9990
a = 1.2433
b = 0.3617

Text 4

x	z	Z
1	150	150.11
2	44	42.42
3	16	19.78
4	13	11.39
5	8	7.37
6	5	5.15
7	3	3.79
8	4	2.90
9	2	2.29
10	1	1.85
11	2	1.52
12	2	1.27
14	1	0.93
15	1	0.80
18	1	0.55
19	2	0.49
23	1	0.3278

Coef. det. D = 0.9988
a = 1.7862
b = 0.0535

Text 5

x	z	Z
1	260	259.86
2	76	78.18
3	41	36.11
4	20	20.24
5	12	12.68
6	7	8.56
7	5	6.09
8	2	4.51
9	4	3.44
10	2	2.69
11	1	2.15
12	1	1.75
13	3	1.44
14	1	1.20
15	2	1.01
18	1	0.64
19	3	0.56
21	1	0.43
28	1	0.20
29	1	0.19

Coef. det. D = 0.9992
a = 1.6242
b = 0.1566

Text 6

x	z	Z
1	323	323.13
2	99	97.67
3	44	43.89
4	17	23.79
5	16	14.42
6	12	9.42
7	7	6.50
8	6	4.67
9	6	3.47
10	4	2.64
11	2	2.06
13	1	1.31
14	1	1.07
15	1	0.88
17	1	0.62
19	1	0.45
21	1	0.34
46	1	0.03
49	1	0.02
54	1	0.02

Coef. det. D = 0.9993
a = 1.5704
b = 0.2247

Text 7

x	z	Z
1	171	170.96
2	57	57.76
3	28	25.19
4	11	12.81
5	5	7.21
6	5	4.37
8	3	1.86
9	1	1.29
11	1	0.67
13	2	0.38
30	1	0.01

Coef. det. D = 0.9992
a = 1.2616
b = 0.4382

Text 8

x	z	Z
1	148	148.02
2	40	39.81
3	17	17.12
4	11	9.09
5	1	5.46
6	3	3.55
7	2	2.45
8	4	1.76
9	4	1.31
10	1	1.00
14	1	0.42
15	1	0.35
16	1	0.29

Coef. det. D = 0.9981
a = 1.7767
b = 0.1701

Text 9

x	z	Z
1	190	189.98
2	39	39.18
3	15	15.91
4	11	8.48
5	7	5.23
6	1	3.54
7	2	2.55
8	1	1.92
9	1	1.50
10	1	1.21
11	1	0.99
12	1	0.83
14	1	0.60

Coef. det. D = 0.9994
a = 2.3122
b = -0.0499

Text 10

x	z	Z
1	287	286.86
2	65	66.97
3	32	30.26
4	20	17.67
5	14	11.81
6	9	8.58
7	4	6.59
8	6	5.27
9	3	4.34
10	2	3.66
12	1	2.75
14	1	2.17
19	1	1.38
23	1	1.05
33	1	0.65

Coef. det. D = 0.9995
a = 2.1869
b = -0.1273

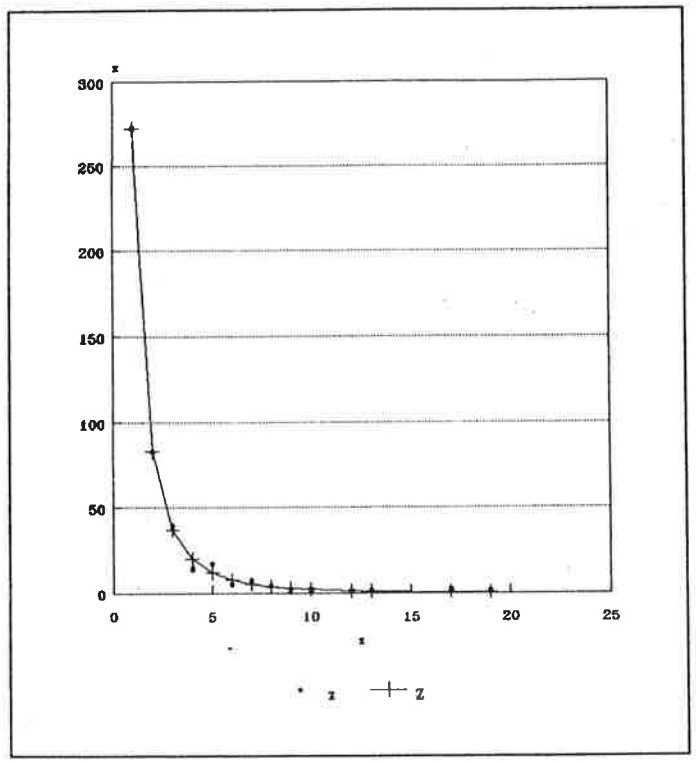

FIGURE 4.3: The distribution of aggregates (Text 1)

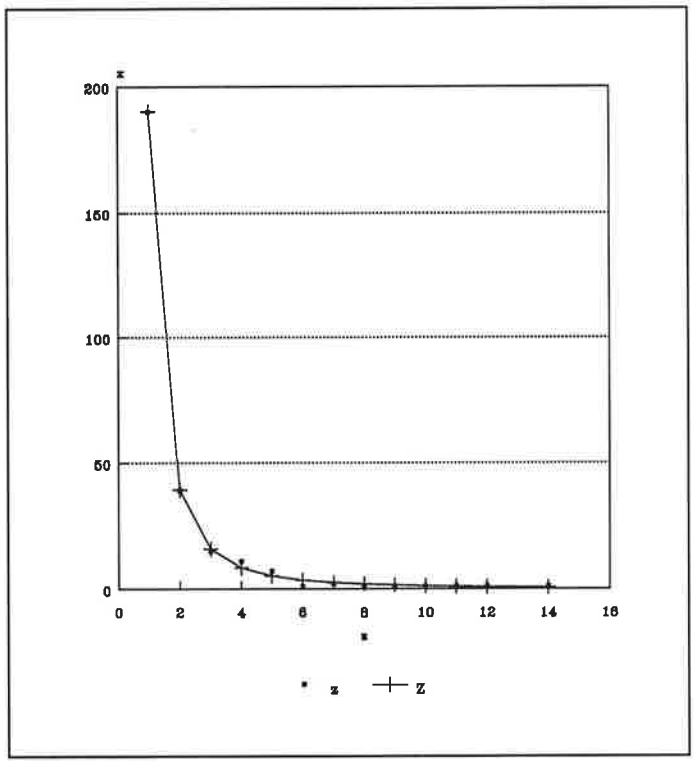

FIGURE 4.4: The distribution of aggregates (Text 9)

4.5. PREDICATION

In connection with the questions concerning the character of the semantic system we wish to discuss one semantic property coming from the categorization of grammatical phenomena and their occurrences in text. Predication is not only one of the characteristics of (sentence) syntax; it also includes inherent semantic qualities from which the formal properties treated by grammarians are deduced.

The MA law applied to aggregates involves two different segmentations of text: into words and into sentences. In this law both these types of segmentation are combined together. Let us stress that this approach represents a specific text-linguistic viewpoint. In any grammar, sentence is described as a phenomenon having its internal structure defined in a specific way. When some linguistic theory starts with text - and this approach seems to be more general than the theories starting with inner sentence structures - the segmentation of a text into structural parts becomes the most important problem which must be solved before the inner sentence structure is analyzed. In text linguistics, the sentence is not only a syntactic unit but also a result of text segmentation. Consequently, for grammarians sentence is a syntactic unit. For text linguists sentence, additionally, is a text segment.

From the viewpoint of syntax, a sentence is a manifestation of predication. This usual way of understanding and defining sentence is correct and need not be changed. Predication is manifested through the occurrence of a verbal construction headed either by finite or also by certain infinitival verbal forms; on this problem, see also Section 3.5. It is logical to change the viewpoint and say that predication takes part in the segmentation of text. The question we pose in this connection is: Does predication really participate in text structuring?

This question has a certain pragmatic background. Classical linguists sometimes argue against quantitative linguistics, saying that it deals with language phenomena in a way applicable to, let us say, timber in forests or to lumber elsewhere. The unsophisticated simplicity of such arguments is based on the conviction that quantities and their general properties are independent on the real items to which they are ascribed. The general properties of quantities are studied by the formal sciences, mathematics and logic. Thanks to this fact, empirical sciences can prove their principles and laws; the specificity of things is thus projected to that general background of cognition with the help of quantities. The following empirical operation can serve as a demonstration of the fact that variables cannot be freely shifted from one item

to another one, from timber in the forest to sentences. In the following experiment it is expected that the MA law will not be valid, and that the sentence with its properties is a unit which cannot be and really is not substituted by the items of lumber or something else.

Text 6, analyzed in Table 3.1.6, was analyzed again, this time, however, the real sentences with their predicational characteristics were neglected and substituted by segments chosen at random. In this procedure the random segments were obtained from random sampling numbers. We proceeded in the sequence of the word forms as they follow in this text and marked a segment after the length indicated by a random number taken from the tables of these numbers. This means that words and their frequencies remained the same as in Table 3.1.6, only the sentences, their limits and their numbers in the text were changed and random "sentences" were obtained. These "sentences" form aggregates with the distribution presented in TABLE 4.3 and 4.4. It is evident that these distributions mutually differ, and they both differ from the observed values of z indicating the real sentences; see Table 4.2 (Text 6). The difference between Table 4.3 and 4.4 consists in the length of "sentences": in the former Table the random numbers have only one integer, while in the latter Table the random numbers have two integers. This means that in the first case the "sentences" are shorter than in the second case.

In each of these Tables we present two sets of results: one for the total observed distribution as it is presented in the whole Table, the second one for the part of the distribution represented by higher values of z (down to the dashed line). In the case of shorter "sentences" (Table 4.3), both the values of Altmann's parameter b are positive; this means that the MA law is not fulfilled. This is in accordance with our expectation.

In the second case (Table 4.4), negative b was obtained for the entire distribution. This is in agreement with the MA law, but not with our expectation. We see that the expected values Y_1 are influenced by the lower values of y for higher x's, i.e. by the periphery of the distribution with the prevalence of random fluctuations. When the parameter is computed for the first 10 values of x, then it is evident that the structure of this text has nothing in common with the MA law: b is positive. This convincingly demonstrates that in the structure of this half natural and half simulated text there can scarcely be found a tendency to comply with the MA law.

The entire argumentation presented has an empirical character. The segmentation of text into sentences should be more deeply studied. Predication is a property of a pure observational nature. The same concept reveals its carriers purely in a semantic way; it evidently cannot be theoretically derived from some strictly formal presumptions. The experiment presented above serves only as evidence of the linguistic character of the MA law, though this law also has properties of a more general character, as will be discussed later. General structures structures observed in languages, and only in languages, are carriers of properties having special inner functions in the language code. The carriers called sentences having the semantic property of predication are constituents of aggregates.

Indeed, it can happen that a certain artificial segmentation of text will end in a negative value of b, as is the case of Y_1 for longer "sentences"; see Table 4.4, where this occurs thanks to the higher marginal values of x. The operation of segmentation is a random act which is subordinated to the laws of chance. Therefore we expect that in the greater part of all theoretically assumed experiments with the random segmentation the results will be in accordance with expectations as is usual in statistical experiments.

TABLE 4.3: The distribution of aggregates based on the "sentences" selected with the help of random numbers (shorter variant).

Text 6

x	z	n	y	Y_1	Y_2
1	333	2186	6.56	6.65	6.60
2	95	1228	6.46	6.66	6.60
3	45	875	6.48	6.66	6.61
4	16	443	6.92	6.66	6.61
5	14	471	6.73	6.66	6.61
6	14	587	6.99	6.66	6.62
7	7	323	6.59	6.66	6.62
8	6	310	6.46	6.66	6.62
9	6	347	6.43	6.66	6.62
10	5	327	6.54	6.67	6.62
12	1	82	6.83	6.67	-
14	1	93	6.64	6.67	-
15	1	111	7.40	6.67	-
16	1	105	6.56	6.67	-
18	1	128	7.11	6.67	-
20	2	260	6.50	6.67	-
45	1	283	6.29	6.67	-
48	1	328	6.83	6.67	-
62	1	398	6.42	6.67	-
Σ	551	-	-	-	-

For x = {1 - 62}: A = 6.6564; b = 0.00057. (Y_1)
For x = {1 - 10}: A = 6.5959; b = 0.00175. (Y_2).

TABLE 4.4: The distribution of aggregates based on the "sentences" selected with the help of random numbers (longer variant)

Text 6

x	z	n	y	Y_1	Y_2
1	345	19159	55.53	58.43	53.41
2	95	10651	56.06	56.64	54.61
3	39	6171	52.74	55.62	55.33
4	20	4462	55.77	54.91	55.84
5	15	3945	52.60	54.37	56.26
6	8	2631	54.81	53.93	56.58
7	5	1942	55.49	53.55	56.86
8	5	2209	55.23	53.24	57.10
9	5	2732	60.71	52.96	57.32
10	1	625	62.50	52.71	57.51
11	1	482	43.82	52.48	-
14	2	1399	49.96	51.92	-
25	1	1293	51.72	50.59	-
26	1	1223	47.04	50.50	-
27	1	1289	47.74	50.41	-
Σ	553	-	-	-	-

For x = {1 - 27}: A = 58.4280; b = - 0.04476. (Y_1)
For x = {1 - 10}: A = 53.4107; b = 0.03213. (Y_2).

4.6 SEMANTIC SPECIFICITY OF AGGREGATES

It was indicated that the MA law has been proved to be valid for language levels in many languages, i.e. for the levels which are intuitively evaluated as levels regardless of the theoretical support of the law. The level of aggregates is the only one derived with the theoretical help of a language law. It was proved for Turkish and German texts.

An argument against aggregates can be stated, namely, that they are based only on certain statistical tricks, for example, on those discussed in Section 3.10. Generally expressed, you can take arbitrary lower units, then make their inventory and from all occurrences paste together a higher unit which can be declared a new level with the help of the statistical tricks of the MA law. If this is correct, then all the above speculations concerning the semantic nature of aggregates have no grounds. This is the leitmotif of the following experiment from which a negative confirmation is expected.

Suppose the following two "aggregates": a syllabic and a morphological one. As for the former, an analyzed text is rewritten into syllables and all those words, in which a given syllable occurs, are hypothetically declared to form a "syllabic aggregate". This "aggregate" is a new language level if it accomplishes the MA law.

The analogy of this "aggregate" is the hypothetical aggregate based on morphemes: a text is rewritten in morphemes and all words in which a given morpheme occurs form a "morphological aggregate". Then two parallel hypotheses can be tested from the viewpoint of the MA law:

The longer a syllabic/morphological "aggregate" in number of words, the shorter its mean syllable/morpheme in number of phonemes.

This is a complete analogy formulated and positively tested for the sentence aggregates. A Turkish text was analyzed and its syllabic aggregates were found. The data are presented in TABLE 4.5. They are evidently intuitively less persuasive than in most cases of the sentence aggregates. The mean syllabic length in column (D) looks like a variable fluctuating without any observable proclivity, and this holds also for the tendency prescribed by the MA law, i.e. the decrease of the mean syllable length. The parameters of the MA law were computed for

the first seven values of x (column A of Table 4.5) which are represented by higher distributional values (column B). The results are not in agreement with the MA law: parameter b is positive and the expected means of the syllable length (column E) are not decreasing.

An analogous experiment with morphemes is presented in TABLE 4.6. The analysis of the morpheme "aggregates" resembles that of their syllabic counterparts. The reason for which only the first seven sizes of these aggregates were taken for computation is the same as in the case of syllables. There is no basis for taking syllable and morpheme "aggregates" as language levels.

This experiment has empirical validity. Nevertheless, it is rational to present its results as support for the speculations about the semantic nature of that structure which is proper to sentence aggregates and also to word associations. These results simply mean that we cannot state aggregates *ad libitum*. Sentence aggregates observed in texts represent a real language construct. The existence of this construct seems to be justified by evident characteristics of text, i.e. by the properties of languages recognized during the history of linguistics as well as by the intuition of language users.

TABLE 4.5: Syllabic "aggregates" in a Turkish text

(A)	(B)	(C)	(D)	(E)
1	132	416	3.15	2.98
2	58	345	2.97	3.10
3	29	257	2.95	3.18
4	16	210	3.28	3.23
5	14	238	3.40	3.27
6	11	206	3.12	3.31
7	5	125	3.57	3.34
8	2	40	2.50	
10	2	63	3.15	
11	6	220	3.33	
12	1	17	1.42	
13	1	46	3.54	
14	4	174	3.11	
15	1	47	3.13	
16	1	58	3.63	
18	3	145	2.69	
19	1	65	3.42	
23	1	71	3.09	
24	1	70	2.97	
Σ	289	2813	-	-

A = 2.98358 b = 0.05723

(A) - the length of syllabic aggregates (in number of words); (B) -the number of aggregates; (C) - the sum of phonemes; (D) - the mean length of syllables (in phonemes); (E) - the expected syllable length.

TABLE 4.6: Morphemic "aggregates"

(A)	(B)	(C)	(D)	(E)
1	123	261	2.12	2.04
2	36	150	2.08	2.20
3	21	149	2.37	2.30
4	16	140	2.19	2.38
5	11	166	3.02	2.44
6	4	42	1.75	2.49
7	6	131	3.12	2.53
8	1	27	3.38	
9	4	90	2.50	
10	2	70	3.50	
11	4	92	2.09	
12	1	40	3.33	
14	1	33	2.36	
15	3	133	2.96	
20	1	64	3.20	
25	1	81	3.24	
Σ	235	1653	-	-

A = 2.03925 b = 0.11087

(A) - the length of aggregates (in number of words); (B) - the number of aggregates; (C) - the sum of phonemes; (D) - the mean length of morphemes (in number of phonemes); (E) - the expected length of morphemes.

The MA law presents rational explanation. The monumentality of this element in our language knowledge can hardly be overvalued. The semantic consequences of the MA law are fascinating. They can be, however, explained in another way.

4.7 AGGREGATES AS CONSTITUENTS

In the preceding sections we tried to indicate that classical linguistic schools (including structuralism) tried to be consistent to the degree that they seek applications of a methodical approach to one level also to the other levels; for example, the approaches to phonology were applied to syntax. Now we try to do something similar with the MA law. The difference, however, consists in the character of our approach (and also in the character of the MA law, as will be seen in the next chapter). While phonological decisions are built on the semantic intuition of linguists or their informants, the position of the MA law is substantially different. Altmann formulated the methodological principle for each scientific law as follows:

> 'Je allgemeiner die Begriffe in einer gesetzartigen Aussage sind, desto mehr Konsequenzen lassen sich durch Einsetzung spezifischer Begriffe ableiten.' (Altmann & Schwibbe 1989, 3.)

His approach to the formulation of the MA law was in agreement with this principle; this has been the step he made from Menzerath's formulation containing concepts such as 'sound' or 'syllable' to the formulation containing the concepts 'construct' and 'constituent'. The general formulation of these methodological principles are presented in Altmann (1988, 6-10); the principles mentioned are quoted with reference to the philosophical expressions by Bunge (1967, 222 ff.). As far as the methodological homogeneity of language levels is concerned, in the same book as cited above (on p. 5) Altmann writes, in connection with his generalized formulation of Menzerath' expression:

> 'Diese Hypothese ist so allgemein, daß sie sogar die Grenzen der Linguistik überschreitet...Wir beschränken uns hier zunächst auf die Linguistik und gehen dabei von dem Prinzip aus, daß die Sprache einheitlich konstruiert ist (vgl. das parallele Prinzip der 'Einheit der Natur', der 'Einheit der Wissenschaft' usw.), d.h., daß ein in der Sprache wirkendes Gesetz auf allen ihren Ebenen wirkt. Prinzipien dieser Art sind metatheoretischer Art (vgl. das Prinzip der geringsten Anstregung) und dienen als Leitfaden beim Aufbau von Theorien.'

Hence it follows that it is our duty to examine the theory presented by the extension of its principles to all language levels which are in connection with aggregates.

The MA law functions on the traditional language levels as well as on the level of sentence aggregates. By its affinity to word associations and to predication it was exhibited that these aggregates have something in common with semantics, with the system of meanings located somewhere in the human brain. From the level of sounds up to this highest level constituted by meanings in the semantic system, we have a string of constructs compounded of the respective constituents. There are, however, two lacunae in this path from sounds to semantics. The first one can be expressed by the following question:

The explanation of the distribution of word associations was grounded on the supposition that word associations have something in common with the units observed in texts as sentence aggregates. Let us suppose an arbitrary association to a word. What do words taking part in this association (i.e. the stated word and its associative echo) represent; are they constituents or constructs? Which one is constituent and which one is construct? If they represent two steps on the ladder of constructs and constituents, which are their neighbouring levels?

It is obvious that these questions require a further experimental investigation in order to obtain data for solving these problems. What is missing here is evident when we compare the tables presenting data of aggregates with those presenting the distribution of word associations. The latter ones contain only two columns: that for x and that for f_x. We lack here an analogue of mean length of constituents in the case of aggregates. It cannot be obtained till specialists in this branch of psychology come up with something that could be called a 'text of word associations'.

In what way can such "text" be obtained? A preliminary conjecture leads to the following way of obtaining it: A word is given to the person taking part in such an experiment. This person pronounces associated word(s) and then each of these associations is again put into the experiment. At first glance it seems that an endless set of associations will be obtained, but this is not correct. It must be expected that certain words will remain without any associative

response. We suppose that due to simple 'psychological' reasons this 'entertainment' must stop, when both experimentalist and the analyzed subject become tired. It is possible that in the psychological literature such experiments are already described.

This is nothing but a suggestion for organizing an experiment. Whether it is possible or not is beyond our ability to guess. Naturally, the fulfillment of the MA law is expected. A similar experiment can be made with sets of texts. For example, several novels of a writer can be analyzed; these novels probably are contained in one book and originated within a relatively short time. They are selected in order to obtain a picture of an (approximately) instantaneous state of the writer's system of meanings. Thus a larger vista of this system can be obtained than in the case where only one text is analyzed. The results of such experiments can be evaluated from different aspects, e.g. from that of the dynamics of meanings or comparison of different authors of texts, etc. Results of such experiments are not presented here, simply because they surpass the aims stated at the beginning of the present work as well as our possibilities for organizing larger experiments at the present time.

The second gap in the string of the hierarchized constituents and their constructs concerns a normal text in a natural language. The question arises as to whether text, if its sentence aggregates are constructs, is also a construct having sentence aggregates as its constituents. This question can be answered only with the help of the MA law, of course. Its application, however, is not simple. For the purpose of solving this problem, let us consider the status of several variables characterizing text:

n_i — the length of the ith sentence (in number of words), $i = 1, 2, \ldots, k$;

$\Sigma_i n_i = n$ — the total length of the text (in number of words);

v — the number of different words, i.e. the number of lexical units.

When we accept the assumption that each lexical unit occurs only once in a sentence, then we can rename the above variables in the frame of the theory of aggregates:

v — the number of aggregates in a text; this is correct, as in one aggregate all occurrences of a given lexical unit are unified; this variable also represents a text length in number of aggregates when text is supposed to be a construct and aggregates its constituents;

n — the sum of lengths of all sentences occurring in a text (in number of words) and, at the same time, the sum of frequencies of all lexical units;

n/k — mean sentence length in the respective text;

n/v — the mean frequency of lexical units and also the mean length of constituents (= aggregates).

Consequently, in the terms of the MA theory, when it is applied to aggregates (= constituents) and text (= construct), $x = v$ and $y = n/v$. Then the MA law (2.2) can be rewritten for text as:

$$\frac{n}{v} = A\, v^b \tag{4.5}$$

The main problem for this expression is the estimation of its parameters. They can possibly be estimated when a set of texts originated by one author within a not very long time interval is analyzed, as has been mentioned above. From such a bundle of texts, the MA parameters can be estimated with the presumption that the circumstances of the production of its texts did not change in a substantial way. Here we can determine parameter b from the perspective of the MA law on two contrastive types of texts which are widely apart from the usual form of texts normally supposed in linguistics.

The first one is a text containing only one word. With $n = v = 1$, from (4.5) it follows that $A = 1$. From (4.5) it can be further deduced that

$$b = \frac{\ln n - \ln A}{\ln v} - 1. \tag{4.6}$$

If $v = 1$, then the logarithm in the denominator of (4.6) equals zero and thus the fraction in this equation converges to zero; then b converges to -1. This means that the MA law is valid for this strange type of text.

Suppose another type of text in which each word has frequency one, i.e. $n = v$. Then from (4.6) it follows that

$$b = -\frac{\ln A}{\ln n}. \tag{4.7}$$

Here also the resulting value is negative and thus in accordance with the MA law.

Text can also be understood as an entity which increases sentence after sentence, and on each step the part already finished also represents a text. This finished part is in fact a sub-text of the increasing text, but still a text. Such an experiment, based on $x = v$ and $y = n/v$ as in (4.5), has been made on a Turkish text (Text 7 of our corpus). The results are presented in in TABLE 4.7.

TABLE 4.7: Text as a construct with aggregates measured by *n/v* ("increasing text")

Text 7

Sentences	$v_{(5)}$	v	$n_{(5)}$	n	n/v
1-5	39	39	42	42	1.08
6-10	42	81	64	106	1.31
11-15	34	115	49	155	1.35
16-20	16	131	30	185	1.41
21-25	7	138	16	201	1.46
26-30	18	156	33	234	1.50
31-35	17	173	38	272	1.57
36-40	14	187	30	302	1.61
41-45	22	209	55	357	1.71
46-50	14	223	44	401	1.80
51-55	18	241	47	448	1.86
56-60	15	256	33	481	1.88
61-65	16	272	49	530	1.95
66-67	13	285	29	559	1.96

$v_{(5)}$ = the number of lexical units in the part of the text containing 5 sentences;
$n_{(5)}$ = the length of each part of the text in 5 sentences (in number of words).

At first glance, it is evident that the variables *v* and *n/v* are not in agreement with the MA law. They do not function as language levels, or more precisely, they do not characterize items representing language levels. It seems to be highly probable that these unsuccessful results follow from the fact that sentences are not involved in the measurement of aggregates the length of which should always be given in number of sentences. Therefore we tried to make use of *n/v* again and relate it to another characteristic of constructs. In order to obtain an increasing progression of values for *x* as argument of the decreasing function *n/v*, let us examine the following model of an increasing text:

The text producer, when coming to the *k*th sentence (which is the last sentence of a text), evaluates the finished total by the value *one*. When a text is finished up to the $(k - 1)$th sentence only, the already finished part of the text also represents a total of one plus one sentence missing to the end of the text; the sum equals two. Similarly, after the $(k - 2)$ sentence, the stage of text origination is evaluated as 3, etc.

This means that we assume the act of text production to be a strategic process tending to a final aim presupposed by the text producer. This idea was introduced into linguistics by J.K. Orlov after the investigation of the lexical structure of text (see J.K. Orlov, M.G. Boroda & J.Š. Nadarejšvili 1982). This idea was also exploited by his collaborators, who formed the Tiflis Quantitative Linguistics group; another conception of this model was presented by Altmann (1988, 59), who says:

> 'Orlov schloß daraus, daß der Verfasser des Textes eine geplante Länge des Textes im Sinne hat und den Informationsfluß auf diese Gesamtlänge zerlegt.'

Thus we obtain a variable characterizing an increasing text:

$$g = \{ k - (k - 1) = 1, k - (k - 2) = 2, \ldots, k - 2, k - 1, k \}.$$

Let the appropriate values of *n/v* be ascribed to *g* which characterizes a planned increase of a text by its producer (in number of sentences). The data taken from the same text as in Table 4.7, naturally, are the same but ordered in the reverse sequence. They are presented in TABLE 4.8.

TABLE 4.8: Text as a construct characterized by *n/v* and *g* ("increasing text")

Text 7

g	N/v	Y	g	N/v	Y
1	1.96	2.16	33	1.57	1.52
3	1.95	1.93	38	1.50	1.49
8	1.88	1.75	43	1.46	1.48
13	1.86	1.66	48	1.41	1.46
18	1.80	1.61	53	1.35	1.44
23	1.71	1.57	58	1.31	1.43
28	1.61	1.54	63	1.08	1.42

A = 2.1586; b = - 0.1012.
Coef. det. D = 0.6645.

The coefficient of determination indicates that only about 66% of the relation between *g* and *n/v* is explained by the MA law. This result is not so bad when we take into account that this construct (text) and its constituents (aggregates in the planned increase *g*) are described with a high approximation represented by the combined variable *n/v*. The correlation between these two variables is significant. This analysis of increasing text indicates that increases can be qualified as constituents and the whole text as a respective construct in the sense of the MA law.

A simpler characteristic can be used for characterization of an increasing text as a construct and of its aggregates. The number of aggregates *v* and the number of sentences *k* are appropriate for this purpose. The fraction *v/k* can be interpreted as "a mean proportion of aggregates appertaining to one sentence." Then it can be observed how this characteristic changes together with the text growth. The same fraction can also be interpreted as the complex fraction:

$$\frac{\frac{n}{k}}{\frac{n}{v}} = \frac{v}{k}$$

This is the relation of mean sentence length to the mean length of an aggregate, or in short: the mean sentence length per mean aggregate. Then again we observe how this mean characteristic changes when a text is increasing sentence after sentence. There are two possibilities for ascribing an argument to the function *v/k*: its argument can be either *v* or *k*. The data with respect to both these possibilities are presented in TABLE 4.9; this analysis is also based on Text 7. In TABLE 4.10 and TABLE 4.11 the data taken from another two Turkish texts are presented.[6]

The variable *v/k* depends on the course of both the variables from which it is composd. In the specific interpretation assigning the meaning of a "measure of the text structure formed by aggregates" to the number of lexical units, this fraction also indicates that the text level which is the highest one in the arrangement of the text units is subordinated to the MA law. At the present moment we do no know any reason for the rejection of the hypothesis saying that text is a language construct in the sense of the MA law. Evidently the testing presented above offers only tentative results because aggregates in the analyzed text are not measured directly. Nevertheless, it is highly probable that text really is a construct composed of aggregates.

Both the computed theoretical values show approximately equal fit to the observed values. The explanatory power of the composed variable *v/k* is not very high, though the results of testing the hypotheses connected with this variable are statistically significant. Nevertheless the entire experiment with text taken as a language construct is not definitive. It deserves to become an object of further investigation, next time with a group of texts produced by one author within a not very large time interval.

[6] The data in Table 4.10 and 4.11 present minor corrections of those published in L. Hřebíček (1993), which is the preliminary study of the problem.

The arrangement of such an experiment is still problematic due to the possibility of style variation by one and the same author even during a short time interval. Other questions connected with increasing text are discussed in Chapter 5.

TABLE 4.9: Text as a construct with aggregates measured by *v/k* ("increasing text")

Text 7

k	v	v/k	$Y_1=Av^b$	$Y_2=Ak^b$
5	23	4.60	4.48	4.52
10	39	3.90	4.27	4.26
15	60	4.00	4.11	4.12
20	131	6.55	5.78	5.97
25	138	5.52	5.68	5.61
30	156	5.20	5.44	5.33
35	173	4.94	5.24	5.11
40	187	4.68	5.10	4.92
45	209	4.64	4.90	4.76
50	223	4.46	4.79	4.62
55	241	4.38	4.66	4.50
60	256	4.27	4.56	4.39
65	272	4.18	4.47	4.29
67	285	4.25	4.39	4.26

Y_1: A = 32.2935; b = - 0.3529; coef. det. D = 0.7344.
Y_2: A = 13.8299; b = - 0.2802; coef. det. D = 0.8464.

TABLE 4.10: Text as a construct (similar analysis to that in Table 4.9)

k	v	v/k	$Y_1=Av^b$	$Y_2=Ak^b$
5	23	4.60	4.48	4.52
10	39	3.90	4.27	4.26
15	60	4.00	4.11	4.12
20	79	3.95	4.01	4.02
25	107	4.28	3.90	3.94
30	126	4.20	3.85	3.88
35	143	4.09	3.80	3.83
40	160	4.00	3.77	3.79
45	173	3.84	3.74	3.75
50	181	3.62	3.73	3.71
55	192	3.49	3.71	3.68
60	204	3.40	3.69	3.66
65	220	3.38	3.66	3.63
70	249	3.56	3.62	3.61
75	273	3.64	3.59	3.59
80	291	3.64	3.57	3.57
85	299	3.52	3.56	3.55
90	313	3.48	3.55	3.53
95	333	3.51	3.53	3.52
100	352	3.52	3.51	3.50
103	364	3.53	3.50	3.49

Y_1: A = 5.9277; b = - 0.0893; coef.det. D = 0.6260.
Y_2: A = 5.1810; b = - 0.0855; coef.det. D = 0.6856.

Text: C. Tanyol, Atatürk ilkeleri. In: Y. Nabi (ed.), *Atatürkçülük Nedir?* Istanbul 1969, 105-109.

TABLE 4.11: Text as a construct (the same approach as in the two preceding Tables)

k	v	v/k	$Y_1=Av^b$	$Y_2=Ak^b$
5	54	10.80	10.85	10.87
10	98	9.80	9.99	9.98
15	143	9.53	9.48	9.49
20	190	9.50	9.11	9.16
25	233	9.32	8.86	8.91
30	255	8.50	8.75	8.71
35	302	8.63	8.54	8.55
40	327	8.18	8.45	8.41
42	343	8.16	8.39	8.36

Y_1: A = 18.8974; b = -0.1390; coef.det. D = 0.8996.
Y_2: A = 13.2511; b = -0.1233; coef.det. D = 0.9225.

Text: Sir James Redhouse (1811-1892). In: *Redhouse Yeni Türkçe-İngilizce Sözlük.* Istanbul 1968, X-XI.

It is quite natural to pose the question as to why the MA law is so general that all language levels from the lowest one up to the text level are organized in accordance with this law. There is no doubt that science should seek such an explanation. In the present chapter we tried to indicate several connections of this law to certain linguistic phenomena. A hypothesis connecting text with the individual human brain into one whole has been presented. Is the assumption reaching down to biological strata of the process in which a text is originated sufficient for the required explanation? Evidently it is not. There must be something deeper in the structure of language and its biological carrier. We try to indicate that there is something that connects such at first glance mutually remote phenomena. They are remote only in their scientific presentation, whereas in reality it is not so, of course. The human brain and language evidently belong to one another. The same cannot be said about biology and linguistics. Fortunately mathematics can put them together in a natural and non-violent way.

5. Levels as a Dynamic System

The following discussion concerns some aspects of the theory of fractals as a theoretical background of the MA law. It represents an attempt at another way of explication offered by this law in connection with text components. During the last decades the concept of fractal has been intensively investigated in mathematics and the natural sciences. This ingenious discovery is sometimes suspected by human scientists to be nothing more than a fashion which after a culmination of interest will soon finish. This, however, is an inappropriate opinion. This theory is one of the greatest discoveries of our time. It was formulated by Benoit B. Mandelbrot; see at least his famous book *The Fractal Geometry of Nature* (1982). Any intellectual afflicted with a lack of interest in this new mathematical understanding of the specific structures described by this theory resembles an educated person of the 3rd century B.C. in Greece without any knowledge of Euclid's *Elements*. The present attempt to apply Mandelbrot's far reaching ideas in linguistics is nothing more than a real venture at understanding these new ideas by a linguist and at their application in the thinking about language.

Background usually means an item on a second plane of interest, which should be taken into account only after the first-plane problems are analyzed. This is not the case in our treating Mandelbrot's theory as a background to language phenomena. Here 'background' means something immense and the linguistic problems assumed represent only a small component standing in connection with it. In linguistics the concept of context is often used. 'Context' usually designates certain language phenomena related to the item actually studied; or it designates an extralinguistic item which can easily be transformed into some language constructs. The background formed by fractal structures represents something broader and

cannot be converted easily into a discursive counterpart. In our treatment the notion of text is connected with a semantic system; and semantic system, as was indicated above, is a real phenomenon, a system proper to each individual user of a natural language. It has also been assumed that this system is placed into certain biological structures of a normal human being. In connection with word associations it was proved that in this specific form of language usage certain properties observed in ordinary texts occur. This concerns the functioning of the MA law. This law functions in so many different language structures as their main principle that it can be designated as omnipresent. Why is it so? We try to indicate in this chapter that the reason has something in common with the background mentioned above.

The relationship between the MA law and fractal theory was already mentioned in Hřebíček (1992, 91-95); see also the paper 'Fractals in language' (1994) by the same author. Certain preliminary ideas mentioned in these works require corrections, as will be seen below. Nonetheless, it seems to be evident that language deserves further investigation from the viewpoints of fractal theory as well as from the perspective of other concepts playing an important role in Mandelbrot's mathematical formulations. One such concept is that of *chaos* discussed in Hřebíček & Altmann (1994).

It can be assumed with a high degree of reliability that meaning and language are not identical phenomena because their structures are not paired in some transparent correlative relations. When meaning is isolated from language in order to understand the system of meanings through language, then what remains? This situation resembles hearing talk in a language unknown for the hearer. The first impression is that of chaos, though soon it will be evident that there is present some principle or principles of organization. A semantic system is often adjusted to linguistic theories in the form of a 'competent informant'. The notion of competence has something in common with that of the individual semantic system, with its properties (especially with those operating with language) and with the biological structures in which this system is embedded.

The sets and structures described in the theory of fractals seem to be the first step towards the future explanation of those mutually combined phenomena. This problem is complex and it cannot be fully disclosed in the following several pages; it requires more detailed analyses than those contained in the sections of this chapter. We still remain at the beginning of the clarification sought.

5.1 LANGUAGE AND FRACTAL THEORY

The problem to be solved is not how to grasp the idea of self-similarity of language subsystems. This notion was already utilized in different linguistic (informal) theories, though this fact is not apparent at first sight.

Let us begin our attempt with presenting the basic concepts of the theory of fractals in an informal way. Our information about the theory is grounded on the basic work by Mandelbrot (1982) quoted above and further on Jens Feder (1988/1991), J.T. Sandefur (1990), M. Barnsley (1988), H.-O. Peitgen & H. Jürgens & D. Saupe (1992) and many other works; in these books, especially in the latter one, a large bibliography of the theory can be found.

The formal definition of fractals stated by Mandelbrot (1982, 15) is based on the idea of topological dimension:

> 'A fractal is by definition a set for which the Hausdorff-Besicovitch dimension strictly exceeds the topological dimension.'

The concept of topological dimension can hardly be applied to language constructs. Jens Feder (1991, 19, Section 2.3) presents the following instructive characteristics:

> 'A structure consisting of parts, which are in a sense similar to the whole, is called "fractal".'

The same author also characterizes fractals as follows:

> 'A fractal looks the same regardless of the scale in which it is observed.'

Essentially, each of these characterizations obviously involves the relation between a construct and its constituents. At first sight, from the words 'part' and 'to consist', it is evident that this topic has something in common with the MA law.

Fractals and their relations are usually illustrated by a set of standard examples to which the example of the coast length also belongs. This line includes a certain similarity between its different pictures taken from different distances. There is a similarity between the shape of a coast line when it is observed from a satellite without details of the minor bays and capes, or from a plane flying at a height of thousands of feet or from several hundred feet. This is one sort of similarity. The notion of *self-similarity* has been introduced by Mandelbrot for the set patterns in which each piece of a shape is similar to the whole. Hence it follows that there

are different kinds of similarity. This becomes evident when the measurement of the supposed shapes is taken into consideration.

Suppose a straight line (for example, a yard-stick) having the length δ applied to a curve (coast line) $N(\delta)$ times. The shorter the straight line the better it covers the curve. The length of the curve is $L = N(\delta)\delta$. The length L is thus a function of δ. This means that when a rod by which a coast is measured shortens *ad infinitum*, i.e. $\delta \rightarrow 0$, the length of the coast grows infinitely.

Our presentation of the issue essentially follows the explication formulated by Feder (1991, 20 f.). He indicates that when we are measuring a curve or other shape representing a set of points, a probationary function is chosen - be it an off-cut of a rod, small square, circle, cube, or ball - which covers the measured set. When the probationary function used is $h(\delta) = \gamma(d)\delta^d$, the total measure is $M_d = \Sigma h(\delta)$. For each straight line the geometrical coefficient $\gamma(d) = 1$. Jens Feder indicates that in general with $\delta \rightarrow 0$ the measure M_d either equals zero or infinity. This result is conditioned by the choice of *d-dimension* of the measure. Then the Hausdorff-Besicovitch dimension D of the measured set is the critical value by which M_d changes its value from zero to infinity. Feder (1991, 22) presents the following formula:

$$M_d = \Sigma[\gamma(d)\ \delta_d] =$$

$$= \gamma(d)\ N(\delta)\ \delta^d \underset{\delta \to 0}{\Rightarrow} \begin{cases} 0, & d > D \\ \infty, & d < D \end{cases}$$

From this relation it follows that for infinitely small δ the length of the curve or set is

$$N(\delta) \sim \frac{1}{\delta^D}.$$

The dimension of a coast line can be measured by the measuring of the angular coefficient of the graph of $ln\ N(\delta)$ as a function of $ln\ \delta$. The straight line in the log × log coordinates corresponds to the relation

$$N(\delta) = a\ \delta^{-D}.$$

For the purpose of understanding the role of dimension in characterization of different sets (or simply: phenomena) such formations as *Koch curve* or *Cantor fractal dust* can serve as the best model. The latter is demonstrated in FIGURE 5.1. This set represents a construction in a steady process of generation. It begins with a straight line called *initiator*, and with a broken line called *generator*. These two lines are the first two steps in the generation of the set. The generation of this set goes on with application of the generator to each part of the line of the preceding step, as is evident from the third and fourth line of Figure 5.1.

Cantor dust is generated in the indicated way infinitely. The question arises as to how the similarity of all these steps can be expressed. Mandelbrot introduced two variables, one characterizing the whole curve as the number of parts into which the whole is divided *(N)*, and the similarity ratio *(r)* involved in all these parts. For the similarity dimension Mandelbrot displays the formula:

$$D_S = \frac{\log N}{\log (1/r)}$$

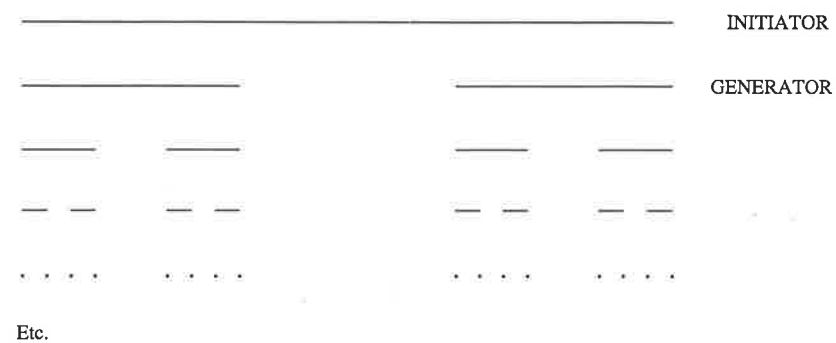

FIGURE 5.1: Cantor fractal dust

When the respective data for Cantor dust are substituted, we obtain:

$$D_S = \frac{\log 2}{\log (1/3)} = -0.63093$$

The similarity dimension thus appears to be an expression characterizing the sought-after invariant of the measured set. For the self-similar fractals, the Hausdorff-Besicovitch dimension D is identical to the similarity dimension D_S.

Now the following property of the analyzed set is substantial for the linguistic application: the *construct* of the set having the length N consists of r *constituents*, as is obvious from Figure 5.1. This indicates that the above formula containing variables N and r can be rewritten with the MA symbols x and y respectively. Thus we obtain:

$$D = \frac{\log x}{\log (1/y)} = \tag{5.1}$$

$$= \frac{\log x}{-\log y}$$

Then

$$\log y = -(1/D) \log x. \tag{5.2}$$

This expression can be supplemented by a correction term A in logarithmic form and $1/D$ can be substituted by b. Both these substitutions are formally correct. Then from

$$\log y + \log A = -b \log x + \log A \tag{5.3}$$

the MA structure is obtained on its right-hand side.

From this transformation it is evident that the concept of constituent in both the theories, fractal and MA, plays an important role. The structures corresponding to this concept are close to each other, though they are different. While on the right-hand side of (5.3) the structure corresponding to the MA law is exposed, on its left-hand side the constituent of the Hausdorff-Besicovitch formula is supplied by log A. Evidently we can then write that

$$\log y + \log A \neq \log y. \tag{5.4}$$

Regardless of their structural similarity, the two items are not identical. Consequently the other variable and parameters of the MA law also differ. While from (5.3) it follows that

$$b = -\frac{\log y}{\log x},$$

in the MA theory this parameter is related with the other parameter A:

$$b = \frac{\log A - \log y}{\log x}$$

This fact observed in languages is a corroboration of the intuitive characterization saying that language subsystems are mutually independent in a certain sense.

5.2 THE STRING OF LEVELS

Suppose x_1 and y_1 are a language construct and its constituents of a higher level, and x_2 and y_2 are the construct and its constituents of the nearest lower level. The supposed relations and their exemplifications are as follows:

x_1 construct of a higher level (for example: sentence length in number of words);
y_1 constituent of the higher level (for example: word length in number of syllables);
x_2 construct of the lower level (for example: word length in number of syllables);
y_2 constituent of the lower level (for example: syllable length in number of phonemes).

It is evident that $y_1 = x_2$ and similarly $y_2 = x_3$. Then the MA law can be written in the form:

$$x = \left(\frac{A}{y}\right)^{\frac{1}{b}}, \tag{5.5}$$

or, when taking into account the indices of the levels and their identities:

$$x_1 = \left(\frac{A_1}{x_2}\right)^{\frac{1}{b_1}},$$
$$x_2 = \left(\frac{A_2}{x_3}\right)^{\frac{1}{b_2}}. \qquad (5.6)$$

The levels can be arranged to the following sequence:

x_1
↕
$y_1 = x_2$
↕
$y_2 = x_3$
↕
$y_3 = x_4$
...

The whole structure is permeated by the relation described in the formula of the MA law. Thus it bears all the consequences of this law. From the string presented it is evident that the entire construction of the levels is contained in the sequence:

$$x_1, x_2, x_3, x_4...$$

With respect to (5.6) we can write:

$$x_1 = \left(\frac{A_1}{x_2}\right)^{\frac{1}{b_1}} = \left(\frac{A_1}{\left(\frac{A_2}{x_3}\right)^{\frac{1}{b_2}}}\right)^{\frac{1}{b_1}} = ... \qquad (5.7)$$

This expression corresponds to the logarithmic polynomial:

$$\log x_1 = \frac{1}{b_1} \log A_1 - \frac{1}{b_1 b_2} \log A_2 + \frac{1}{b_1 b_2} \log x_3..., \qquad (5.8)$$

where an arbitrary member i of this polynomial is

$$\frac{1}{b_1 b_2 ... b_i} \log A_i$$

(which is positive for odd i and negative for even i) and the last member $i = m$ of this polynomial is

$$\frac{1}{b_1 b_2 ... b_m} \log x_{m+1}$$

(which is positive for even m and negative for odd m), where $i = 1, 2,..., m$.

The reason for which this complicated structure of parameters is presented here consists in the convincingly demonstrated fact that:

1. each pair of parameters A_i and b_i represents their specific combination;
2. each b_i appears to be an approximation from a chain of the type $b_1.b_2...b_i...$
3. each two neighbouring levels i and $i + 1$ of the polynomial (5.7) need not inevitably represent two neighbouring levels in a linguistic sense.

In our empirical argumentation two neighbouring levels are characterized by parameters with values which are only their estimates; they change when a new level is inserted between the two former neighbours. The scheme of linguistic levels in an arbitrary form is nothing but a classification of language units. Any classification represents a relationship between the classifier and his knowledge about linguistic units and their relations. This is, for example, the case of the sentence aggregates.

When we assume an imaginary ultimate level x_1 we assume that it is probably the level of meanings embodied in a semantic system in the human mind. When the relation of a language construct and its constituents is tested, their respective polynomial is, according to the MA law, defined by the relation:

$$\log x_m = \frac{1}{b_m} \log A_m - \frac{1}{b_m} \log x_{m+1} \qquad (5.9)$$

When the two supposed levels are really neighbouring, then it is evident that their parameters are affected by the levels in their surrounding. All the levels in languages seek their mutual equilibrium and this seeking is of a stochastic character exploiting specific language laws.

Feder (1991, 184) defines an *affine transformation* of a point $x = (x_1, ..., x_E)$ into the point with coordinates $x' = (r_1x_1, ..., r_Ex_E)$, where not all the coefficients of similarity $r_1, ..., r_E$ are equal. Consequently, the system of language levels (including all text levels) appears to be a particular sort of affine structure in which the parameters on an arbitrary level are able to characterize constructs and constituents on the levels not being explicitly taken into account as levels m and $m + 1$, cf. (5.9). This sort of scaling contained in levels deserves a deeper analysis. It is also described by the concept of *scaling*; from the linguistic viewpoint this concept is analyzed by J. Králík (1993).

5.3 RELATIONS OF LEVELS OBSERVED

Whenever we prepare an experiment concerning neighbouring levels, the fact mentioned above must be stressed again: we cannot list the final set of language levels. The following demonstration of the relations between sentences, words, syllables and phonemes trying to aim at some wider characterization of their mutual relationships does not mean, and cannot mean, that other levels are not involved.

We tried to prove equation (5.8) with data obtained from Text 1, where:

x_1 is sentence length in number of words;
y_1 is word length in number of syllables;
x_2 is word length in number of syllables;
y_2 is syllable length in number of phonemes.

In accordance with the identities of levels in their string presented above we can put $y_1 = x_2$ and $y_2 = x_3$. The observed data for this experiment are presented in TABLE 5.1 and TABLE 5.2.

TABLE 5.1: Sentence length and word length

Text 1

x_1	z_1	Total words	Total syllables	x_2	X_2
1	6	6	19	3.17	3.04
2	12	24	66	2.75	2.97
3	13	39	113	2.90	2.93
4	22	88	264	3.00	2.90
5	12	60	177	2.95	2.88
6	16	96	276	2.86	2.86
7	11	77	214	2.78	2.84
8	7	56	161	2.86	2.83
...	...	...	...	...	...
Σ	128	966	2845	-	-

A_1 = 3.0435; b_1 = -0.0348; Wilcoxon T = 11.5 > $T_{0.05}(8)$ = 4.
$E(x_1)$ = 966/128 = 7.55.

x_1 = sentence length in number of words;
z_1 = the number of sentences having the respective length x_1;
x_2 = word length in number of syllables;
X_2 = the expected word length calculated according to the estimates of A_1 and b_1 and the MA law.

TABLE 5.2: Word length and syllable length

Text 1

x_2	z_2	Total syllables	Total phonemes	x_3	X_3
1	90	90	236	2.62	2.71
2	283	566	1340	2.37	2.45
3	292	876	2045	2.33	2.31
4	213	847	1955	2.31	2.22
5	66	330	751	2.28	2.15
6	18	108	262	2.43	2.09
7	4	28	45	1.61	2.05
Σ	966	2845	6634	-	-

A_2 = 2.7101; b_2 = -0.1443; Wilcoxon T = 12.5 > $T_{0.05}(7)$ = 2.
$E(x_2) = 2845/966 = 2.95$; $E(x_3) = 2.3318$.

x_2 = word length in number of syllables;
z_2 = the number of words having the respective length x_2;
x_3 = mean syllable length in number of phonemes;
X_3 = the expected word length calculated according to the estimates of A_2 and b_2 and the MA law.

With

$x_1 = -0.0348$ $A_1 = 3.04$ $x_1 = 7.55$
$b_2 = -0.1443$ $A_2 = 2.71$ $x_3 = 2.3318$

we obtain the polynomial:

$$\log 7.55 \doteq \frac{1}{0.0348} \log 3.04 - \frac{1}{0.0348 \,.\, 0.1443} \log 2.71 +$$

$$+ \frac{1}{0.0348 \,.\, 0.1443} \log 2.3318 = 7.52$$

The data are very sensitive to approximations even in the higher decimal places; therefore the result is only approximate. Let us note that in (5.8) the values of parameter *b* must be substituted by the respective values in their positive form as it follows from (5.5).

An analogical experiment can be made with aggregates, sentences, words and syllables. The data obtained from Text 1 are presented in Table 3.1.1 and in Table 5.1. In this experiment the symbols mean:

x_1 the mean length of aggregates in number of sentences;
$y_1 = x_2$ the mean sentence length of aggregates expressed in number of words;
x_3 the mean word length in number of syllables.

The values of the respective variables and parameters are:

$b_1 = -0.0459$ $A_1 = 14.90$ $x_1 = 2.01$
$b_2 = -0.0348$ $A_2 = 3.04$ $x_3 = 2.95$

With the observed mean $x_3 = 2.95$ we did not obtain a satisfying result. When, however, this result is changed to $x_3 = 2.7703$, the result is better. It can be supposed that the estimate of the mean word length in number of syllables is afflicted with a (not very high) error. The polynomial corresponding to (5.8) is as follows:

$$\log 2.01 \doteq \frac{1}{0.0459} \log 14.90 - \frac{1}{0.0459 \,.\, 0.0348} \log 3.04 +$$

$$+ \frac{1}{0.0459 \,.\, 0.0348} \log 2.7703 =$$

$$= 25.56 - 302.3 + 277.0433 = 0.3033 = \log 2.01$$

Let us stress that all the observed values are at least influenced by an error originating from approximations of the relative values and parameters obtained from these values.

Finally, the two preceding experiments with Text 1 can be combined together, so that a slightly longer string of levels is obtained. Our analysis contains the following values:

x_1 the length of aggregates in number of sentences;
x_2 the length of sentences in number of words;
x_3 word length in number of syllables;
x_4 the length of syllables in number of phonemes.

The values to be substituted for the quantities of formula (5.8) are:

$A_1 = 14.90$	$b_1 = -0.0459$	$x_1 = 2.01$
$A_2 = 3.04$	$b_2 = -0.0348$	$x_4 = 2.3318$
$A_3 = 2.71$	$b_3 = -0.1443$	

If the value of x_4 is corrected from the value 2.3318 to the value 2.339440835 (and this correction represents the change of the approximative value 2.33 to 2.34), the values of the polynomial corresponding to (5.8) are:

$$\log 2.01 \doteq 25.56 - 302.3 + 1878.45 - 1601.40 \doteq$$
$$\doteq 0.3032 \doteq \log 2.01$$

With regard to the sensitivity of the values, the results of this testing can be taken as an approval of the structure derived with the help of the MA law. It is evident that this law formulates the principle of harmony proper to the language and text levels. Naturally, these results must be proved also on other languages.

5.4 THE DEEPNESS OF TEXT STRUCTURE

Text structure (or text system) naturally cannot be fully understood with the help of the knowledge of mutual relations existing between and among text levels. We want to stress that text structure is so deep that it is in fact infinite. It cannot be supposed that sometime we will say that now we know everything about texts and we must turn our attention to another topic. This cannot turn out to be true. If text is supposed as one unity together with the respective communicators it changes into an infinite totality, into a universe of information carried by text structure or text structures.

Now the question arises as to whether a general theory of text can be constructed. Can we expect that sometime in the future such a theory will be formulated? As in other sciences, each future general theory will be surpassed by a more general theory. Theoretical linguistics will make progress and each general theory will be replaced by a more general treatment of text and language.

Let us stress our opinion that the MA theory together with all its consequences represents the contemporaneous general theory of these phenomena. If it is so, then each new theory should seek its relation to this general theory with the purpose of supplementing or replacing it by something more general. Nevertheless, the MA theory is not a barrier for investigation thanks to the infiniteness of language and text structures. Let us indicate some topics recently appearing in text linguistics proving the correctness of our opinion.

As an example, let us mention the study of text dynamics based on indices such as the verb-adjective-ratio or type-token-ratio enabling us to model surprising regularities and mathematical functions emerging from texts, see R. Köhler & M. Galle (1993). Different analyses of frequency distributions enable us to open doors the existence of which was until now unknown; see R.J. Chitashvili & R.H. Baayen (1993). Text can be observed from aspects originating in practical purposes; see, for example, the studies by J. Tuldava (1993a, b)

concerning text readability and difficulty. All these studies testify that the science having text as its object of investigation works on a very large, possibly infinite, field.

A very instructive example of this property is the study of the relation between word length and word frequency done, with a lot of inventiveness by Rüdiger Grotjahn (1982) on the text level. He starts with an observation made by G.K. Zipf (1965): The more frequent a word, the lower its average length. The correctness of this law was proved by Zipf for many languages. And it appeared that it is not that frequency is determined by word length, but that word length is determined through word frequency. Thus such a simple characteristic as word length has social and cultural connotations. On the other hand, the principles of synergy and self-regulation work on relationships between the language subsystems which induce the assumption that the form of mutual dependence is not simple and that the relations between these variables are reciprocal. In the paper quoted, R. Grotjahn writes:

> 'Die Wortfrequenz hängt auf der Textebene wiederum von Variablen wie Kommunikationsintention, Thematik oder intendierter Leser ab. Das kann dazu führen, daß in der Sprache seltene und damit lange Wörter in einem Text überdurchschnittlich häufig verwendet werden.'

Grotjahn presents a critical analysis of the so-called Fucks' model in the work quoted above; see Fucks (1955a, b). This model is based on the Poisson distribution. Grotjahn discovered that it is not adequate for data obtained from texts. He interpreted its parameter as a variable with gamma distribution. Thus Grotjahn obtained the negative binomial distribution. It is, however, the MA law which offers an explanation of consequences for the relation between word length and word frequency in texts. (According to a personal communication, a larger study of the same problem is in preparation by an international group of linguists coordinated by K.-H. Best.)

All this is evidence for the inexhaustible richness of text structures explaining the applicability and flexibility of natural languages to all situations in which human beings and their communities find themselves. In the following sections we try to apply certain ideas testifying to this assertion.

5.5 *R/S* AND THE HURST LAW

One important problem has not been discussed as yet in connection with sentence aggregates, namely, the problem of their location in the text. A member of a set called *aggregate* can be placed anywhere in the sequence of sentences of a text. From a certain viewpoint this assertion is correct; however, it is not correct in general as we cannot arbitrarily change the sequence of sentences. Such a change evidently destroys a given text and another text or a non-text is thus obtained.

Let us suppose an arbitrary aggregate as a text unit, the position of which is fixed by the first occurrence of the lexical unit defining the supposed aggregate. The sentence in which a lexical unit occurs for the first time is taken as an indication of the place in which the respective aggregate occurs in text. It is nothing but a first occurrence of a lexical unit in the text.

Text structure can be imagined as a flow of structural parts which can be described by measurements made at certain time intervals or after a certain number of some structural parts accede to the increasing text. In the following demonstration of the ideas presented, a new measurement is made after each sentence.

A text to be produced represents for a text producer at a time $t = 0$ some vessel or reservoir which should be filled by a liquid or water. This ideal model has been deeply studied on natural water reservoirs by H.E. Hurst (1951, 1965). We try to explain the principles of the Hurst empirical law in accordance with J. Feder (1991, 151 f.).

Hurst's method is called the *method of normalized range* or the *R/S method.* The task solved by H.E. Hurst is the measurement of the outflow of a lake as a function of time. The task consists of achieving an optimal amount of the reservoir according to the measured outflow from the lake. At each year t such a reservoir receives inflow $\xi(t)$; the regulated outflow is represented by the mean value $\langle\xi\rangle$. Thus the mean value of inflow during τ years equals

$$\langle \xi \rangle_\tau = \frac{1}{\tau} \sum_{t=1}^{\tau} \xi(t).$$

The *cumulated deviation* from the mean influx is

$$X(t,\tau) = \sum_{u=1}^{t} [\xi(u) - \langle\xi\rangle_\tau].$$

The difference between the maximal and minimal X is the *range* R. It represents the extensiveness needed for a mean outflow at a given interval:

$$R(\tau) = \max_{1\le t\le\tau} X(t,\tau) - \min_{1\le t\le\tau} X(t,\tau)$$

It is evident that R is a function increasing with τ. Hurst indicated that the dimensionless relation R/S, where S is the standard deviation, can be used for the comparison of ranges of different phenomena:

$$S = \left(\frac{1}{\tau}\sum_{t=1}^{\tau} [\xi(t) - \langle\xi\rangle_\tau]^2\right)^{1/2}$$

Hurst discovered that the relation R/S suits the description of many time series when the following empirical relation is used:

$$R/S = (\tau/2)^H$$

Hurst discovered that in different natural processes the values of H are symmetrically distributed around the value 0.73.

Now let us return to text. We can state that any text is like a lake into which the language units of different levels flow. Or we can - and this is more interesting - take data observed in texts as outflow from a reservoir existing in the text producer's head. Language units appearing as language constructs signal a certain outflow from that reservoir. This means that different units are coming from one and the same reservoir and their occurrence in the measurements as time series should indicate the same range of that "lake".

We will use Hurst's notation with several modifications. We measure texts after each sentence, so that instead of t we write here i. The mean value of the data obtained is computed for entire texts, so that instead of τ we write now k , as is usual in the present work in the sense of text length (in number of sentences). In the texts analyzed, the time series (or better "sentence series") indicate the number of phonemes (p), syllables (s),

morphemes (m) and words (n) flowing out of the producer's head with the ith sentence. Together with these variables, the number of sentence aggregates introduced into the text by the ith sentence was also ascertained.

For a relatively great number of tables, the data observed in two Turkish texts are presented in the Appendix at the end of this book (see Table I and Table II). In TABLE 5.5 the observed data for the variables mentioned above in these two texts are presented.

TABLE 5.5: Indices and other data for the Hurst analysis applied to two Turkish texts

Index	Text I	Text II
$R(k)_p$	216	175
S_p	34.18936	45.39970
E_p	48.53061	87.79487
$(R/S)_p$	6.31776	3.77158
H_p	0.40205	0.44690
$R(k)_s$	92	76
S_s	14.43043	20.26598
E_s	20.55102	37.53846
$(R/S)_s$	6.37542	3.75013
H_s	0.40403	0.44498
$R(k)_m$	78	65
S_m	12.36990	17.69067
E_m	17.46939	32.41026
$(R/S)_m$	6.30563	3.67425
H_m	0.40163	0.43810
$R(k)_n$	31	23
S_n	4.69161	6.36569
E_n	6.73469	12.20513
$(R/S)_n$	6.60753	3.61312
H_n	0.41183	0.43245
$R(k)_a$	23	15
S_a	2.85748	3.68981
E_a	4.55612	6.97436
$(R/S)_a$	8.04904	4.06524
H_a	0.45487	0.47215
E(H)	0.41488	0.44692

p = phonemes, s = syllables, m = morphemes, n = sentence length in words,
a = aggregates (number of new lexical units in the *i*'th sentence); E = the mean value of the variable indicated by the respective index or argument.

The other indices are described in the present section.
The data of Text I and II are presented in Appendix, Tables I and II.

The values of H presented above indicate that the results are close to each other regardless of the measured level, regardless of the texts being from different authors and different functional styles. While in natural processes the values of H are symmetrically distributed around the value 0.73, the hypothesis can be stated that in Turkish texts the respective values of H are symmetrically distributed around the value 0.43. This property deserves a deeper analysis in many texts of many languages. With reference to the relation derived by Feder (1991, 185) for self-affine curves as

$$D = 2 - H,$$

we can expect that the similarity dimension of the system formed by the language levels in Turkish texts (if the condition of self-similarity is fulfilled) is

$$D = 2 - 0.43 = 1.57.$$

Hurst's index H is thus a means for obtaining the value of the dimension D in self-affine systems.

6. Strategy in Text

It can be assumed that when a text is produced, its producer has in mind a certain final appearance of the text. We can ascertain that there is a functional relation between the beginning of a text and its total structure. This relation is of a complicated nature. The same is valid when we suppose a text at its arbitrary phase of origination and an arbitrary subsequent phase. The text structures supposed in the process of their production are mutually related. A structure (language items, the organization of their constituents, or simply texture) occurring at a place in a text enables a more satisfactory way to some final structure which is the strategic aim of the producer. These relations can be characterized as the producer's intent. It need not be discernible even for the text producer at the interval of production. Sometimes the final appearance of a text seems to be fabricated step-by-step and the final structure, or final organization of the language phenomena, becomes existent only at its end; similar results also play the role of producer's intent. The way leading from state to state represents a communicative intent and can be denoted as strategy. In order to exclude from our suppositions the parts of text which are not sentences or parts of sentences, let us suppose that these states are reached or fulfilled only at the end of an arbitrary sentence in their sequence in a text.

The same process occurs also on the receiver's side of the communication channel. The receiver is active as an interpreter, i.e. as a producer of an interpretative text. This new text can be long, longer than the interpreted text, or short, possibly also extremely short when it contains a semantic equivalent of one sentence only: 'This does not interest me.' Two different receivers can interpret the same texture in different ways, as has been stressed in the preceding chapters. An interpretative text also has its intent (or strategy). The process of text construction and its reception is full of mutual connections which can be characterized as

producer's or receiver's intent. Any model operating with this concept is sufficiently general; this concept is evidently applicable to anybody who is in communicative contact with a text.

The idea of strategy in connection with texts is not new. Let us mention at least the concept of process-grammar in which strategy has a significant function. This concept was applied to the process of text comprehension by T.A. van Dijk (1980, 1985) and by T.A. van Dijk & W. Kintsch (1983). In these works strategy is comprehended as a conceptual complex connected with other subordinated conceptual complexes, so that the whole reality investigated is framed by certain axiomatic expressions. They are not, however, supplemented by a theoretical structure containing a testable hypothesis. The epistemological process of this investigation is not a movement from assumptions to rejectable consequences. This model remains on the same level where it has been placed by a set of definitions.

The mathematical theory of games presents an exact scientific treatment of the concept of strategy; see, for example, J. von Neumann & O. Morgenstern (1953). In this theory, strategy is an element of the set of acts which can be selected by a player in a step of a game. As far as we know, this remarkable idea has not as yet been applied in linguistics; the problem consists in the complicated structure of the language strategies. In the case of a "text game" these strategies represent sets which are too large, if not infinite. For this reason, in the present work, strategy is simply understood as a synonym of 'intent' and we do not try to define it in a more sophisticated way.

This all means that text is a process moving from a certain point to another point carried by a sequence of language structures.

6.1 SENTENCE LENGTH IN A TEXT STRATEGY

An analysis of a text taken as a process should reflect the fact that the object of analysis represents a special transcription of structures. The construct written on a paper or tape should reveal relations, impulses or processes which put their residues into a written form. Quantitative characteristics are then transcriptions of these events observed on a paper or tape. The values of these characteristics change as the respective text increases. During this increase the successive states and then the final state of the text are reached. While the text is growing, the values of a characteristic change, and we can analyze them at an arbitrary state of the

analyzed text in the same way as at its final state. Constants characterizing the whole text thus become variables describing it as a process. The producer's (or receiver's) strategy consists in the selection of the path or trajectory going through certain values of constants and variables. Each producer (or receiver) produces a given text (or text interpretation) with the intention of achieving at the end a certain value of a variable.

Sentence length is selected in order to examine the model described by testing the preliminary ideas with which the model is constructed. It enables us to obtain data which can be simply handled in statistical experiments. Another reason consists in the relevance of sentence length in the supra-sentence text level, i.e. in sentence aggregates. Different statistical characteristics can be analyzed in a similar way and used for testing text as an increasing phenomenon.

Suppose two means (in number of words) characterizing sentence length in a text:

1. the mean sentence length of the whole text (*finite mean*);
2. the mean sentence length of the increasing text measured at the end of each sentence (*sequential mean*).

Using the same symbols as in the preceding chapters we can define:

$$n = \sum_{i=1}^{k} n_i, \qquad i = (1, 2, \ldots, k),$$

where n_i = the length (in words) of the ith sentence;
n = the length of the whole text in number of words;
k = the text length in number of sentences.

Consequently, i is the rank number of a sentence and the maximal value of i is k which is the rank number of the last sentence.

Let us introduce the symbol N_i as the cumulative text length (in number of words) of that part of a text beginning at the first sentence (inclusive) and ending at the ith sentence. Further, let us introduce two deviations:

1. the deviation of a sentence length from the finite mean

$$D_i = n_i - \frac{n}{k};$$

2. the deviation of a sentence length from the sequential mean

$$d_i = n_i - \frac{N_i}{i}.$$

The difference between the two deviations is:

$$D_i - d_i = (n_i - \frac{n}{k}) - (n_i - \frac{N_i}{i}) =$$

$$= \frac{N_i}{i} - \frac{n}{k} \qquad (6.1)$$

We can write the following progression:

$D_1 = n_1 - n/k$
$D_2 = n_2 - n/k$
...
$D_i = n_i - n/k$

The sum of this progression is:

$$\sum_i D_i = N_i - i\frac{n}{k} \qquad (6.2)$$

It can be easily proved that the mean value of D_i equals the difference (6.1):

$$\frac{1}{i}\sum_i D_i = \frac{N_i}{i} - \frac{n}{k} = D_i - d_i \qquad (6.3)$$

Thus the mean of deviation D_i can be estimated as the difference of the sequential and final means. Each text reaching its end also reaches its strategic aim. This is the way in which the variables are defined. If $i = k$, from (6.2) it follows that the sum of D_i equals zero:

$$\sum_{i=1}^{k} D_i = n - k\frac{n}{k} = 0 \qquad (6.4)$$

6.2 THE INCREASE OF SEQUENTIAL MEAN

The increase of sequential mean r_i can be defined in the following way:

$$r_i = \frac{N_i}{i} - \frac{N_{i-1}}{i-1} =$$

$$= \frac{i(N_i - N_{i-1}) - N_i}{i(i-1)}, \qquad i > 1. \qquad (6.5)$$

As evidently $N_i - N_{i-1} = n_i$, we can write:

$$r_i = \frac{i\,n_i - N_i}{i(i-1)} = \frac{1}{i-1}\left(n_i - \frac{N_i}{i}\right) =$$

$$= \frac{1}{i-1}\,d_i, \qquad i > 1. \qquad (6.6)$$

Let us introduce the variable:

$$R_i = \sum_i r_i, \qquad i \neq 1.$$

The following progression can be derived on the basis of (6.5) and (6.6):

$r_2 = N_2/2 - N_1/1 = N_2/2 - n_1;$	consequently	$N_2/2 = r_2 + n_1.$
$r_3 = N_3/3 - N_2/2 = N_3/3 - (r_2 + n_1);$	consequently	$N_3/3 = r_3 + r_2 + n_1.$
$r_4 = N_4/4 - N_3/3 = N_4/4 - (r_3 + r_2 + n_1);$	consequently	$N_4/4 = r_4 + r_3 + r_2 + n_1.$
... ...	consequently	$N_i/i = \sum_i r_i + n_1 = R_i + n_1.$ (6.7)

For the whole text with $i = k$ it holds that

$$\frac{n}{k} = R_k + n_1 \qquad (6.8)$$

The last two formulas enable us to characterize the sentence-length text strategy as two choices determining the process of reaching the strategic aim at the end of a text. Primarily, the length of the first sentence is selected and then the sum of increases of the sentence length. It seems, however, to be more adequate to assume the second choice as the selection of the mean sentence length which will be obtained at the end of the produced text. When this strategy is selected at the beginning, the following principle can be formulated:

> The sum of increases equals the difference of the mean value of the tested variable and the value of the same characteristics proper to the first sentence.

This is the verbal expression contained in (6.7). The strategic process of a text production, from the viewpoint of sentence length, is determined by two constants: n_1 and n/k.

6.3 AN APPLICATION TO TEXT

Texts I and II in the Appendix were statistically analyzed from the viewpoint discussed above. The observed values of the related variables are presented in Tables III and IV in Appendix. Some interesting properties of the variables emerge from this experiment. It becomes evident that the final sum of increases R_k in a certain sense is present already at the first sentence of each text; see the absolute value of $|R_k| = 3.27$ in Table III, which equals the difference $N/i - n/k$ for $i = 1$. This is the consequence of the idea of strategy. In Table IV (Appendix) the same value equals 2.21.

According to our opinion, this idea should not be rejected right away. Sentence length is a characteristic firmly tied together with the semantic level of a text, as is testified by the theory of aggregates; it is also connected with the state of the individual semantic systems of language users.

In connection with (6.7) and (6.8), let us return to Section 3.8 where a modification of the MA law was derived in formula (3.6). It defines the mean sentence length in such a (theoretically assumed) aggregate which is based on a lexical unit occurring in each of k sentences of a text. From (3.7) it is obvious that the mean sentence length is closely connected with the structure of sentence aggregates and thus also with the lexical structure of a text.

If we can take as valid that

$$y_k = A\, k^b = \frac{n}{k},$$

then with respect to (6.8) it holds that

$$A\, k^b = n_1 + R_k. \qquad (6.9)$$

For each sub-text of an increasing text it can also be written:

$$A\, i^b = n_1 + R_i \qquad (6.10)$$

Here Altmann's b should be written as b_i and its value estimated as follows:

$$b_i = \frac{\log (n_1 + R_i) - \log A}{\log i}, \quad i = (1,2,...,k). \qquad (6.11)$$

The values of b_i obtained from Text 2 (see also Tables I and III in the Appendix) are presented in TABLE 6.1. The values of b_i observed in Text 9 (see also Tables II and IV in the Appendix) are presented in TABLE 6.2.

The final values of b are not identical with those presented in Tables 3.1.2 and 3.1.9. The approach to their estimation differs a little, especially in the definition of word forms; the difference, however, is not substantial. The observations presented above demonstrate one very interesting general property of text:

If we analyze the sequence of values A_i in both texts, see Tables 6.1 and 6.2, it is evident that there is a great dissimilarity in their course. This is evident from FIGURE 6.1 and FIGURE 6.3. On the other hand, the respective values of b_i of both these texts exhibit one common quality: both these curves descend in a way making them similar to the MA curve; see FIGURE 6.3 and FIGURE 6.4.

TABLE 6.1: Altmann's b_i as a variable of an increasing text (see also Tables I and III in the Appendix)

Text 2

i	10	20	30	40	50	60
A_i	9.57	12.07	11.84	12.47	12.09	11.73
b_i	-0.01249	-0.05126	-0.07844	-0.08849	-0.08228	-0.08551
i	70	80	90	100	110	120
A_i	11.15	11.42	11.30	11.09	10.87	10.75
b_i	-0.07191	-0.08096	-0.08172	-0.07972	-0.08108	-0.08187
i	130	140	150	160	170	180
A_i	10.84	10.78	10.54	10.43	10.39	10.33
b_i	-0.08161	-0.08195	-0.08038	-0.08092	-0.08049	-0.07850
i	190	196				
A_i	10.22	10.16				
b_i	-0.07893	-0.07798				

TABLE 6.2: Altmann's b_i as a variable of an increasing text (see also Tables II and IV in the Appendix)

Text 9

i	5	10	15	20	25	30
A_i	6.94	10.55	11.05	10.65	12.13	12.52
b_i	0.08821	-0.01471	-0.01549	-0.03986	-0.04066	-0.02241
i	35	39				
A_i	12.26	13.88				
b_i	-0.02972	-0.03514				

We tried to analyze the values of b_i presented in Tables 6.1 and 6.2. With the purpose of obtaining positive values, we shifted the observed curves straight up along the axis b_i by a constant C; this constant can obtain an arbitrary positive value, in this case $C = 1$. These shifted curves were fitted by the MA curve with $b^* = -0.01495$ in the case of Text 2 and $b^* = -0.04936$ in the case of Text 9. The calculated theoretical values were greater by 1 than those of the original curves. Analogous experiments were made with several other texts.

This finding seems to be interesting. If it is correct that text structure on all its levels (including the level of semantic structure formed by aggregates) is in agreement with the MA law , then it is also correct that this structure increases in a way which is in accordance with the same law. Text structure thus appears to be a phenomenon enfolded in itself. We can try to formulate the following principle:

> When each part of a text beginning with the first sentence and ending with an arbitrary sentence is called a sub-text of a given text, then each sub-text appears to be a constituent of all its higher sub-texts as well as of the respective total text; all these constituents and constructs are in agreement with the MA law.

This principle is presented here only as a hypothesis. Nevertheless, this hypothesis is quite well justified and there is an expectation of obtaining new results concerning text structure. In other words, this principle requires further investigation, after which it should become a firmly stated language law. One can ask, for example, whether the values of b_i are tending to a limit and what is this limit. According to the MA law, it seems to be prohibited for this value to be higher than zero or to equal zero. When this happens, text obtains a destroyed structure and can scarcely be supposed to be a text.

There are many other questions, for example, whether sub-texts taken as constituents form a smooth curve b_i and thus the fluctuations of its values have to be taken as consequences of some random events, or whether they can be interpreted as functional breaks forming a linear text structure. These and many other problems deserve solution in future experiments.

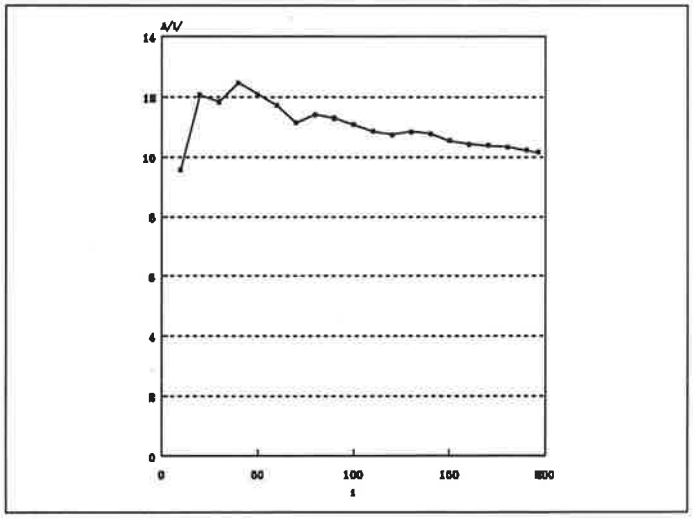

FIGURE 6.1: A_i in the increasing Text 2

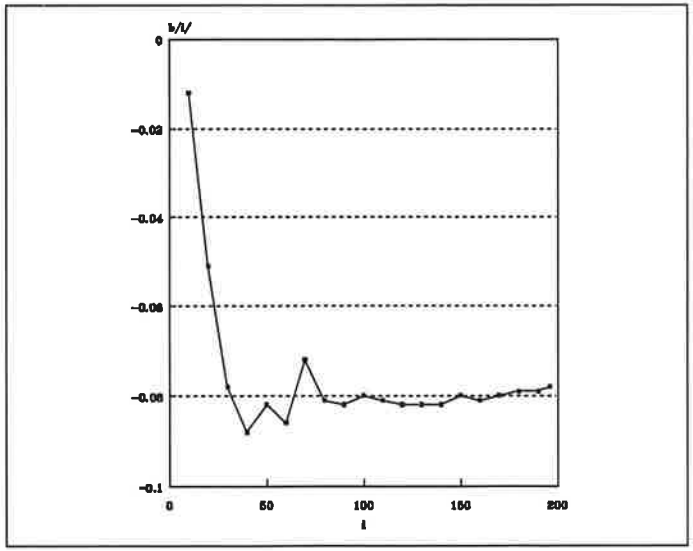

FIGURE 6.2: b_i in the increasing Text 2

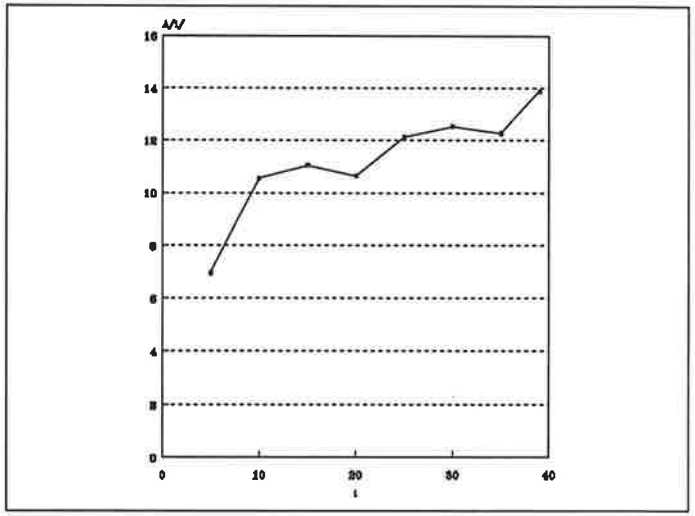

FIGURE 6.3: A_i in the increasing Text 9

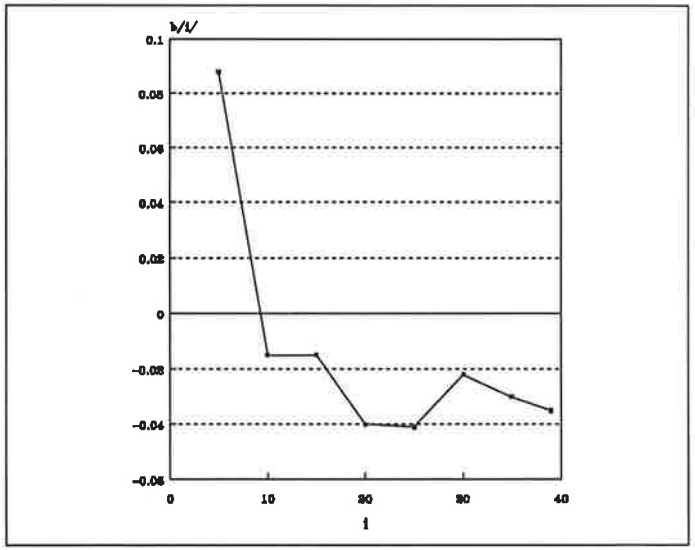

FIGURE 6.4: b_i in the increasing Text 9

Several Proposals for Further Experiments

Instead of a summary which is, in fact, presented at the beginning, let us sum up several proposals for future development of the theory and its experimental confirmation. It is believable that all the possibilities hidden in the idea of the MA law are far from being exploited. The list of proposals presented here is not definite or complete; it is limited to some consequences of the results mentioned in the preceding chapters.

1. The idea of the closest connection between text structure and the mind of a person who is in contact with it deserves a detailed investigation in linguistics. This seems to be the task of psycholinguistics in particular. The connections between the physiological carrier of the human ability to speak and create texts should soon become topical, if it is not already.

2. New linguistic levels should be investigated. This is, in fact, an act of classification and it bears all the consequences which are proper to each classification in sciences. The MA law enables us to describe levels and their strings. Sentence syntax especially, represents a space for investigation of levels. We can ask, whether, for example, the cuts between subject and predicate (or between nominal phrase and verbal phrase) represent limits between different levels. This will influence the understanding of sentence aggregates and their structure. For the present, aggregates seem to be non-structured. The future investigation of aggregates should indicate their semantic itemization.

3. A theoretical principle should be sought representing a criterion for identification of the exactly neighbouring levels. We insist on the thesis that such a criterion is not at hand as yet.

4. Increasing text, i.e. text supposed as a growing phenomenon, is tested in the present work only from the viewpoint of sentence length. We can expect that the other qualities and variables will also offer interesting results.

5. The investigation of the spoken forms of texts is very promising. This especially concerns the psycholinguistic and physiological aspects of these phenomena when they are observed from the viewpoint of different constructs and constituents.

6. The special concern with dialogue is not new in linguistics. This investigation should encompass not only an instant production and reception by participants but also their instant interpretation of the commonly produced dialogue text. Both written and spoken dialogues are worthy of scientific interest. One can ask, for example, whether the participants produce language constituents in relation to the (assumed) constructs formulated by their opposites in the dialogue.

7. According to our conviction, the examination of the abstract structures analyzed in the theory of fractals seems to be profitable for text linguistics. It must be stressed that between this theory and the theory of communication there are important ties which should be exploited by linguistics.

8. The theory of language levels in the conception of the MA law and the semantic consequences of this theory deserves a little more interest on the part of the specialists in belles-lettres. Artistic creativity based on natural languages also submits to the laws researched in linguistics.

The set of such proposals can be further enlarged. Let it be stressed that the properties of text structure tested in one language only cannot be freely transferred to other languages without experiments and testing. Text studies require empirical approaches with quantitative evaluation and testing of theoretical assumptions. This is the only way to obtain some deeper knowledge of languages and texts.

References

Alekseev, P. (1978): O nelinejnych formulirovkach zakona Zipfa. *Voprosy kibernetiki 41*, 53-65.

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

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

Altmann, G. (1992): Two models for word association data. In: *Glottometrika 13.* Bochum, Brockmeyer, 105-120.

Altmann, G. (1993): Science and linguistics. In: Köhler & Rieger (1993, 3-10).

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

Barnsley, M. (1988): *Fractals everywhere.* Boston, Academic Press.

Bunge, M. (1967): *Scientific research I.* Berlin, Springer.

Bunge, M. (1983): *Exploring the world. (Treatise on basic philosophy, Vol. 5)* Dordrecht-Boston-Lancaster, Reidel.

Chitashvili, R.J. & Baayen, R.H. (1993): Word frequency distributions. In: Hřebíček & Altmann (eds.)(1993, 54-135).

Chomsky, N. (1957/1962): *Syntactic structures.* The Hague, Mouton. [Russian edition in: *Novoe v lingvistike II.* Moskva, Izd. innostrannoj literatury, 1962, 412-527.]

Čapek, K. (1934): [A quotation from the Epilogue to his *Trilogy.*]

Dijk, T.A. van (1980): *Textwissenschaft.* Tübingen, Niemeyer.

Dijk, T.A. van (1985): Strategic discourse comprehension. In: T.T Ballmer (ed.), *Linguistic dynamics.* Berlin-New York, de Gruyter.

Dijk, T.A. van & Kintsch, W. (1983): *Strategies of discourse comprehension.* New York, Academic Press.

Dolinskij, V.A. (1988): Raspredelenie v eksperimentach po verbal'nym associacijam. *Acta et Commentationes Universitatis Tartuensis 827*, 89-101. [Quoted from Altmann 1992.]

Feder, J. (1988/1991): *Fractals.* New York, Plenum Press. [Russian edition: Moskva, Mir, 1991].

Fenk, A. & Fenk-Oczlon (1993): Menzerath's law and the constant flow of linguistic information. In: Köhler & Rieger (1993, 11-31).

Fucks, W. (1955a): *Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen.* Köln/Opladen, Westdeutscher Verlag. [Quoted from Grotjahn 1982.]

Fucks, W. (1955b): Theorie der Wortbildung. *Mathematisch-physikalische Semesterberichte 4*, 195-212. [Quoted from Grotjahn 1982.]

Furth, H.G. (1969): *Piaget and knowledge. Theoretical foundations.* Englewood Cliffs (N.J.), Prentice-Hall.

Grotjahn, R. (1982): Ein statistisches Modell für die Verteilung der Wortlänge. *Zeitschrift für Sprachwissenschaft 1*, 44-75.

Haight, F.A. (1966): Some statistical problems in connection with word association data. *Journal of Mathematical Psychology 3*, 217-233.

Haken, H. (ed.) (1973): *Synergetics. Cooperative phenomena in multi-component systems.* Stuttgart, Teubner.

Haken, H. (1978): *Synergetics. An introduction.* Heidelberg-New York, Springer.

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

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

Horvath, W.J. (1963): A stochastic model for word association tests. *Psychological Review 70*, 361-364.

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

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

Hřebíček, L. (1993): Text as a construct of aggregations. In: Köhler & Rieger (1993, 33-39).

Hřebíček, L. (1994): Fractals in language. *Journal of Quantitative Linguistics 1, 1, 82-86.*

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

Hřebíček, L. & Altmann, G. (1994): Levels of order in language. *Glottometrika 15, 28-60.*

Hurst, H.E. (1951): Long-term storage capacity of reservoirs. *Trans. Am. Soc. Civ. Eng., 116*, 770-808. [Quoted from Feder 1991, 252-253.]

Hurst, H.E. (1965): *Long-term storage: an experimental study.* London, Constable. [Quoted from Feder 1991, 253.]

Jenkins, J.J. (1964): *Word association norms. Grade school through college.* Minneapolis, University of Minesota Press.

Katz, J.J. & Postal, P.M. (1964): *An integrated theory of linguistic descriptions.* Massachusetts Institute of Technology. [Czech edition: Praha, Academia, 1967.]

Köhler, R. (1982): Das Menzerathsche Gesetz auf Satzebene. In: *Glottometrika 4*, 103-113.

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

Köhler, R. (1989): Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsmechanismus. In: *Altmann & Schwibbe (1989, 108-112).*

Köhler, R. & Galle, M. (1993): Dynamic aspects of text characteristics. In: Hřebíček & Altmann (eds.)(1993, 46-53).

Köhler, R. & Rieger, B.B. (eds.)(1993): *Contributions to quantitative linguistics.* Dordrecht-Boston-London, Kluwer.

Králík J. (1993): Probabilistic scaling of texts. In: Köhler & Rieger (eds.)(1993, 227-240).

Kubánková V. & Hendl J. (1986): *Statistika pro zdravotníky.* Praha, Avicenum.

Mandelbrot, B.B. (1982): *The fractal geometry of nature.* New York, Freeman.

Menzerath, P. (1954): *Die Architektonik des deutschen Wortschatzes.* Bonn, Dümmler.

Neumann, J. von & Morgenstern, O. (1953): *Theory of games and economic behavior.* Princeton, Princeton University Press.

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

Palermo, D.S. & Jenkins, J.J. (1964): *Word association norms. Grade school through college.* Minneapolis, University of Minnesota Press. [Quoted from Altmann 1992.]

Peitgen, H.-O. & Jürgens, H. & Saupe, D. (1992): *Fractals for the classroom.* Part one. Part two. New York, Springer.

Popper, K.R. (1963): *Conjectures and refutations. The growth of scientific knowledge.* London-Henley, Routledge and Kegan Paul.

Reisenauer, R. (1970): *Metody matematické statistiky a jejich aplikace.* Praha, SNTL.

Sandefur, J.T. (1990): *Discrete dynamical systems. Theory and applications.* Oxford, Clarendon Press.

Schwarz, C. (1992): Zur Verteilung von Aggregaten in Texten. Ruhr-Universität Bochum. [Unpublished seminar work.]

Tuldava, J. (1993a): The statistical structure of a text and its readability. In: Hřebíček & Altmann (eds.)(1993, 215-227).

Tuldava, J. (1993b): Measuring text difficulty. In: *Glottometrika 14*, 69-81.

Zipf, G.K. (1965): *The psycho-biology of language.* Cambridge (Mass.), Addison Wesley.

Corpus of Turkish Texts

Text 1 Demir Özlü, *Kanallar.* [Chapter 1, 7-13] Istanbul, Can.

Text 2 Demir Özlü, *Bir yaz mevsimi romansı.* [Chapter I: Kuzey Avrupa'da bir kahve,7-13] Istanbul, Ada, 1990.

Text 3 Demir Özlü, Cristina Nilsson'u aramak. In: D. Özlü, *Stockholm öyküleri.* Istanbul, Ada, 1988, 15-19.

Text 4 Aziz Nesin, Kelepir bir işçi. In: A. Nesin, *Geriye kalan.* Istanbul, Cem-May, 1982, 26-29.

Text 5 Serap Yılmaz, Osmanlı İmparatorluğu'nun Doğu ile ekonomik ilişkileri. *Belleten, LVI,* 215, 31-40. [Introductory section of this article.]

Text 6 Aydın Sayılı, Atatürk ve millî kültürümüzün temel unsurlarından bilim ile entellektüel kültür ve teknoloji. *Erdem 3,* 9, 1987, 609-621. [Introductory part of this text.]

Text 7 Ertuğrul Özkök, Vatandaşın gözü, Kaya Erdem'in odasında. *Hürriyet 12.3.1990,* 17.

Text 8 Maliye, Istanbul'un 50 milyarını vermiyor. *Hürriyet 13.3.1990,* 4.

Text 9 Yaşar Nabi, 1967'ye toplu bir bakış. In: *Varlık yıllığı 1968,* 5-7.

Text 10 Yaşar Nabi, 1967'ye toplu bir bakış. In: *Varlık yıllığı 1968,* 5-8. [The same text as T[Text 9 enlarged by the next section.]

Yunus Emre [poem LXXXI] In: Abdüllâh Gölpınarlı (ed.), *Yunus Emre, Risâlat al-Nushiyya ve Dîvân.* Istanbul, Garan, 1965, 81.

Cahit Tanyol, Atatürk ilkeleri. In: Yaşar Nabi (ed.), *Atatürkçülük nedir?,* Istanbul, Varlık, 1969, 105-109.

Sir James Redhouse (1811-1892). In: *Redhouse yeni Türkçe-Ingilizce sözlük.* Istanbul, Redhouse Yayınevi, 1968, X-XI.

Subject Index

Index of Names

Appendix

TABLE I: Number of phonemes (p), syllables (s), morphemes (m), words (sentence length in number of words - n) and new lexical units (v) in each *i*th sentence

Text 2

i	p	s	m	n	v
1	64	26	22	10	10
2	43	19	17	5	5
3	63	25	23	7	7
4	84	38	32	13	12
5	48	20	18	7	6
6	101	42	42	13	12
7	16	8	7	3	3
8	105	43	42	15	11
9	92	39	39	11	10
10	83	35	32	9	8
11	37	15	13	4	0
12	53	21	20	7	3
13	111	47	38	16	13
14	66	28	24	10	8
15	53	21	21	7	7
16	50	21	18	6	4
17	43	18	15	7	6
18	152	61	53	18	13
19	118	50	45	17	7
20	139	59	51	22	17
21	8	3	3	2	1
22	6	2	2	1	0
23	30	13	11	3	2
24	50	20	19	9	4

i	p	s	m	n	v
25	56	23	22	6	5
26	28	11	9	3	3
27	29	11	8	3	1
28	55	25	18	7	4
29	98	42	34	14	11
30	116	49	44	17	12
31	20	8	8	2	1
32	78	32	26	11	6
33	52	22	16	8	3
34	68	29	26	9	4
35	197	82	60	26	17
36	21	9	8	3	2
37	47	21	14	7	3
38	95	39	33	13	7
39	16	6	5	2	1
40	55	25	24	7	2
41	130	56	41	16	9
42	26	10	6	4	2
43	96	39	29	13	10
44	59	24	17	11	9
45	26	11	8	4	3
46	92	36	27	11	5
47	26	11	9	5	3
48	22	9	6	4	2
49	50	21	19	6	4
50	21	7	8	4	3
51	32	14	12	4	1
52	50	19	16	6	2
53	87	35	28	12	7
54	40	17	17	6	3
55	38	16	14	7	3

i	p	s	m	n	v
56	20	8	4	4	2
57	34	14	15	5	1
58	37	16	12	5	1
59	33	15	14	4	2
60	29	13	13	5	2
61	124	51	45	20	8
62	84	35	29	11	6
63	46	21	18	6	3
64	43	18	16	5	2
65	71	31	30	12	4
66	37	16	13	6	5
67	26	11	10	4	2
68	33	14	13	6	2
69	7	3	3	1	0
70	65	27	21	8	4
71	221	94	79	32	9
72	29	12	12	6	1
73	51	21	17	7	2
74	8	3	2	1	1
75	11	5	4	2	1
76	38	16	14	6	5
77	15	6	5	2	0
78	13	5	5	1	0
79	37	16	12	6	3
80	22	10	9	3	1
81	80	33	27	10	4
82	37	16	14	5	4
83	61	26	16	13	9
84	36	16	12	3	1
85	21	9	8	3	0
86	13	5	4	2	0

i	p	s	m	n	v
87	27	11	7	5	0
88	14	6	4	2	1
89	65	29	22	9	6
90	79	33	25	11	6
91	63	27	19	9	1
92	45	20	15	5	3
93	37	17	16	5	2
94	32	13	9	4	2
95	48	20	14	6	3
96	33	14	9	4	1
97	58	24	20	8	4
98	37	14	11	4	0
99	57	24	17	8	2
100	66	29	27	11	5
101	57	23	20	8	3
102	30	12	13	4	1
103	50	21	21	8	4
104	38	16	15	5	2
105	35	17	14	5	2
106	36	16	13	5	1
107	13	7	6	3	0
108	29	14	12	3	3
109	21	9	7	4	0
110	32	13	11	4	2
111	19	7	5	2	1
112	51	23	18	8	2
113	50	22	22	8	2
114	16	7	5	3	0
115	28	11	10	3	1
116	24	10	8	4	2
117	14	6	4	2	1

i	p	s	m	n	v
118	40	16	13	5	2
119	125	55	49	15	3
120	33	15	13	5	1
121	41	18	15	5	1
122	11	5	4	2	0
123	177	76	65	21	10
124	55	23	16	7	2
125	52	23	18	6	0
126	41	17	14	5	3
127	35	14	14	5	2
128	73	34	31	11	2
129	25	11	9	3	2
130	56	24	21	10	5
131	27	12	11	4	1
132	17	7	8	3	0
133	28	13	10	5	0
134	32	12	10	4	1
135	21	9	9	3	0
136	42	17	16	5	0
137	74	32	26	9	2
138	62	26	24	8	1
139	68	29	27	10	4
140	46	20	18	9	2
141	23	13	10	4	0
142	37	14	13	7	5
143	49	20	16	7	2
144	23	9	7	3	3
145	12	6	5	2	2
146	43	17	14	5	4
147	34	14	14	4	1
148	26	12	11	4	2

i	p	s	m	n	v
149	71	30	30	8	5
150	51	22	19	6	1
151	67	28	25	11	2
152	22	10	7	3	0
153	40	17	14	5	2
154	15	7	4	3	1
155	66	29	27	9	2
156	7	3	2	1	0
157	15	6	4	3	0
158	45	20	18	5	1
159	50	22	17	5	3
160	37	15	13	5	3
161	67	30	25	9	2
162	48	20	17	8	7
163	14	6	4	2	0
164	102	44	41	15	2
165	5	2	1	1	1
166	5	2	1	1	0
167	11	5	5	2	1
168	99	40	38	12	4
169	64	29	23	9	2
170	11	5	5	2	0
171	25	10	6	3	1
172	68	28	24	9	3
173	38	16	12	6	1
174	34	15	11	5	0
175	59	26	23	7	5
176	73	31	26	9	5
177	37	16	12	5	1
178	92	40	31	13	4
179	92	39	35	11	2

i	p	s	m	n	v
180	10	4	4	1	0
181	14	5	4	2	1
182	26	10	10	3	2
183	41	18	16	5	2
184	35	15	13	4	2
185	71	30	27	10	4
186	12	5	4	2	0
187	19	8	9	3	1
188	10	4	3	1	0
189	82	33	29	10	3
190	49	21	21	6	3
191	44	19	15	5	2
192	20	9	10	3	0
193	57	25	22	8	2
194	63	29	20	11	3
195	27	12	12	3	2
196	47	21	20	7	0
Σ	9512	4028	3424	1320	616

TABLE II: Number of phonemes (p), syllables (s), morphemes (m), words (sentence length - n) and new lexical units (v) in each *i*th sentence

Text 9

i	p	s	m	n	v
1	66	27	25	10	10
2	96	41	43	14	10
3	40	15	14	5	5
4	44	19	12	7	5
5	37	15	14	4	4
6	109	49	34	17	11
7	76	33	27	11	7
8	46	19	18	7	4
9	94	38	33	13	13
10	94	40	32	14	9
11	85	38	33	12	8
12	76	34	30	12	9
13	90	38	34	11	6
14	105	44	41	15	8
15	47	21	18	7	2
16	94	38	35	12	7
17	29	12	11	3	3
18	35	15	12	4	2
19	47	20	20	8	4
20	24	10	8	3	1
21	100	41	39	13	5
22	66	28	26	9	8
23	104	48	37	16	11
24	155	65	62	22	14

i	p	s	m	n	v
25	115	48	44	17	6
26	56	23	22	7	6
27	141	60	54	20	7
28	126	56	43	18	10
29	189	84	69	26	13
30	87	36	30	11	8
31	50	22	16	8	1
32	31	13	9	5	4
33	43	18	13	5	4
34	123	52	42	14	7
35	37	16	14	6	1
36	166	72	59	22	10
37	199	86	73	26	16
38	124	54	48	18	5
39	178	76	70	24	9
Σ	3424	1464	1264	476	271

TABLE III: Sentence-length-strategy (see also Table I in Appendix)

Text 2

i	n_i	N_i	N_i/i	D_i	d_i	r_i	R_i	$N_i/i-n/k$
1	10	10	10.00	3.27	0.00	-	-	3.27
2	5	15	7.50	-1.73	-2.50	-2.50	-2.50	0.77
3	7	22	7.33	0.27	-0.33	-0.17	-2.67	0.60
4	13	35	8.75	6.27	4.25	1.42	-1.25	2.02
5	7	42	8.40	0.27	-1.40	-0.35	-1.60	1.67
6	13	55	9.17	6.27	3.83	0.77	-0.83	2.44
7	3	58	8.29	-3.73	-5.29	-0.88	-1.71	1.56
8	15	73	9.13	8.27	5.87	0.84	-0.87	2.40
9	11	84	9.33	4.27	1.67	0.20	-0.67	2.60
10	9	93	9.30	2.27	-0.30	-0.30	-0.70	2.57
11	4	97	8.82	-2.73	-4.82	-0.48	-1.18	2.09
12	7	104	8.67	0.27	-1.67	-0.15	-1.33	1.94
13	16	120	9.23	9.27	6.77	0.56	-0.77	2.50
14	10	130	9.29	3.27	0.71	0.06	-0.71	2.56
15	7	137	9.13	0.27	-2.13	-0.16	-0.87	2.40
16	6	143	8.94	-0.73	-2.94	-0.19	-1.06	2.21
17	7	150	8.82	0.27	-1.82	-0.12	-1.18	2.09
18	18	168	9.33	11.27	8.67	0.51	-0.67	2.60
19	17	185	9.74	10.27	7.26	0.41	-0.26	3.01
20	22	207	10.35	15.27	11.65	0.61	0.35	3.27
21	2	209	9.95	-4.73	-7.95	-0.40	-0.05	3.22
22	1	210	9.55	-5.73	-8.55	-0.40	-0.45	2.82
23	3	213	9.26	-3.73	-6.26	-0.29	-0.74	2.53
24	9	222	9.25	2.27	-0.25	-0.01	-0.75	2.52
25	6	228	9.12	-0.73	-3.12	-0.13	-0.88	2.39
26	3	231	8.88	-3.73	-5.88	-0.24	-1.12	2.15
27	3	234	8.67	-3.73	-5.67	-0.21	-1.33	1.94
28	7	241	8.61	0.27	-1.61	-0.06	-1.39	1.88

i	n_i	N_i	N_i/i	D_i	d_i	r_i	R_i	$N_i/i-n/k$
29	14	255	8.79	7.27	5.21	0.18	-1.21	2.06
30	17	272	9.07	10.27	7.93	0.28	-0.93	2.34
31	2	274	8.84	-4.73	-6.84	-0.23	-1.16	2.11
32	11	285	8.91	4.27	2.09	0.07	-1.09	2.18
33	8	293	8.88	1.27	-0.88	-0.03	-1.12	2.15
34	9	302	8.88	2.27	0.12	0.00	-1.12	2.15
35	26	328	9.37	19.27	16.63	0.49	-0.63	2.64
36	3	331	9.19	-3.73	-6.19	-0.18	-0.81	2.46
37	7	338	9.14	0.27	-2.14	-0.05	-0.86	2.41
38	13	351	9.24	6.27	3.76	0.10	-0.76	2.51
39	2	353	9.05	-4.73	-7.05	-0.19	-0.95	2.32
40	7	360	9.00	0.27	-2.00	-0.05	-1.00	2.27
41	16	376	9.17	9.27	6.83	0.17	-0.83	2.44
42	4	380	9.05	-2.73	-5.05	-0.12	-0.95	2.32
43	13	393	9.14	6.27	3.86	0.09	-0.86	2.41
44	11	404	9.18	4.27	1.82	0.04	-0.82	2.45
45	4	408	9.07	-2.73	-5.07	-0.11	-0.93	2.34
46	11	419	9.11	4.27	1.89	0.04	-0.89	2.38
47	5	424	9.02	-1.73	-4.02	-0.09	-.98	2.29
48	4	428	8.92	-2.73	-4.92	-0.10	-1.08	2.19
49	6	434	8.86	-0.73	-2.86	-0.06	-1.14	2.13
50	4	438	8.76	-2.73	-4.76	-0.10	-1.24	2.03
51	4	442	8.67	-2.73	-4.67	-0.09	-1.33	1.94
52	6	448	8.62	-0.73	-2.62	-0.05	-1.38	1.89
53	13	460	8.68	5.27	3.32	0.06	-1.32	1.95
54	6	466	8.63	-0.73	-2.63	-0.05	-1.37	1.90
55	7	473	8.60	0.27	-1.60	-0.03	-1.40	1.87
56	4	477	8.52	-2.73	-4.52	-0.08	-1.48	1.79
57	5	482	8.46	-1.73	-3.46	-0.06	-1.54	1.73
58	5	487	8.40	-1.73	-3.40	-0.06	-1.60	1.67
59	4	491	8.32	-2.73	-4.32	-0.08	-1.68	1.59
60	5	496	8.27	-1.73	-3.27	-0.05	-1.73	1.54
61	20	516	8.46	13.27	11.54	0.19	-1.54	1.73
62	11	527	8.50	4.27	2.50	0.04	-1.50	1.77
63	6	533	8.46	-0.73	-2.46	-0.04	-1.54	1.73

i	n_i	N_i	N_i/i	D_i	d_i	r_i	R_i	$N_i/i-n/k$
64	5	538	8.41	-1.73	-3.41	-0.05	-1.59	1.68
65	12	550	8.46	5.27	3.54	0.05	-1.54	1.73
66	6	556	8.42	-0.73	-2.42	-0.04	-1.58	1.69
67	4	560	8.36	-2.73	-4.36	-0.06	-1.64	1.63
68	6	566	8.32	-0.73	-2.32	-0.04	-1.68	1.59
69	1	567	8.22	-5.73	-7.22	-0.10	-1.78	1.49
70	8	575	8.21	1.27	-0.21	-0.01	-1.79	1.48
71	32	607	8.55	25.27	23.45	0.34	-1.45	1.82
72	6	613	8.51	-0.73	-2.51	-0.04	-1.49	1.78
73	7	620	8.49	0.27	-1.49	-0.02	-1.51	1.76
74	1	621	8.39	-5.73	-7.39	-0.10	-1.61	1.66
75	2	623	8.31	-4.73	-6.31	-0.08	-1.69	1.58
76	6	629	8.28	-0.73	-2.28	-0.03	-1.72	1.55
77	2	631	8.19	-4.73	-6.19	-0.09	-1.81	1.46
78	1	632	8.10	-5.73	-7.10	-0.09	-1.90	1.37
79	6	638	8.08	-0.73	-2.08	-0.02	-1.92	1.35
80	3	641	8.01	-3.73	-5.01	-0.07	-1.99	1.28
81	10	651	8.04	3.27	1.96	0.03	-1.96	1.31
82	5	656	8.00	-1.73	-3.00	-0.04	-2.00	1.27
83	13	669	8.06	6.27	4.94	0.06	-1.94	1.33
84	3	672	8.00	-3.73	-5.00	-0.06	-2.00	1.27
85	3	675	7.94	-3.73	-4.94	-0.06	-2.06	1.21
86	2	677	7.87	-4.47	-5.87	-0.07	-2.13	1.14
87	5	682	7.84	-1.73	-2.84	-0.03	-2.16	1.11
88	2	684	7.77	-4.47	-5.77	-0.07	-2.23	1.04
89	9	693	7.79	2.27	1.21	0.02	-2.21	1.06
90	11	704	7.82	4.27	3.18	0.03	-2.18	1.09
91	9	713	7.84	2.27	1.16	0.02	-2.16	1.11
92	5	718	7.80	-1.73	-2.80	-0.04	-2.20	1.07
93	5	723	7.77	-1.73	-2.77	-0.03	-2.23	1.04
94	4	727	7.73	-2.73	-3.73	-0.04	-2.27	1.00
95	6	733	7.72	-0.73	-1.72	-0.01	-2.28	0.99
96	4	737	7.68	-2.73	-3.68	-0.04	-2.32	0.95
97	8	745	7.68	1.27	0.32	0.00	-2.32	0.95
98	4	749	7.64	-2.73	-3.64	-0.04	-2.36	0.91

i	n_i	N_i	N_i/i	D_i	d_i	r_i	R_i	N_i/i-n/k
99	8	757	7.65	1.27	0.35	0.01	-2.35	0.92
100	11	768	7.68	4.27	3.32	0.03	-2.32	0.95
101	8	776	7.68	1.27	0.32	0.00	-2.32	0.95
102	4	780	7.65	-2.73	-3.65	-0.03	-2.35	0.92
103	8	788	7.65	1.27	0.35	0.00	-2.35	0.92
104	5	793	7.63	-1.73	-2.63	-0.02	-2.37	0.90
105	5	798	7.60	-1.73	-2.60	-0.03	-2.40	0.87
106	5	803	7.58	-1.73	-2.58	-0.02	-2.42	0.85
107	3	806	7.53	-3.73	-4.53	-0.05	-2.47	0.80
108	3	809	7.49	-3.73	-4.49	-0.04	-2.51	0.76
109	4	813	7.46	-2.73	-3.46	-0.03	-2.54	0.73
110	4	817	7.43	-2.73	-3.43	-0.03	-2.57	0.70
111	2	819	7.38	-4.73	-5.38	-0.05	-2.62	0.65
112	8	827	7.38	1.27	0.62	0.00	-2.62	0.65
113	8	835	7.39	1.27	0.61	0.01	-2.61	0.66
114	3	838	7.35	-3.73	-4.35	-0.04	-2.65	0.62
115	3	841	7.31	-3.73	-4.31	-0.04	-2.69	0.58
116	4	845	7.28	-2.73	-3.28	-0.03	-2.72	0.55
117	2	847	7.24	-4.73	-5.24	-0.04	-2.76	0.51
118	5	852	7.22	-1.73	-2.22	-0.02	-2.78	0.49
119	15	867	7.29	8.27	7.71	0.07	-2.71	0.56
120	5	872	7.27	-1.73	-2.27	-0.02	-2.73	0.54
121	5	877	7.25	-1.73	-2.25	-0.02	-2.75	0.52
122	2	879	7.20	-4.73	-5.20	-0.05	-2.80	0.47
123	21	900	7.32	14.27	13.68	0.12	-2.68	0.59
124	7	907	7.31	0.27	-0.31	-0.01	-2.69	0.58
125	6	913	7.30	-0.73	-1.30	-0.01	-2.70	0.57
126	5	918	7.29	-1.73	-2.29	-0.01	-2.71	0.56
127	5	923	7.27	-1.73	-2.27	-0.02	-2.73	0.54
128	11	934	7.30	4.27	3.70	0.03	-2.70	0.57
129	3	937	7.26	-3.73	-4.26	-0.04	-2.74	0.53
130	10	947	7.28	3.27	2.72	0.02	-2.72	0.55
131	4	951	7.26	-2.73	-3.26	-0.02	-2.74	0.53
132	3	954	7.23	-3.73	-4.23	-0.03	-2.78	0.50
133	5	959	7.21	-1.73	-2.21	-0.02	-2.80	0.48

i	n_i	N_i	N_i/i	D_i	d_i	r_i	R_i	N_i/i-n/k
134	4	963	7.19	-2.73	-3.19	-0.02	-2.82	0.46
135	3	966	7.16	-3.73	-4.16	-0.03	-2.85	0.43
136	5	971	7.14	-1.73	-2.14	-0.02	-2.87	0.41
137	9	980	7.15	2.27	1.85	0.01	-2.86	0.42
138	8	988	7.16	1.27	0.84	0.01	-2.85	0.43
139	10	998	7.18	3.27	2.82	0.02	-2.83	0.45
140	9	1007	7.19	2.27	1.81	0.01	-2.82	0.46
141	4	1011	7.17	-2.73	-3.17	-0.02	-2.84	0.44
142	7	1018	7.17	0.27	-0.17	0.00	-2.84	0.44
143	7	1025	7.17	0.27	-0.17	0.00	-2.84	0.44
144	3	1028	7.14	-3.73	-4.14	-0.03	-2.87	0.41
145	2	1030	7.10	-4.73	-5.10	-0.04	-2.91	0.37
146	5	1035	7.09	-1.73	-2.09	-0.01	-2.92	0.36
147	4	1039	7.07	-2.73	-3.07	-0.02	-2.94	0.34
148	4	1043	7.05	-2.73	-3.05	-0.02	-2.96	0.32
149	8	1051	7.05	1.27	0.95	0.00	-2.96	0.32
150	6	1057	7.05	-0.73	-1.05	0.00	-2.96	0.32
151	11	1068	7.07	4.27	3.93	0.02	-2.94	0.34
152	3	1071	7.05	-3.73	-4.05	-0.02	-2.96	0.32
153	5	1076	7.03	-1.73	-2.03	-0.02	-2.98	0.30
154	3	1079	7.01	-3.73	-4.01	-0.02	-3.00	0.28
155	9	1088	7.02	2.27	1.98	0.01	-2.99	0.29
156	1	1089	6.98	-5.73	-5.98	-0.04	-3.03	0.25
157	3	1092	6.96	-3.73	-3.96	-0.02	-3.05	0.23
158	5	1097	6.94	-1.73	-1.94	-0.02	-3.07	0.21
159	5	1102	6.93	-1.73	-1.93	-0.01	-3.08	0.20
160	5	1107	6.92	-1.73	-1.92	-0.01	-3.09	0.19
161	9	1116	6.93	2.27	2.07	0.01	-3.08	0.20
162	8	1124	6.94	1.27	1.06	0.01	-3.07	0.21
163	2	1126	6.91	-4.73	-4.91	-0.03	-3.10	0.18
164	15	1141	6.96	8.27	8.04	0.05	-3.05	0.23
165	1	1142	6.92	-5.73	-5.92	-0.04	-3.09	0.19
166	1	1143	6.89	-5.73	-5.89	-0.03	-3.12	0.16
167	2	1145	6.86	-4.73	-4.86	-0.03	-3.15	0.13
168	12	1157	6.89	5.27	5.11	0.03	-3.12	0.16

i	n_i	N_i	N_i/i	D_i	d_i	r_i	R_i	$N_i/i-n/k$
169	9	1166	6.90	2.27	2.10	0.01	-3.11	0.17
170	2	1168	6.87	-4.73	-4.87	-0.03	-3.14	0.14
171	3	1171	6.85	-3.73	-3.85	-0.02	-3.16	0.12
172	9	1180	6.86	2.27	2.14	0.01	-3.15	0.13
173	6	1186	6.86	-0.73	-0.86	0.00	-3.15	0.13
174	5	1191	6.84	-1.73	-1.84	-0.02	-3.17	0.11
175	7	1198	6.85	0.27	0.15	0.01	-3.16	0.12
176	9	1207	6.86	2.27	2.14	0.01	-3.15	0.13
177	5	1212	6.85	-1.73	-1.85	-0.01	-3.16	0.12
178	13	1225	6.88	6.27	6.12	0.03	-3.13	0.15
179	11	1236	6.91	4.27	4.09	0.03	-3.10	0.18
180	1	1237	6.87	-5.73	-5.87	-0.04	-3.14	0.14
181	2	1239	6.85	-4.73	-4.85	-0.02	-3.16	0.12
182	3	1242	6.82	-3.73	-3.82	-0.03	-3.19	0.09
183	5	1247	6.81	-1.73	-1.41	-0.01	-3.20	0.08
184	4	1251	6.80	-2.73	-2.80	-0.01	-3.21	0.07
185	10	1261	6.82	3.27	3.18	0.02	-3.19	0.09
186	2	1263	6.79	-4.73	-4.79	-0.03	-3.22	0.06
187	3	1266	6.77	-3.73	-3.77	-0.02	-3.24	0.04
188	1	1267	6.74	-5.73	-5.74	-0.03	-3.27	0.01
189	10	1277	6.76	3.27	3.24	0.02	-3.25	0.03
190	6	1283	6.75	-0.73	-0.75	-0.01	-3.26	0.02
191	5	1288	6.74	-1.73	-1.74	-0.01	-3.27	0.01
192	3	1291	6.73	-3.73	-3.73	-0.01	-3.28	0.01
193	8	1299	6.73	1.27	1.27	0.01	-3.27	0.00
194	11	1310	6.75	4.27	4.25	0.02	-3.25	0.02
195	3	1313	6.73	-3.73	-3.73	-0.02	-3.27	0.00
196	7	1320	6.73	0.27	0.27	0.00	-3.27	0.00

$n_1 + R_k = n/k = 10 + (-3.27) = 6.73$

TABLE IV: Sentence-length-strategy in Text 9 (see also Table II in Appendix)

Text 9

i	n_i	N_i	N_i/i	D_i	d_i	r_i	R_i	$N_i/i-n/k$
1	10	10	10.00	-2.21	0.00	-	-	-2.21
2	14	24	12.00	1.79	2.00	2.00	2.00	-0.21
3	5	29	9.67	-7.21	-4.67	-2.33	-0.33	-2.54
4	7	36	9.00	-5.21	-2.00	-0.67	-1.00	-3.21
5	4	40	8.00	-8.21	-4.00	-1.00	-2.00	-4.21
6	17	57	9.50	4.79	7.50	1.50	-0.50	-2.71
7	11	68	9.71	-1.21	1.29	0.21	-0.29	-2.50
8	7	75	9.38	-5.21	-2.38	-0.33	-0.62	-2.83
9	13	88	9.78	0.79	3.22	0.40	-0.22	-2.43
10	14	102	10.20	1.79	3.80	0.42	0.20	-2.01
11	12	114	10.36	-0.21	1.64	0.16	0.36	-1.85
12	12	126	10.50	-0.21	1.50	0.14	0.50	-1.71
13	11	137	10.54	-1.21	0.46	0.04	0.54	-1.67
14	15	152	10.86	2.79	4.14	0.32	0.86	-1.35
15	7	159	10.60	-5.21	-3.60	-0.26	0.60	-1.61
16	12	171	10.69	-0.21	1.31	0.09	0.69	-1.52
17	3	174	10.24	-9.21	-7.24	-0.45	0.24	-1.97
18	4	178	9.89	-8.21	-5.89	-0.35	-0.11	-2.32
19	8	186	9.79	-4.21	-1.79	-0.10	-0.21	-2.42
20	3	189	9.45	-9.21	-6.45	-0.34	-0.55	-2.76
21	13	202	9.62	0.79	3.38	0.17	-0.38	-2.59
22	9	211	9.59	-3.21	-0.59	-0.03	-0.41	-2.62
23	16	227	9.87	3.79	6.13	0.28	-0.13	-2.34
24	22	249	10.38	9.79	11.62	0.51	0.38	-1.83
25	17	226	10.64	4.79	6.36	0.26	0.64	-1.57
26	7	273	10.50	-5.21	-3.50	-0.14	0.50	-1.71
27	20	293	10.85	7.79	9.15	0.35	0.85	-1.36
28	18	311	11.11	5.79	6.89	0.26	1.11	-1.10

29	26	337	11.62	13.79	14.38	0.51	1.62	-0.59
30	11	348	11.60	-1.21	-0.60	-0.02	1.60	-0.61
31	8	356	11.48	-4.21	-3.48	-0.12	1.48	-0.73
32	5	361	11.28	-7.21	-6.28	-0.20	1.28	-0.93
33	5	366	11.09	-7.21	-6.09	-0.19	1.09	-1.12
34	14	380	11.18	1.79	2.82	0.09	1.18	-1.03
35	6	386	11.03	-6.21	-5.03	-0.15	1.03	-1.18
36	22	408	11.33	9.79	10.67	0.30	1.33	-0.88
37	26	434	11.73	13.79	14.27	0.40	1.73	-0.48
38	18	452	11.89	5.79	6.11	0.16	1.89	-0.32
39	24	476	12.21	11.79	11.79	0.32	2.21	0.00

$n_1 + R_k = n/k = 10 + 2.21 = 12.21$