QUANTITATIVE LINGUISTICS

Volume 52



L. Hřebíček, G. Altmann (eds.)

# Quantitative Text Analysis

WVT Wissenschaftlicher Verlag Trier

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

**Quantitative text analysis** /L. Hřebíček; G. Altmann (ed.).-
Trier: WVT Wissenschaftlicher Verlag Trier, 1993
(Quantitative linguistics; Vol. 52)
ISBN 3-88476-080-7
NE: Hřebíček, Luděk [Hrsg.]; GT

Umschlaggestaltung
Brigitta Disseldorf
smart graphik & design, Trier


ISBN 3-88476-080-7



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

# Preface

*The deepest qualitative knowledge is a byproduct of some quantitative knowledge.*
M. Bunge, Exploring the World 1983: 188.

A text has so many facets that its examination nowadays makes up a vast scientific area in which researchers from several disciplines are engaged. The present volume is a rather modest survey of some specific domains that have been made accessible to the use of quantitative methods. Many other domains are merely indicated as possibilities: There is a wealth of mathematical methods with the aid of which particular problems could be solved, or which could trigger the construction of partial theories, but one waits for hypotheses from textologists.

The scope of these possibilities is enormous: Texts or their parts have physical, sociological, psychological, linguistic, semiotic and information theoretical aspects; in addition, in the course of text generation emergent textological patterns arise, e.g. rhyme, chaotic sequences, frequency distributions, aggregates, style, etc. And, as is a good custom of the reality surrounding us, texts behave like systems in all their aspects.

Since nobody knows everything, textological research must be performed *viribus unitis*. There must be researchers dealing with concept formation, as can be found here in the contributions by Gordesch/Zapf and Marcus. However, the romantic prose of concept formation by qualitative textologists should be shunned. There must be researchers dealing with problems of measurement, like Köhler/Galle, Tuldava, Roos and Mikk/Elts in this volume. There must be descriptive text analysts, engaged in examination of special properties of individual texts or genres, setting up local hypotheses and testing them objectively, as Uhlířová, Těšitelová and Baumann here, and in turn those who classify the evaluated texts and contribute to our taxonomic knowledge, like Liiv/Tuldava in this volume.

Some researchers prefer to concentrate on classes of texts, e.g. myths, narratives or dialogues, like Marcus, Wildgen, Møller, Těšitelová and Baumann. Others try to find models for particular text phenomena, like Chitashvili/Baayen, conceive or compare quantitative descriptive means, like Roos and Köhler/Galle, or search for connections of texts to other modes of communication, like Fenk.

Those who try to elaborate an ontology of texts (like Hřebíček/Altmann or Hřebíček), or a theory of some text aspects based on systems theory, will al-

ways be a small oppressed minority attacked from all sides, while the other extreme, developing computation tools, like Klein (whose INTEXT was awarded with the software prize of German Universities), enabling the textologists to perform numerical empirical investigations and to scrutinize their hypotheses, will soon become the welcome overwhelming majority if a sufficient number of textologists is ready to conceive hypotheses and to use these facilities.

In order to show many different patches on the textological palette, we present a variegated and we hope representative sample of European quantitative text research encompassing several countries reaching from the Netherlands to Georgia and from Denmark to Austria. The Russian contributions to this field will appear in a separate volume. The New World participated in this volume in the person of Sheila Embleton who tried hard to change our Pidgin English into readable English without injuring our sacred personal styles, and to whom all authors are immortally indebted.

We hope that the ideas and methods presented here in modest mathematical garments but in a rigorous diction will suggest avenues for further research.

L.H.
G.A.

# CONTENTS

## *TEXT READABILITY*



## *DESCRIPTION AND CLASSIFICATION*



## *SOFTWARE*





# Prospects of Text Linguistics

*An Essay*

*Luděk Hřebíček, Prague*
*Gabriel Altmann, Bochum*

*A start in mathematization or mathematical modelling, however unrealistic, is better than either a prolix but unenlightening description or grandiose verbal sketch.*
Mario Bunge 1967: 469.



### Introduction

Many linguists intending to investigate texts always feel the need to solve questions like the following ones:

- What is text?
- In which parts is it to be segmented?
- Why are definitions and opinions differing from his/her own so incorrect?

These and many other questions indicate that the concept of text - just as that of any real entity - is quite vague and that difficulties are connected with any attempt to define it. While on lower linguistic levels everything seems to be simpler and clearer, and the pertinent units can (seemingly) be identified with-

out great difficulties, the variability of texts exceeds any imaginable limit. The task of finding entities common to all possible texts at first sight seems to be impracticable. But this is the very reason for the existence of text linguistics, one of its main tasks.

Every text can doubtlessly be considered as representing a system having structures and being generated by processes. System, process and structure are fundamental concepts of the sciences in our century; this also holds for linguistics. These universal concepts represent unremovable constituents of our thinking. All conceptual universals of this sort should become clearer when scientific work achieves a more mature level.

The following chapters represent rather a free discussion, provoking, disarranged and incomplete, concerning the fundamentals of quantitative text linguistics and its technical tools which are at linguists' disposal.

There are many problems worthy of discussion in connection with this topic. Several of them have the power necessary to provoke a deeper reasoning concentrated on the increase in the scientific status of linguistics. They are mentioned in this paper. The decisive arguments in favour of quantitative approaches in linguistics can be found in the works presenting the actual formulations of scientific theories, part of which is contained in the present volume.

## Text as an object of scientific observation

It seems to be more difficult to find a definition of text than of sentence, once so popular in linguistics. Since all definitions are conventions, it can appear to be reasonable to make an agreement that everybody knows what is to be meant by the term in question. Let us try to formulate such a convention saying that

> *text is a continuous formation in a natural language that can be segmented into sentences and words.*

Definitions of this kind are only partly operational. Can one say that texts consist of phonemes, morphemes, syllables,...,sentences? Yes and no a text consists of as many types of entities as we are able to conceive, i.e. it depends on our hypotheses which entities it consists of. We want to indicate that making definitions is an activity far from leading to truth. Each phenomenon can be viewed from different sides and aspects; thus the number of definitions can be unlimited.

Many other properties could be added in order to obtain some strictly formulated idea of text. For example, it can be said that text always originates in some natural conditions.

No observed text can be obtained by a random or non-random choice from a set of some prefabricated constituents since text is observed as unity. On the other hand, text can be formed by an individual sentence or even by an individual word or phoneme if such a text represents a unity observed under circumstances that can be supposed to be "natural". Text can have a sound form or be set down in letters or in another way of fixing.



Not only in linguistics but in almost all human sciences the tendency to define more and more general concepts can be observed. The purpose of this activity is to construct conceptual hierarchies in which the more general concepts "explicate" the more specific ones. This possibly is the case of the terms 'utterance' and 'discourse' in linguistics (the latter concept being used by some new philosophical schools). Definitions are always conventions: they represent an encroachment of an observer into the observed reality. To define something does not mean 'to be objective'. The construction of a conceptual net, as it often occurs in human sciences, cannot be taken as a decisive step towards the formulation of a theory. It is merely a necessary but not a sufficient condition, the first step towards the determination of the universe of discourse. The nucleus of a theory is to be found in other constituents constructed from this net, namely in testable hypotheses and laws.

When the notion of text is subordinated to a higher concept, be it 'utterance' or 'discourse' or something similar, and when these concepts are used as substitutes for 'text', we can hardly suppose that the status of the subordinated phenomenon becomes more clear or more evident. The superordinated concepts cannot make an object more clear or lucid if theoretical knowledge is not at hand. In most cases it is merely a classificatory process without any epistemological consequences. Therefore we decided not to use and discuss here the more abstract terms.

The name 'text linguistics' was coined at approximately the end of the sixties. This branch is naturally rooted in the earlier history of linguistics, but in the indicated decade a new substantial idea became popular, namely the idea of *text coherence* based on *text references*. Before this important progress text was naturally an object of linguistic reasoning, and several philological branches had text as their object. Let us mention stylistics and poetics which stand closer to present-day text linguistics than other philological branches. They present sets of instructions for the use of language phenomena for certain purposes, for example, to become a novel writer; yet they can scarcely be assumed to be a systematic insight into text structures. Representatives of these branches understand text as a number of sentences ordered with some extralinguistically defined purposes in a sequence. They investigate special characteristics of texts written by different authors, especially of belles-lettres and poetry, while nothing substantial is known about the general properties of texts. This approach is very useful and instructive, but it is not scientific as no testable hypothesis is set up saying why a text is organized in the observed way. Linguistic as well as extralinguistic contexts are taken into account in non-systematic ways with the purpose of bringing some knowledge about individual texts. 'This poet writes

so and so' is the general form of their statements. And neither their methodological position nor their statements are quite lucid; abstract and general concepts dominate in their terminology, while observational concepts are lacking. The aesthetic relevance of belletristic and poetic texts justifies this approach. The linguistic aspects of the problem unfortunately remain untouched.

The attempts to explain some aspects of texts from their external conditions or circumstances are acceptable only when these circumstances can be captured in a rational way. However, this is a very rare case. Facts like author's age, social status, and time and place of the origination of texts can be largely commented upon and semantically interpreted, but they can scarcely be included in some testable scientific theory. Under very favourable conditions they can merely explain the variation of some parameters in the theory. However a phenomenon cannot be rationally explained by irrationalities or arguments themeselves waiting for rational explanation.

It is evident that text is not only a sequence of sentences. And 'sentence' itself is a formation with a high variability. Thousands of years of history of linguistics is full of investigations concerning sentences, their inner relations and forms of their expressions. The assertion that almost all of linguistics is in fact a science concerning grammar is - unfortunately - not very far from the truth. This is the reason why we know so little about texts. However, this is not the only reason, and we want to touch here on some other ones, too.

It is a great paradox both of the human and biological sciences that human beings use their brains and the natural instruments of their bodies without having knowledge about these gifts. The same holds for natural language. Each communicator is able to construct necessary texts. Writers can write texts on the theme 'how I wrote my texts'; their explanations are interesting, sometimes amusing, but never rational and complete.

Needless to say, the necessity to understand text is identical with the necessity to understand language. This is not an artificial need. In many human activities language is used by authorities as something evident and with far-reaching consequences for lives of human individuals. Jurisprudence and jurisdiction can serve as a typical example of the sovereignty in the interpretation of language expressions and texts.

## Philosophy of science and linguistics

There are sciences defining themselves as possessing a special methodology singular in the world of science. On the one hand, this is absolutely correct for each science. On the other hand, this is an alibi for methodological indolence.

For several centuries, natural sciences tending to mathematization have understood their common methodological fundamentals with a greater ease. However, a sharp dividing line between natural sciences and the humanities is



not acceptable. Human knowledge cannot be divided into separate sections, as each human individual, who is the only bearer of knowledge, cannot be torn into incoherent parts.

Traditional as well as modern linguistics is often supposed to be a special branch with incomparable aims and procedures. There is no reason to refuse spiritual or philosophical approaches to anything including language. However, if a scientific explanation becomes the aim of the epistemological activity, then not every interpretative treatment of the objects studied can be taken as a scientific explanation. For example, a set of instructions for generating correct sentences in linguistics can hardly be understood as an explanation. Linguistics cannot be separated from the whole of human knowledge and, consequently, from the general principles of methodology comprised under the term 'science'.

Whenever human beings seek solutions to problems or puzzles, they begin to ascribe names to phenomena they consider as relevant. This first step is called concept formation, but if subsequently no testable hypetheses are set up, it is merely word magic. Names usually are supposed to belong to some semantic categories. Semantic entities imply classifications and a classification sometimes contains an explication. This explication, however, is not identical with the scientific explanation.

Everybody knows what Noun, Verb, Present Tense and Past Tense are. All these and similar categories are useful for a rational contact with languages on the metalanguage level of thinking; there is no reason to replace them by something else.

However, we try to construct something else that is capable of yielding explanation. When describing languages our first task is to differentiate between science and non-science. Yet 'to be non-science' does not mean 'to be bad or false or useless', etc. It merely means 'to be different from science'.

We do not need to seek some original criteria for making this difference. This question has been answered by all schools of metascience - even if not uniformly. Nevertheless, all of them agree that an empirical science must be based on a theory being a system of testable hypotheses some of which must have the status of laws (cf. Bunge 1967). No explanations are possible without such a theory.

Now it is evident that our reflections have to do merely with one part of linguistics, namely theoretical linguistics. Unfortunately, only some parts of existing theoretical linguistics fulfill the above criteria; it is not difficult to identify these parts. On the other hand, a great part of linguistics is captured by tradition. And these orthodox parts dispensed - and are steadily dispensing - with the formulation of testable theories.

As far as (the orthodox) modern linguistics is concerned, it stated its goal in setting up formal rules and instructions for creation of language formations. The metalanguage describing these rules is often very formalized. But the level of formalization is not decisive for the quality of a theory and rules and instruc-

tions are not explanatory statements. Rules are empirical generalizations that can never achieve the status of laws. Rules demarcate the limits of investigation which can never be surpassed. The qualities of instructions determine the quality of descriptions; the way in which instructions are obtained can be arbitrary and the system of instructions, their mutual relations and everything they can tell about language is then deduced from the postulated arbitrary instructions. Consequently, the explanatory power of systems of this kind is not high. Instructions serving for creation of language constructs may be written in words or in structural formulas, a grammar can be written or a software program may be put together for the purpose of constructing correct sentences, but all this does not enhance our knowledge of language. Surely, in this way an abstract space, an artificial world, is created; but this world is completely independent of the real language systems to the same degree as the machinery of the tape-recorder is independent of the systems of reception proper to living recipients of language structures.

Our aim is the construction of empirical linguistic theories.

According to our presumptions language is a phenomenon observable in its consequences or outputs. To understand linguistics as an abstract science like for example logic is an attempt directed to something quite different from natural languages. Abstract sciences and their results can be applied to empirical sciences; however, they cannot replace them.

The criteria mentioned above are of high relevance for text linguistics. It can hardly be imagined to be a branch built up as an abstract science able to generate all correct texts of a natural language. This aim appears to be a fantasy. Text linguistics should encompass the entire empirical knowledge concerning language and perhaps something more: it should surpass the boundaries of pure linguistics.

## Text linguistics and quantitative linguistics

The generally accepted requirement of testability is a call for constructing theories on a knife-edge. The conditions and the range of validity of a theory must be delimited in a clear way, if this criterion is to be fulfilled. It is evident that the quantitative approach represents the broadest way towards this aim, and after a short hesitation it becomes evident that this is the only way to construct testable theories. Of course, the possibility of finding other ways to testable theories in the future cannot be excluded, but for the present time there is no suitable substitute. The only alternative is represented by free philosophical reflections about language, which, of course, cannot claim any theoretical status.

The transformation of observed data into quantities, i.e. variables and constants, has a high epistemological value esteemed in all more mature sciences; measurement, if made correctly, represents a standardized approach to observation, it enables us to perform experiments in branches where they represent something unusual. Measurement is a way of control, establishing a direct connection to certain testing methods. The quantitative approach enables us to use different branches of mathematics and logic. It opens the door to the possibility of formulation of testable (and of course, also non-testable, but this is not interesting) hypotheses. In any discipline one can formulate hypotheses without limitations, but those requiring measurement and quantities for their testing have a special quality increasing the scientific status of the respective discipline. The foundations of mathematics, mathematical statistics, probability theory, theory of games, graph theory, systems theory and many other branches represent a large stock of methods at the disposal of the investigator who decides to construct testable theories in his/her discipline.



Let us stress one important aspect of these facts: When a linguistic theory is built up, i.e. when after the formulation of a conceptual net, stating of axioms and conventions (definitions, operations, etc.), hypotheses are pronounced (or derived) and successfully tested, in no way are the traditional approaches of classical linguistics eliminated from the sphere of intellectual enterprise. Everybody can pronounce free reflections about language and text after his/her own heart. The necessity of commenting on all inputs and outputs of theories is not annihilated by testing. However, if a testable theory is missing, the scientific status of the ideas presented is lowered. This holds for every empirical science and linguistics does not represent any exception. Thus the quantitative approach supplies a linguistic theory with an explanatory apparatus.

In order to anticipate the possible objections with which quantitative linguists are well acquainted (how monotonous they are!), let us add some self-evident notes: The quantitative approach is not synonymous with an automatically correct approach. A work presenting some quantities and their relations does not bear a hallmark of correctness. A quantitative approach without an attempt to construct a theory is stigmatized with a lowered scientific status regardless of formulas, graphs or block-schemes. In the earlier decades of the second half of the 20th century, linguistic journals and books were full of tables presenting frequencies, percents and graphs indicating the values of variables. Only a "trifle" was lacking: a theory - or at least a modest hypothesis - saying what all this means. It cannot be said that all such work is senseless, but its sense has not been indicated. Many young linguists often come to their older colleagues with mountains of numerical computer outputs asking what they could do with these numbers. The answer is stereotypical: Throw them into the waste-paper basket. The usual procedure is to start from a linguistic hypothesis, to translate it into mathematical language, to expose it to a test based on a probability model, to translate the result into natural language and to interprete it linguistically (cf. Altmann 1972).

Another objection to the application of the quantitative approach in human sciences suggests that not quantities but qualities are substantial. However,

when one looks around elsewhere in the world and in the surrounding universe, neither qualities nor quantities can be found; qualities and quantities do not exist at all. They are always human constructions, or in a sense relations between the observed reality and the observing individual. Avoiding the quantitative aspects of phenomena means a lot of abstraction and a loss of information. A "qualitative approach" usually means a verbal characterization of certain phenomena and as such it represents a simplification directed against the construction of testable theories (cf. e.g. Bunge 1961).

General methodology stresses the fact that testing hypotheses does not mean the way for reaching unshakable truth. Every theory can and certainly will be refuted, or it will be included in a more general theory or amended by introducing some new presumptions. The quantitative approach is motivated not merely by the desire to be more precise, since it can also be used for approximations or simplifications which, however, in contrast with the semantic approximations, are usually under the control of the investigator.

The rise of text linguistics and its constitution as a scientific branch certainly does not mean the end of the classical branches dealing with texts, e.g. stylistics or literary criticism. The same is valid for the quantitative approaches in linguistics. Essayistic treatment of language and its phenomena is not touched by stressing the methodological criteria for empirical sciences. However, our common aim is to establish text linguistics as a theoretical discipline.

The other aspect of our reasoning is represented by the fact that without taking text into the focus of interest theoretical linguistics will have fundamental problems. Text was always passed over by linguistic theoreticians because of its variability and polymorphous character. In a number of works quantitative text linguistics has manifested its ability to get along with these problems. If the examination of text ought to become the aim of theoretical activity in linguistic science, then testability of its theories and, consequently, a quantitative approach are unavoidable.

## Text as a structure

Several trends in modern linguistics are expounded in such a way that they present a main idea and all the other ideas are semantically related to, or derived from, this idea. The difference between this approach and the building of hypotheses in actual scientific theories is testability. Semantic relations are not testable, they represent rational axioms or even principles. Basing explanations on relations of this kind occurs in historical or philosophical literature (e.g. the struggle of nations or social classes as the general principle of human history; or the principle of being thrown into existence in philosophy). Karl Popper (1963: 72) characterized this approach in the following way:



> 'It is very interesting that the imitators were always inclined to believe that the 'master' did his work with the help of a secret method or a trick. It is reported that in J.S.Bach's days some musicians believed that he possessed a secret formula for the construction of fugue themes.
> It is also interesting to note that all the philosophies which have become fashionable (so far as I am aware) have offered their disciples a kind of method for producing philosophical results. This is true of Hegelian essentialism which teaches its adherents to produce essays on the essence or nature or idea of everything - the soul, the universe...'

And we are inclined to add: ... language. Such a trick is now often regarded in the formulation of rules, and formerly this role was played by the principle of structure.

Language is a tool of communication. This means that it serves for transmission of information. Under information we understand the properties defined by C.E. Shannon (cf. Shannon & Weaver 1949), N. Wiener and other founders of the mathematical theory of information.

The concept of structure can be intuitively imagined as a property of an articulated, organized phenomenon. Then it can be asserted that the vehicle bearing information must always be a structured phenomenon. Non-structured phenomena are always maximally homogeneous.

The general concept of information theory is that of entropy. It represents a quantity expressing the degree of arrangement existing among constituents of a phenomenon measured by entropy. It indicates a degree of organization. One can suppose that the extent of structuring corresponds to the ability to bear information. This measure is inversely correlated with the amount of entropy.

Consequently, every structure is predetermined to function as an information vehicle, and information is always bound to a structure. Meaning seems to be a sort of semantically interpreted information.

It can be supposed that the predominance of structuralism in the history of sciences consisted in its non-explicit giving of importance to information through the concept of structure. The necessity to connect structuralism (and the other schools following structuralism) with the general theory of communication can be regarded as a logical step in the further development of linguistic thinking. Only this connection is able to make the real functions of structures explicit.

Language as a means of communication can be used in an infinite number of situations and for an infinite number of purposes. This possibly is the reason for the complexity of its structure; language bears many different items of communication and it is adapted to this functioning. On the other hand, its structure must be intuitively lucid for its users, even for little children. This intuitive understanding is based on semantic evaluations of the structural parts of the communication vehicle made instantly and permanently by the language users. Scientific explanation must differ from the users' understanding mentioned

above. Explanation in sciences means 'coordination to testable theories', or according to Salmon (1984), it means the embedding of a phenomenon in a discernible nomological structure of the world.

Besides the rules stated by each mother teaching her child to speak, or by an Academy of Sciences, or by teachers at schools, there are elements in languages which deserve explanation. Unpredictable fluctuations occur at all times. And chance is a source of dynamics in languages, too; this kind of flexibility makes it possible to use language in all possible situations with which human beings are faced.

The complicated structures of texts are supplied with something that is not defined as a part of the construction existing in language. A language cannot be described only by a set of instructions and rules, there is something that can be called 'structural possibility' and each analytical approach to text should take this property into account. Chance as well as choice inseparably belong not only to *parole* but also to *langue*. Any structural description omitting chance in fact neglects Man, the user and interpreter of the language formations. This means that linguistics is not in a special methodological position as compared with the other empirical sciences. The concept of chance and randomness enables us to apply the method elaborated in mathematical statistics and probability theory, and in the related sciences as well. This circumstance facilitates the formulation of testable theories in linguistics.

The situation described makes one think that the entire text is not contained in its visible or audible form. What is written on paper or on tape is not a whole text; it cannot be preserved as a whole and cannot be called forth in some original shape of all its structures. This seems to be evident and this formulation is not surprising in modern linguistics (cf. Miller 1977).

However, under these circumstances a new understanding of text can be created: we can take text as an interaction between a language formation and a communicator coincident with this formation, the communicator being the producer or receiver of the formation. This relation is illustrated in Figure 1.

A certain language formation actually represents two texts which can turn out to be identical or different. There are different possible results of the semantic interpretations by two communicators. The number of communicators is not limited. Each interpretation can be treated as an estimation of semantic structures, i.e. of those invisible correlates of the visible (audible) language vehicles that emerge from the practical knowledge of individual communicants. This procedure is full of fluctuation possibilities. Chance is in play. And one of the aims of text linguistics should become the estimation of conditions under which the probability of difference between the two (or more) interpretations can be minimized.

The scheme presented in Figure 1 has a complicated content. It puts questions as yet unsolved in linguistics and in other sciences describing human behavior; there is no doubt that interpretation is a kind of (language) behavior.



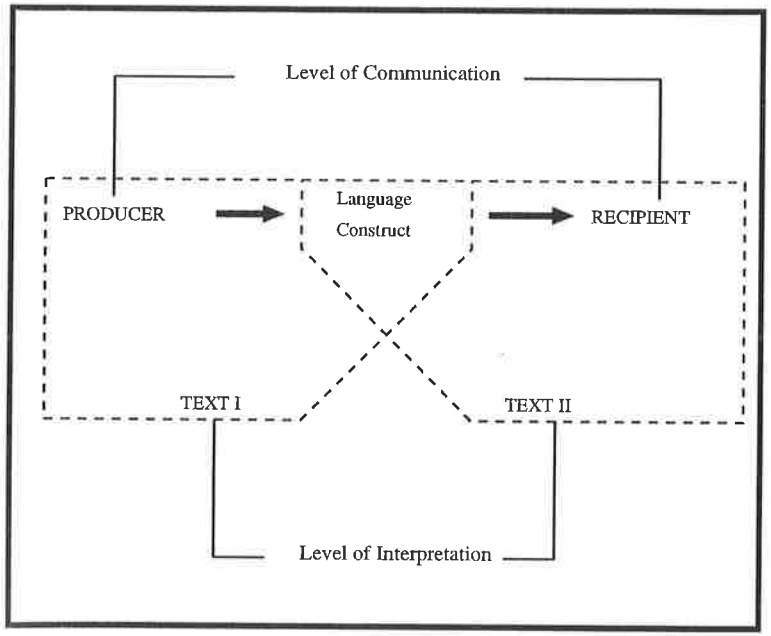



Figure 1. One and the same language construct in different interpretations forming two texts

Under the assumption that each communicator is in possession of a *communication memory*, this phenomenon can be taken as a generator of chance having consequences for the choice of language expressions. This memory is a representation of the obscure semantic structure which is formed, changing and emerging in communication processes preceding a supposed communicative act. Each communicator together with his/her communication memory is an integral part of text; it cannot be assumed to exist without a communicator.

Communicators interpret not only entire texts but also their parts and constituents having the quality of sign. Text linguistics will doubtlessly seek new approaches to some interception of the semantic interpretation mentioned above and the treatment of its structures. Here we can give but an outline of an approximative approach to the problem.

Let us suppose that each communicator producing or receiving a text decides whether a given unit on the level of words is identical with, or different from all other units he/she is able to take into account during this procedure. This is a simplification of the semantic interpretation.

With the help of this procedure a certain lexical unit occurring in different places in a text can be diversified and interpreted by communicators as different

lexical units; or different lexical units occurring in different parts of a text can be interpreted as identical units; or, finally, different lexical units are interpreted as different lexical units and nothing is changed by interpretation. Hence it follows that there exist two kinds of text analyses, one operating with interpreted, and the other with non-interpreted, lexical units. Consequently, two different vocabularies of a text are assumed, having different numbers of units and different frequencies.

If the interpreting communicator is treated as a subsystem of the text system, then one and the same language construct stimulates a number of possible interpretations. An investigator should work with a group of informants in order to obtain some standardized results of interpretation. On the other hand, each individual communicator can interpret an arbitrary text according to his/her will, determination or personal knowledge. No general criteria for preferences between and among interpretations can be given in advance.

Many interpretations are predictable with a certain probability, but the existence of unpredictable possibilities of interpretation cannot be denied. Also the predictable ones must be considered as probability functions and not as some strictly formulated rules. The generator of chance produces such fluctuations of the language structures. The intention of linguistics to exclude elements of chance from its suppositions, qualifying them as non-standard elements is, in fact, directed against the aims of linguistics.

One substantial characteristic should be stressed in connection with the concept of structure, namely its unfolding in time. Text is a composite sign, a complicated structure, and all this complex is able to go forth through an infinitely short time interval. When at a moment a language unit is exposed to interpretation, some of the related entities are inside, and the other ones are beyond the horizon of the communicator's memory.

This arrangement resembles a network moved through a little hole, thread after thread, none being broken, and after going through, the whole structure blooms in the original shape and then escapes out of sight. Language structure is usually treated without regard to its relation to time. Language structure at an infinitely little period of time is something completely senseless. As the time shortens, this structure disappears and then, thanks to the memory, it becomes visible again. This is one of the fascinating mysteries of language.

What is the sense of the existence of this abundance of structures and subsystems? Each language construct can be considered to be a challenge to interpretation. The construct itself is a result of some stimulation, a response to this stimulation. For recipients text represents a challenge to interpretation. One can say that each structure has this property when coming into contact with the preliminarily prepared and adequately organized communication memory. Structure is a challenge to interpretation - this poetical expression characterizes one of the most important properties of structure in communication. Also the sense of complexity of a structure can be characterized in this way.



A communicator's position in relation to an arbitrary structure can be briefly characterized as an exposition of the structure to his/her choice. The selected (i.e. reflected) entities are related to the structure of communicator's memory, i.e. to meanings.

Text is always an interpreted formation; it undergoes a revival in a direct interaction with a communicator who selects sign vehicles and subjects them to his/her interpretation. Interaction with a communicator is unavoidable for the existence of a text. Something written on a paper or on a tape is not a text. It begins to exist in contact with human beings.

## Text as a repetition

A repertory of entities cannot be treated as a structure; structure always means an articulated phenomenon. The units or constituents of phenomena are characterized by quantities distributed in space and time in different places or positions of a phenomenon. In short, they are repeated; they have *frequency*. Repetition is an important property of text structures which cannot be neglected in linguistic analyses. Here this property is briefly commented on, for its detailed quantitative treatment cf. Altmann (1988).

Linguistic units are repeated on all linguistic levels since they are elements or constituents of the language constructs. This salient property is the basis of their ability to function as instruments of communication.

To be individual, to have a unique occurrence, is, in fact, an irrelevant property from the standpoint of the language inventory. However, due to the existence of levels, the final products of speaking and writing are often characterized by uniqueness. Communicators often manifest their desire to present their texts as something individual and unique. This paradox is due to the flexibility of languages - and we do not hesitate to say - the beauty of its systems. As a matter of fact, it is one of the basic Bühlerian functions of language.

Any repetition inside a structure requires not only the qualities of mutual identity and difference, but especially those of similarity. Unique elements are always members of some classes and each class taken as an individual phenomenon represents an element of a higher level. And this higher element is also characterized by frequency. For example, in texts, there are words having frequency equal to one; they belong to different word categories which also possess frequencies. Words consist of morphemes, syllables or phonemes, and these units are also subordinated to the principle of repetition.

One could expect that repetition will become an important position in the philosophy of knowledge or in philosophy in general. This property of phenomena is observed and treated with appropriate consequences only by empirical sciences.

We have stressed that language can hardly be treated theoretically as a

jigsaw with preconceived rules for the procedure of stacking. Grammatical rules and lexical meanings are mere untheoretical abstractions; they summarize certain semantic intuition obtained from texts which, however, is not the actual intuition of the communicators. They are results of practical knowledge and are applicable only with an assessed enumeration of exceptions. All this is correct and necessary in practical life. Semantic treatment of an arbitrary repertory, so frequent in linguistics, is a simplification of the language system; it represents an unsystematic approach. Such simplification is useful for communicators but it has nothing in common with theories and scientific explanations.

Segmentation of arbitrary language constructs represents the reverse side of repetition. Whatever is identified as a repeated entity by a communicator is automatically taken by him as a segment. Both these aspects, repetition and segmentation, are related to the degree that it is impossible to indicate which one represents a basis in relation to the other one. The ability to produce segments is thus connected with semantic interpretation; repetition and frequency are properties closely bound to semantics.

Let us assume a segmented entity containing no repeated constituents. Such an entity is maximally non-homogeneous. (It is evident that the boundaries for dividing the constituents from each other are not part of the repertory, otherwise they should be understood as a repeating phenomenon.) On the other hand, a segmented entity consisting of units of the same kind can be supposed. This entity represents a maximal homogeneity. Within these extreme limits repetition occurs. Repetition is a concept logically connecting repertory with segmentation.

The reason for considering unique occurrences as repetitions is evident: they are parts of the *distribution* of the respective units; thus distribution is another concept implied by repetition. Consequently, to occur once is a sort of repetition.

The idea of text coherence or *cohesion* (cf. e.g. Halliday & Hasan 1976) is inseparably connected with the idea of text as a linguistic unit, similar to that of phoneme, morpheme, syllable, etc. Regardless of the fact that text is something like a universe composed of different entities, linguistics must be consistent in searching for constructs whose components are texts.

Text linguistics seeks the principles of arrangements of units on different levels and those of coordination among levels. Linguistics alone is still not able to capture these principles in their totality. The image of text structures presented by linguists is a decomposition of language constructs into levels. The instructions and rules for generating correct sentences are based on this decomposition. The measure of orderliness called 'entropy' is bound to repertories, too, but it is derived from the probabilities of the units of the respective repertory. The future task of text linguistics is to join the language levels and to offer explanations of text phenomena based on text as a language construct and linguistic unit.



From the set of already well-known principles of text structures let us mention here the famous Zipf's law. This law originated in linguistics, and later on it was shown to be relevant for different entities in other sciences, too. This law establishes a functional relation between the frequency of a unit and its rank number in the ordered set of units arranged according to their frequencies, the unit with the highest frequency having rank one.

This law was criticized for different reasons (cf. e.g. Herdan 1966: 87-91). It is true that the status of the independent and dependent variable in Zipf's original formulation is not clear. It seems that in a certain sense the independent variable is derived from the dependent one. Thus the law seems to be a consequence of a supplementary arrangement. Besides, the parameters of the law appear to depend on sample size.

From different approaches showing that ranking is a rational procedure displaying latent structures (cf. e.g. Arapov 1988, Altmann 1991), let us merely mention the argument of Orlov and Boroda who ascertained that the law holds only for the entire text, but not for samples taken from it. This is valid both for texts in natural languages and for musical "texts" as well.

If this is true then it can be deduced that there is a certain text length to which the text producer relates the distribution (repetition, frequency) of lexical units he/she uses and thus controls the flow of information (cf. Orlov, Boroda & Nadarejšvili 1982). It was proposed to call this length 'Zipf-Orlov's length'.

Thus it is evident that there are relations in texts that can be damaged by a random choice of sub-texts, i.e. by sampling. Text is a complex hierarchy of systems of subsystems, of structures of structures, a complicated arrangement of inhomogeneities; the occurrences of text units are controlled by stochastic laws.

On the other hand, the following property must be stressed: Any continuous fragment of a text inevitably possesses some characteristics of the whole text - of course, not all of its characteristics. Well, then which ones? In this connection we can mention the analogy with water: A set of molecules of water is called 'ocean', another one 'lake' and a further one 'puddle'. To differentiate between different sets of the same entities is the problem with which text linguistics is faced. Which 'molecules' are contained in texts?

Laws like Zipf-Mandelbrot's law - which removes some criticized points of Zipf's law - yield the first clues to the solution of the problem mentioned above, as well as some other ones. However, the prospects for this research can hardly be predicted. The first steps for linguistics and text linguistics to become a real empirical science are promising. Linguists must learn to operate with testable hypotheses and to realize the fact that units and levels are our conceptual constructs. If a testable hypothesis is at hand and the range of its validity is known, then the unique problem of the investigator is to use his/her own scientific imagination (e.g. concept formation) in order to obtain new information.

## Text and linguistic levels

The representatives of different human sciences, when analyzing text, regularly pose the question: What is the specific property of the given text? Such a question, however, is meaningful only if the properties of all existing and potential texts are known, or if the laws controlling text production were deciphered. Text linguistics still seems to be at the beginning of its discoveries; for the time being not very much is known about the generalities of texts.

However, an important discovery is already at hand; unfortunately, its knowledge remained latent for a long time. In the 19th century as well as in the first half of our century, several linguists - especially phoneticians - noticed that *the longer the words, the shorter their syllables.* Systematic attention to this phenomenon was paid by P. Menzerath (1928, 1954).

It is generally accepted that in language there exist several levels; no doubts were expressed against the idea according to which the units of a higher level are composites of the units of lower levels. This step-like functioning mechanism can be applied in natural languages up to the level of sentences. On this very level many problems and obstacles emerged for linguists, the greatest one being the problem of obtaining an operational definition of sentence as a language unit.

And what about higher levels? Can we hypothetize them as constructs consisting of units of lower levels?

The idea elaborated largely by P. Menzerath was formulated in a mathematical form by G. Altmann (1980; cf. also the extensive clarification presented in Altmann & Schwibbe 1989). Menzerath´s principle was generalized to an arbitrary language *construct* and its *constitutents;* this principle turned out also to be valid for non-linguistic phenomena. It is not surprising that it is parallel to the "allometric law" used above all in biology. They differ merely in the sign of the exponent: with the allometric law, concerning systems (constructs) with non-homogeneous subsystems (constituents), it is positive; with Menzerath´s law, holding for systems with homogeneous subsystems, it is negative.

This law, called Menzerath´s law, was corroborated on different levels of different languages. Therefore we can consider it as a well attested principle of construction of linguistic units. Substantive contribution to the interpretation and application of this law in linguistics was presented by R. Köhler (1984, 1986).

This law can be used as a criterion for the construction of linguistic levels, and for the time being there is no reason to deny it. The question arises whether there are also some levels higher than that of the sentence, and we may ask whether text is a construct in relation to some lower linguistic levels.

Menzerath´s law prescribes that larger constructs have smaller constituents. Are sentences *immediate* constituents of texts in the sense of this law? If so then sentences of larger texts should be shorter than those of shorter texts. Admittedly, as far as we know, this phenomenon has not been tested, but intuitively it is hardly acceptable: it can hardly be supposed that in larger novels sentences are shorter than in shorter stories of the same author. Thus sentences do not seem to be immediate constituents of texts in the sense of Menzerath´s law.



Now, the principle of linguistic levels, or in other words, the relation of constructs and constituents, should be generally applicable; it can be understood as one of the basic language principles. Then an important aim of text linguistics is to seek immediate constituents for the construct called 'text'. Of course, it is possible to set up different hypotheses of this kind and test them with the help of Menzerath´s law.

Let us assume that sentences are constituents of formations called 'text aggregations'. Every aggregation is a set of sentences containing a certain lexical unit. The number of lexical units (lexeme types) in a text is equal to the number of aggregations. The number of sentences in an aggregation equals the frequency of the lexical unit on which the respective aggregation is based - if it is counted only once in a sentence. The number of sentences in an aggregation represents the size of the construct, while the size of its constituents is expressed as the mean sentence length in the given aggregation. This hypothesis has been corroborated by a number of texts from a limited number of languages. This application of Menzerath´s law showed that sentences of a text are constituents of aggregations and aggregations are constructs consisting of sentences in the sense of the same law. The details of this hypothesis are presented in L. Hřebíček (1990a,b, 1992). The hypothesis that text is a construct composed of text aggregations as its constituents was published in L. Hřebíček (1991).

These two hypotheses arising from the same theory represent a step towards the knowledge of the semantic framework of texts. All other approaches to this problem are - as far as we know - based only on the semantic intuition of individuals. The way indicated by Menzerath´s law can be applied to setting up other hypotheses concerning units and constructs forming text, i.e. the law can be used as a hypothesis generator. In this way deeper and testable information concerning text structures can be obtained.

The basis for aggregations is the ability of words to have contexts. Here the context is defined in advance as the sentence in which the given word-unit occurs. This ability was systematically investigated by R. Köhler (1986) and his results were incorporated in his theory under the name *polytexty.* Polytexty of a lexical unit means the number of contexts in which it can occur. This property is correlated with *polylexy*, the number of different meanings of a lexical unit. Both these properties are correlated with word frequency and word length. These characteristics are mutually related by means of Menzerath´s law. This excellent theory is a clue that can help us to understand the formal properties of semantic interpretation of lexical units by communicators. The theory also puts together Menzerath´s law with the Zipf-Mandelbrot law.

Sciences make progress by getting over contradictions and by formulation of more general theories. The thesis of classical linguistics concerning the relative independence of levels, regardless of the weakening through the adjective 'relative' - is a *contradictio in adiecto.* It implies the absence of a system, an inconsistency of the description starting from a whole and passing to subsystems which are not able to form a complete system. In such a description, i.e. in classical approaches to understanding language constructs, phenomena do not represent a unity. Phonology, morphology, syntax, etc. are at hand; however, these levels are not able to make up a whole. Of course, they are (relatively) independent, in the sense that all subsystems of a system strive for independence, but the integrating tendency of the system is not touched at all in orthodox linguistic descriptions. We are not far from the truth saying that according to its content orthodox theoretical linguistics is still very restricted. It shows merely little islands of theoretical knowledge but their surrounding represents a mere protoscientific knowledge, beautiful and inspiring, but still non-theoretical.

Let it be stressed that this portrait of linguistics is highly optimistic: there are large fields waiting to be cultivated by theoretical linguists.

## Self-similarity and self-organization

There are some indications enabling us to assume the existence of a certain parallelism between theoretical linguistics and new trends in geometry. We have in mind the fractal geometry built up by B. Mandelbrot (1977), who introduced the concept of self-similarity into mathematics.

It can be preliminarily assumed that the geometrical imagination is unserviceable in linguistic applications. Nevertheless, it can be used, as is indicated by the following ideas concerning constructs and constituents:

Let us imagine an arbitrary language construct having a certain size measured in the number of its constituents. The size of the construct may be equal to one, two, three, etc. What is the greatest size of the construct? Evidently that one containing the entire text - for example, the length of a word measured in morphemes covering the whole text. A better example is that using text aggregations as representatives of a certain construct. A text can be decomposed into its aggregations: At first, aggregations having length one (in number of sentences) are marked off, then those of length two, three, etc. What is the greatest length of an aggregation in a text having $k$ sentences? Theoretically, it is that of the aggregation having length $k$. This result is surprising since one of the results of the decomposition of a construct is the construct itself.

At first sight, this result seems worthy of refutation as being illogical. Mandelbrot's theory sheds new light on this problem. The aggregation of length $k$ is not identical with the decomposed text as it represents only a set of all its sentences missing any arrangement. However, between the maximal aggregation



of the size $k$ and the respective text, there is an analogy which can be described by the concept of self-similarity. In the original theory this property is explained as invariability in relation to the change of scale.

This viewpoint can probably be used in text linguistics too. Each linguistic level represents a point on a scale which is applied to text. From the viewpoint of text, all levels and their units have the property of self-similarity. A possible task of text linguistics is to ascertain a kind of invariability proper to the subsystems of language.

Language as a phenomenon with a stochastic character can only for special purposes (i.e. in suppositions serving for the construction of different hypotheses) be presented as something extrapolated to infinity. Text is a finite phenomenon even if it is supposed to be an entity in growth. Even being in growth, it has a finished or complete sentence (and also complete syntagm, word, morpheme, syllable or phoneme) in its ultimate position. Text is a phenomenon wrapped up into its own structure. This property corresponds with its self-similarity; it emerges from its invisible state only when the unit 'text' occurs in the theoretical reflections concerning language.

There is another kind of similarity to be mentioned in this connection, namely the similarity between the expressions concerning fractal structures and the structure expressed by Menzerath's law. Compare the two formulas occurring in these theories:

$$L(\epsilon) = F\epsilon^{1-D} \qquad y = Ax^{b}$$

L = coastline length
ε = length of a stick used for the measurement of coastline
F, D = characteristic constants

y = length of the constituent
x = length of construct
A, b = characteristic constants (according to Menzerath's law $b$ is negative)

The formula on the left side is the famous Richardson formula analyzed by Mandelbrot in his work quoted above. It expresses the length of a coast measured by a stick of certain length. On the right side the formula of Menzerath's law is presented. The difference between these two structures consists in the value of their exponents: $b$ must be negative. If the coastline is approximated by a polygon, the number of its sides is $N = F\epsilon^{-D}$. Thus we obtain a completely analogical structure of the two mathematical expressions.

It is not difficult to recognize in the exponents of both expressions the variables interpretable as dimensions of the supposed phenomena - a dimension of a coast and a dimension of language level. The exponents express the degree to which the respective phenomenon is broken into constituents, i.e. how complicated it is. As far as language is concerned this is proper to the constant $b$. While in fractal geometry in the exponent the Hausdorff dimension $D$ was

determined by Mandelbrot, and this dimension is always positive, the negative value of $b$ seems to represent a problem for interpretation. However, in the fractal expressions the general consideration starts with constituents (represented by the length of stick) and arrives at constructs (the length of coastline), Menzerath´s law proceeds in the oppsite direction: it goes from construct down to mean constituents. If the formula is reversed, the exponent $b$ proves to be the inverse absolute value of $D$ and obtains a positive value. This dimension expresses the degree to which one of the investigated levels is broken into constituents. It expresses the inner dimension of language. ´To be broken´ means ´to have a structure´, ´to be diversified into elements´ or something in this sense. And thus again, the idea of communication is touched: structure bears or transmits information.

This can be demonstrated on the subsystems of aggregations:

A text having $k$ sentences is a system with Boltzmann entropy $k_B ln\ k$, where $k_B$ is Boltzmann´s universal constant. The same text contains $n$ words and, consequently, the entropy $k_B ln\ n$. These two entropies must be equivalent, as they concern one and the same system, so that we can write:

$$b\ ln\ k = ln\ n,$$

where $b$ is a coefficient of proportionality. This formula needs two kinds of corrections:

(1) There are $k$ sentences in the text in which the units are organized according to grammatical rules, therefore the value $ln\ k$ must be subtracted from the right-hand side of the formula.

(2) We know that the value for x = 1 plays a very important role in the system, and this value is expressed by the constant $A$ of Menzerath´s law. It occurs in all formulas for all sizes of the construct. Therefore $ln\ A$ has to be added to the left-hand side of the formula.

Both supplemented terms, $ln\ A$ and $ln\ k$, concern subsystems of the same system. Their entropy must be supposed to have the universal constant too. Thus they are expressed as $k_B\ ln\ A$ and $k_B ln\ k$. And finally, constant $k_B$ can be eliminated from the resulting equation. Thus we obtain

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

The same formula can be obtained when we proceed in the following way: The assumption of the maximal aggregation concerns all sentences of a given text and its mean sentence is equal to the mean sentence of the text. Menzerath´s law $y = Ax^b$ can be written for this case as $n/k = Ak^b$. Its logarithmic transformation results in the above formula, i.e. in $ln\ A + b\ ln\ k = ln\ n - ln$ k. Consequently, the relation between construct and constituents has a communicative interpretation in terms of entropy.



Self-similarity represents a kind of constant feature in text dynamics; self-similarity is a stem bearing the whole organization of the phenomenon. The task of science is to present an explanation saying why the organization has the observed form.

The semantic analogy between the concepts of self-similarity and self-organization may become the first step to the future unification of these two theories.

The difference between organization and *self-organization* is explained by H. Haken (1978: 191) using the example of a group of workers:

> ´Each worker acts in a well-defined way on given external orders, i.e. by a boss. Thus regulated behaviour results in a joint action. The same process is self-organized if there are no external orders, the workers work together by some kind of mutual understanding´.

In general, the dynamic systems are self-organized in the case when nothing intervenes in their course and arrangement from outside.

It was indicated above that language structures and systems cannot be explained in detail when communicators and their communication memories are not taken into account. Through communicators the language system is opened to external influences. Language constructs, and thus also texts, are systems always containing a communicator as its subsystem. Only in the case when communicators are considered together with the respective texts does the obtained system get closed to external influences, like the self-organized group of workers. This is true regardless of the fact that each communicator is open to all possible external factors through his/her communication memory. In this sense language represents a self-regulating system. Changes in the communicator´s mind have changes in the language constructs as their consequences. Changes in an arbitrary language structure are followed by changes of another structure of language. Dynamic structures and systems may vary and variations in one part imply changes in variations of other parts, etc.

This mechanism has its regularities. Classical linguistics behaves in relation to language as if it is completely explicable by a set of instructions given in advance to its users. However, language is opened to its users and thus subjected to fluctuations of different kinds. Communicators are generators of random events influencing language structures and systems. Self-similarity taken as a principle of language structures has as its consequence some total reactions to interventions coming from the generators of chance. Text as a ultimate unit encompassing the communicator as its subsystem can be described with the help of stochastic instruments. What else is applicable by text linguistics in this situation?

## Sequential character

Language texts are presented in sequences of sounds, letters, words, sentences, feet, strophes, etc. and these units, or parts of a sequence or the whole sequences fulfill a certain purpose. The purpose need not be known in special kinds of analysis, though its knowledge is advantageous. The units constituting the sequences have as many properties as we are able to conceive.

Since our reality takes its course in time we find time-dependent sequences everywhere. Most similar to language texts is a *film*, drafted first in the form of a language text. Its parts (units) have a lot of measurable properties (cf. Birett 1988).

A *melody* is a sequence of units having many analogies with words, phrases, sentences (cf. Orlov, Boroda & Nadarejšvili 1982). Even sequences of tones produced by birds are texts of their own, and there is merely a small step towards considering *behavior sequences* of living beings as texts. Animals have smaller inventories of behavior units than man but in both cases they can give rise to texts (cf. e.g. Pike 1967). If the number of degrees of freedom (i.e. the freedom of concatenating many different units) is very large then a sequential description of behavior is very circumstantial. In order to avoid this problem, we define greater units that allow us to make a lucid description of a sequence. Consequently, units represent our concepts; they result from our interaction with reality (cf. Salthe 1985).

In this sense we can consider *history* and, even the whole *evolution* as texts, since they consist of sequences of units (that were chosen or defined by us). Evolution is our longest text producing in turn means for producing other texts yielding its own description or reconstruction (language). Language in turn enables us to produce means for the description of these new language texts (mathematics), etc. In linguistics we have arrived at the point on the spiral line at which we try to catch language texts by means of mathematics and entrust other problems of evolution to other sciences.

There are of course also sequences having spatial form, e.g. the double helix, or sequences in many dimensions such as painting, architecture, sculpture, etc. etc.

In mathematics and statistics a number of methods for analyzing sequences has been developed, for example the theory of runs, the theory of discrete dynamic systems, the theory of Markov chains, time series, etc. It must be noted that it is possible to capture sequences of any linguistic units merely with 'stochastic' means though there are concatenation rules that seem to be almost deterministic, e.g. phonotactic, morphological, syntactical rules; however, they can be considered merely as means for increasing or decreasing the transition probabilities in the sequence and thereby generating redundancy. Thus grammar is merely one of the many sequential facets of the text.



## Multidimensional probability distributions

We do not always analyze a linguistic text sequentially, i.e. we are not always interested in the reconstruction of its birth and unfolding, but consider it as a ready whole. In this case it can be viewed as a multidimensional distribution of linguistic (and probably other) units since all its entities display frequency distributions. The values of random variables may be real numbers, integers, rankings or nominal classes, e.g. sound length measured in milliseconds, sentence length measured in number of clauses, ranks of units or classes of units. All these distributions are more or less strongly interrelated. Since it will probably never be possible to take into account all dimensions of textual random variables at the same time, we shall be forced to content ourselves with marginal distributions. Their examination is however of the utmost importance. The theory of probability provides us with an arsenal of distributions (cf. Johnson & Kotz 1969; Patil, Boswell, Joshi & Ratnaparkhi 1984; Altmann & Hammerl 1989; Altmann & Zörnig 1992), out of which several have already been used in text analysis.

Probability distributions are results of text processes and they represent an equilibrium state of these processes. Every property of text units has its own probability distribution.

## Processes and self-regulation

At first glance it seems that examining sequences and distributions would be a sufficient contribution of quantitative linguistics to text theory. Yet both sequences and distributions are generated by processes that are not directly observable. We can imagine that processes operate in a black box the opening of which is the aim of text research. In order to be able to look into this black box and bring forth explanations we must set up hypotheses about the underlying processes.

Since these processes jointly generate the text they are not independent of one another. Either they cooperate in the same direction or they compete with one another and thus set up a synergetic whole which is based on a harmonic self-regulation. The means of operation of these processes (Pr) can be deciphered in two ways: either by induction, starting from texts and looking at how they are constructed, or by deduction, starting from axioms telling how they should be constructed (cf. Fig. 2). The best way is of course to join these approaches. The mechanism hidden in the black box, i.e. a system of self-regulating processes, transforms an input into an output (text). However, while the output is observable and well known, the input can be captured only with the utmost difficulties since it consists of situations, cultural background, speaker's knowledge of language, linguistic and non-linguistic requirements of the

speaker, the kind and state of the hearer, the kind of channel, etc. All this is, of course, an object of research but lies at the periphery of text analysis (and linguistics). If we want to decipher the domain of processes, i.e. the mechanism operating in the black box we must set up simplified, very general assumptions (playing the role of axioms) by means of which at first isolated processes can be captured, and try to verify the results on many texts. The testing by means of statistics is a very delicate problem. At the beginning we shall notice that no simple model will be quite adequate, we shall always find deviations. Only when reaching a more mature state of research shall we be able to ascertain whether these deviations represent the omnipresent fluctuations or (at least in part) are results of other interfering processes.

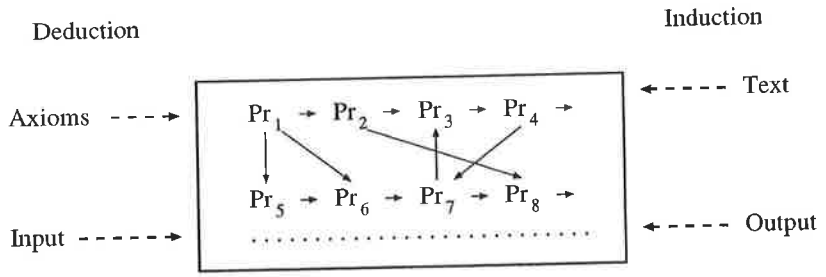



Figure 2. Approaches to text analysis

The processes in the black box build several control cycles which are, of course, connected with one another and at the same time are parts of superimposed control cycles. The processes generating texts are lawlike, i.e. they are present and identical with the production of any text. The outcomes of these processes are different because of boundary conditions which may be linguistic or non linguistic (physical, social, psychological, etc.) Their knowledge is of primary importance because they must be taken into account in modelling any text phenomena. None of the text processes is deterministic.

If we consider isolated dependences or processes we are seduced to the opinion that we must think causally, and our first successes in modelling encourage us to follow this path. Also the fact that with an isolated view we must take into account only a few boundary conditions can bring further succes for our explanation trials. However, if we consider language as a self-regulating system we come unavoidably to the result that our causal ideas can be successful only because we look merely in one direction and take into account only cutouts from a complex net. "Die Realität bestätigt, daß die exekutive Kausalvorstellung nur Fäden in den Netzmaschen, kaum die Maschen selbst und nie das ganze Netz abbildet" (Riedl 1982: 140). In a net no direction is more



important than any other. Consistent thinking must bring us to consider individual dependences and processes merely as functional components of self-regulating cycles. This view was initiated by G.K. Zipf (1949) and formally founded in linguistics by R. Köhler (1986, 1987a,b, 1989, 1990, 1991, cf. also Hammerl & Maj 1989). The first 'textual' control cycle comes from Nöth (1974, 1983).

The idea of self-regulation is extremely prolific since it can be applied to the whole reality (cf. Jantsch 1984) and its applicability depends merely on our ability to capture the necessary control cycle within the framework of our discipline. In this sense physicists have to cope with 'easier' problems than text scientists who must search for the components of their control cycles in psychology, culture, semiotics, linguistics, etc. The difficulty is enhanced by the fact that these 'boundary' disciplines have to go a long way to attain maturity in order to be applicable in text science.

Mathematics supplies us with a voluminous battery of methods applicable to processes and systems, such as the theory of stochastic processes, differential and difference equations, control theory, synergetics, etc. Rich literature can be found in any mathematical library.

## Requirements

Texts come into existence in order to fulfill certain purposes; they satisfy certain requirements or needs. These requirements are present with any act generating text. They are satisfied in that they initiate a process, modify a running process or even stop it. In Köhler's synergetic model the requirements are firm components of his system and enter it in the form of constants or functions (cf. Köhler 1986-1990). Processes effect here individual system quantitites and lead to a certain result. In language analysis this result is a change of language state, in text analysis it is a special text formation. With any text generation several requirements effect several processes at the same time. The result (text), even a not yet ready one, and its reception or author's assumption about its reception, operate as feedback. This leads to the rise of a synergetic whole.

If we want to build a theory of texts then all these background factors must be taken into account. It is not sufficient to perform concept formation i.e. to give names to requirements and processes. In the second step we must set up hypotheses about the effect of requirements on processes, about the possible kinds of processes, and about the changes or states which are the results of these processes. In the third step every deduction must be tested empirically. This way is evidently very long and more complicated than everything that has been done in theoretical linguistics so far.

It would, of course, be ideal to begin with requirements and to follow the path

requirements ⇒ processes ⇒ quantities ⇒ texts

but for the time being this way surpasses the possibilities of text analysts. Even requirements build a synergetic net and their investigation is in an embryonic state. Thus we are condemned to small steps.

## Conclusion

The eternal endeavour of man is to investigate how things are put together. Human beings really seek substances or nuclei; they break things into their parts and constituents in order to find some energetic surplus hidden in their interior, or something equivalently valuable.

The future task of text linguistics seems to be quite different. It tries to get together what was dissociated by classical linguistics in its analyses. To understand language with the help of simple semantic instruments, with the help of semantic intuition is possible; however, it was shown that it brings no answers to fundamental questions. We tried to indicate here that quantitative approaches are promising because they impose no prefabricated solutions.

Since questions concerning the *noumenon* or essence of things (texts) are either irrelevant or frustrating or classificatory or leading to conventional one-sided definitions, it is more expedient to choose an aspect and to try to set up a theory starting from the given point. Since linguists are (usually) no mathematicians, the aspect should be chosen at the beginning (a) in connection with a well developed statistical or mathematical discipline which removes the necessity of developing new mathematics (cf. Gordesch 1991), (b) *per analogiam* i.e. asking what phenomena can be considered as texts, what is present in texts that has been found or established elsewhere, how texts behave, etc. This enables us to make generalizations going beyond the domain of texts and removes the necessity of developing new methodologies. (c) To start from Köhler's synergetic model and extend it to texts. Whatever way we choose no future theory of texts will be possible without considerations analogous to those of Köhler.

## References

**Altmann, G.** (1972). Status und Ziele der quantitativen Linguistik. In: Jäger, S. (ed.), *Linguistik und Statistik.* Braunschweig, Vieweg 1972: 1-9.

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

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



**Altmann, G.** (1991). Modelling diversification phenomena in language. In: Rothe, U. (ed.), *Diversification processes in language: grammar.* Hagen, Rottmann 1991: 33-46.

**Altmann, G. & Hammerl, R.** (1989). *Diskrete Wahrscheinlichkeitsverteilungen I.* Bochum, Brockmeyer.

**Altmann, G. & Zörnig, P.** (1992). *Diskrete Wahrscheinlichkeitsverteilungen II.* Bochum, Brockmeyer.

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

**Arapov, M.V.** (1988). *Kvantitativnaja lingvistika.* Moskva, Nauka.

**Birett, H.** (1988). Statistische Filmästhetik oder Ist Absicht zufallsverteilt? In: Bluhme, H. (ed.), *Beiträge zur quantitativen Linguistik.* Tübingen, Narr 1988: 180-182.

**Bunge, M.** (1961). The weight of simplicity in construction and assaying of scientific theories. *Philosophy of Science 28, 120-149.*

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

**Gordesch, J.** (1991). *Statistische Datenverarbeitung in der Textanalyse.* Berlin, Freie Universität, Institut für Soziologie.

**Haken, H.** (1978). *Synergetics.* Berlin-Heidelberg-New York, Springer.

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

**Hammerl, R. & Maj, J.** (1989). Ein Beitrag zu Köhlers Modell der sprachlichen Synergetik. *Glottometrika 10, 1-31.*

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

**Hřebíček, L.** (1990a). Menzerath-Altmann's law on the semantic level. *Glottometrika 11, 47-56.*

**Hřebíček, L.** (1990b). The constants of the Menzerath-Altmann law. *Glottometrika 12, 61-71.*

**Hřebíček, L.** (1991). Text as a construct of aggregations. In: *QUALICO 91.* Abstracts. Trier, University of Trier: 36-41.

**Hřebíček, L.** (1992). *Text in communication: suprasentence structures.* Bochum, Brockmeyer.

**Jantsch, E.** (1984). *Die Selbstregulation des Universums.* München, DTV.

**Johnson, N. L. & Kotz, S.** (1969). *Discrete distributions.* Boston, Houghton Mifflin.

**Köhler, R.** (1984). Zur Interpretation des Menzerathschen Gesetzes. *Glottometrika 6, 177-183.*

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

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

**Köhler, R.** (1987b). Sprachliche Selbstregulation als Mechanismus des Sprach-

wandels. In: Boretzky, N., Enninger, W., Stolz, Th. (eds.), *Beiträge zum 3. Essener Kolloquium über Sprachwandel und seine bestimmende Faktoren.* Bochum, Brockmeyer 1987, 185-200.

**Köhler, R.** (1989). Linguistische Analyseebenen, Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation. *Glottometrika 11, 1-18.*

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

**Köhler, R.** (1991). Diversification of coding methods in grammar. In Rothe, U. (ed.), *Diversification processes in language: grammar.* Hagen, Rottmann, 47-56.

**Mandelbrot, B.** (1977). *The fractal geometry of nature.* New York, Freeman.

**Menzerath, P.** (1928). Über einige phonetische Probleme. In: *Actes du premier congrès international de linguistes.* Leiden, Sijthoff: 104-105.

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

**Miller, G.A.** (1957). Some effects of intermittent silence. *The American Journal of Psychology* 70, *311-314.*

**Miller, G.A.** (1977). Practical and lexical knowledge. In: Johnson-Laird, P.N. & Wason, P.C. (eds.), *Thinking. Readings in Cognitive Science.* Cambridge, Cambridge University Press 1977: 400-410.

**Nöth, W.** (1974). Kybernetische Regelkreise in Linguistik und Textwissenschaft. *Grundlagen der Kybernetik und Geisteswissenschaften 15, 75-86.*

**Nöth, W.** (1983). Systems theoretical principles of the evolution of the English language and literature. In: Davenport, M., Hansen, E. & Nielsen, H.G. (eds.), *Current topics in English historical linguistics.* Odense UP, 103-122.

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

**Patil, G.P., Boswell, M.T., Joshi, S.W. & Ratnaparkhi, M.V.** (1984). *Dictionary and classified bibliography of statistical distributions in scientific work.* Fairland, Maryland, International Co-operative House.

**Pike, K.L.** (1967). *Language in relation to a unified theory of the structure of human behavior.* The Hague, Mouton.

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

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

**Salmon, W.C.** (1984). *Scientific explanation and the causal structure of the world.* Princeton, Princeton University Press.

**Salthe, S.N.** (1985). *Evolving hierarchical systems. Their structure and representation.* New York, Columbia UP.

**Shannon, C.E. & Weaver, W.** (1949). *The mathematical theory of communication.* Urbana, The University of Illinois Press.

**Zipf, G.K.** (1949). *Human behavior and the principle of least effort.* New York, Hafner.

# Computer-Aided Formation of Concepts

*Johannes Gordesch & Antje Zapf, Berlin*

## 1 Introduction

Cognition comprises knowledge, knowledge acquisition, and knowledge processing. Cognitive processes pertain to the transformation, reduction, processing, storage, and reactivation of knowledge. They are, in general, accompanied by or related to language. Linguistics is the professional study of language and provides valuable tools for information processing. Thinking in particular is based on chunks of information which contain concepts as well as a variety of references connected with it, and on various types of logical operations joining the elements of information together. Text corpora are an outstanding basis for investigating language-related aspects of cognition, and techniques developed in corpus linguistics form the skeleton of many methods of processing and analyzing texts.

## 2 Formation of concepts

**2.1 Meaning.** Casually speaking, *meaning* is connected with the various aspects of our understanding of words, sentences, and texts, and our ability to endow them with a symbolic function. *Confirmation* is the relation between propositions when one supports or adds credence to another. It is restricted by the quest for established criteria to decide what kind of propositions is confirmable, and ends up with Feyerabend's epistemological anarchy that meaning is totally dependent on context.

Another aspect of meaning is its connection with pragmatic aspects such as wanting or intending, and with human conventions and rules. Thus meaning is found by analyzing how it is normally used by those intending to bring about a certain result in an audience.

Thirdly, meaning is connected with other semantic notions, such as reference or truth. Model theory defines validity, consequence, independence etc. via set theoretic interpretations. It equates meaning with truth-conditions, and the recognition of the meaning of grammatically complex sentences is reduced to rule.

The word 'semantics' shows a broad spectrum of meaning. It comprises such diverse notions as linguistic semantics, philosophical and logical semantics,

psychological semantics, semantics as used in computer science or in communication studies. For the purpose of knowledge representation, the more pragmatic integrative approaches have proven useful. Two of them, however, seem to fit best to the intuitive comprehension of meaning, namely in a first approximation predicate semantics, and a more refined approach based on AI techniques (i. e. knowledge representation). Predicate semantics is closely related to the enumeration of the elements of an ordered assemblage of semantemes (informemes, attributes), the standard definition of entries in a lexicon. This fact also corresponds to the classical theory of concepts which distinguishes between intension (or connotation, i. e. the set of defining attributes) and extension (or denotation, i. e. the set of characterized objects). Since predicates (in the sense of logic) rather than single words or lexemes correspond to concepts, phrases (in the sense of linguistics) are used, thus comprising also syntactical knowledge.

The AI approach of meaning ("meaning is knowledge of a certain subject") seems most natural: we possess detailed knowledge of a certain subject denoted by a single lexeme or a complex phrase, and techniques of knowledge representation furnish appropriate tools for implementing the knowledge we have.

**2.2 Frege's resp. Russel's theory of concepts.** A *concept* is identified with a *propositional function* A(x) which is true for the elements of the extension of the concept, or with a predicate (or class). An *attribute* consists of a lexeme (or set of synonyms) and/or its fuzzy value (cf. the relevant section below) plus grammatical, semantic etc. tags and/or their fuzzy values.

**2.3 Paul Lorenzen's operational theory of concepts.** *Proper names* are words denoting objects, while *predicators* are words that are assigned (attributed) or not assigned (attributed) to proper names. This process is called *predication* (cf. case grammar). Predication exists in all languages. It may be introduced by examples or by rules (e. g. using implication or other logical operators; cf. knowledge representation by rules). Concepts are equivalence classes of predicators (equivalence of meaning). Two predicators P, Q are of the same meaning if $P \rightarrow Q$ and $Q \rightarrow P$ (or $P \equiv Q$).

## 3 Knowledge representation and conceptual analysis

**3.1** Artificial Intelligence techniques penetrate more and more every field of electronic data processing, and knowledge representation is in the centre of it. AI aims at the investigation of cognitive abilities of human beings and its representation and modelling on a computer. It heavily depends on fossilized interaction between human beings and their environment.

*Knowledge* consists of true propositions that can be 'explained' or 'justified', i. e. (more or less convincing) reasons can be given for their truth. *Knowledge representation* is the representation of real world facts in some formalism chosen so that information processing can be performed as efficiently as possible. The most severe problems arise with knowledge that cannot or need not be communicated since it seems natural (self-evident) or remains unconscious.



There is more than one type of knowledge representation, and no general recipe exists. Knowledge representation is a rapidly developing topic, yielding rather intermediate than final results. Not without reason, the popular text book on AI written by Alan Rich and Kevin Knight defines AI as "... the study of how to make computers do things which, at the moment, people do better."

Knowledge representation is not identical with the development of new data structures, for the solution of classes of problems shall be sought for independently from special computer implementations. Knowledge representation also differs from databases: the absence of appropriate techniques of abstraction separate databases and knowledge representation systems.

**3.2** The computer lacks the intelligence human beings possess. In order to compensate for this deficiency, two main approaches exist: annotation, i.e. putting the information into the text, and artificial intelligence, introducing knowledge once and for all by conceiving appropriate rules etc. The second approach looks more promising because of its convenience; unfortunately, sophisticated AI systems are rather difficult to devise, and their results are often disappointing.

Additionally, texts may have to be reformulated in a nearly formal language, e.g. by some kind of tagging (grammatical, semantic, ...), thus removing ambiguity or vagueness and contradictions.

Not every type of grammar is well apt for automatic analysis. In this context, cf. notional grammar, word grammar, constituents grammar, transformational grammar, case grammar, dependency (valency) grammar, functional grammar, head-driven phrase-structure grammar (HPSG), unification grammar, augmented transition network grammar (ATNG). Certain unification-based grammars seem to fit best to knowledge representation on a computer, whereas ATNGs are most appropriate to programming techniques in general.

**3.3** Techniques of knowledge representation comprise relatively stable and basic structures mapping reality. Some of them are well known in psychology (schemes, scripts) or other scientific disciplines, while others are only known to experts in AI. A short overview of the most common types of representation is given below. It suffers, however, from the fact that no strict and compelling definitions exist, and rather different types of knowledge representation are summarized under the same name.

**3.4 Conceptual Dependency.** *Conceptual dependency* utilizes the information contained in natural language textemes (often sentences). It is a means to represent the knowledge in a way that

- facilitates drawing inferences;
- is independent of the language in which the textemes were originally stated.

To reach independence from the choice of a particular language, it grounds on semantemes (concepts, constructs, ...) rather than words etc. Conceptual dependency is related to semantic networks; it is, however, a much more powerful tool, especially in natural language processing.

Conceptual dependency as a means of knowledge representation was introduced in Schank 1973, and has been developed further in Schank 1975.

**3.5 Construct.** A *construct* consists of the following components:

(1) A structured set of concepts which are conceived to be used for theoretical purposes as well as in empirical research. Thus it clearly distinguishes from *concept*, i. e. the representation of a variety of entities by a single denotation, and which results from abstraction and generalization.

(2) In the case that the concepts of the construct do not serve as operationalization, they must be made observable, i. e. operationalized.

(3) Pertaining to a certain system of concepts, a description or explanation of complex processes, actions, or states that are considered to form a functional unit. In this respect a construct resembles a model or a theory.

Constructs related to a fixed system form a lattice. If an element of the lattice is determined by its predecessor and its successor, it plays the role of an intervening variable (MacCorquodale & Mechl 1948).

**3.6 Frames.** A *frame* is a collection of attributes (in AI often called slots) and associated values (and possibly, constraints or values) that describe some entity in the real world. Frames were introduced by Minsky 1975. *Frame* is a rather general, not to say vague, concept. It is of best use if an attribute of a frame is a frame itself (inheritance principle). Hence a frame is a class (or set) or an instance (an element of a class), and the governing relations are 'is a subset of' ('is a', a transitive relation) and 'is an element of' ('instance'). Vector spaces where coordinates may be vectors themselves (property of *inheritance*) are a particularly useful type of frame in the case of hybrid grammars.

**3.7 Logic.** Logic may be used in two distinct ways. *Declarative representation* (knowledge representation by logic in the strict sense) specifies knowledge only, whereas the use to which that knowledge is to be put is given in a separate program ('logic as data'). Knowledge may also be encoded by using rules (*procedural representation* or *representation by rules*, 'representation by program').



Rules formulated with the aid of propositional logic and predicate logic constitute, at least at first sight, an appealing means of knowledge representation, for they suggest the deduction of new knowledge, and they reflect certain syntactic structures quite well. For the purpose of natural language processing, however, they are often superseded by other techniques.

**3.8 Schemes.** In psychology, a (cognitive) *scheme* is a general cluster of stored knowledge that controls cognitive processes and anticipates knowledge of objects, human beings, and situations. Schemes are thought to be the primary units of information processing systems, and are connected with general categories.

**3.9 Scripts.** A *script* is a cluster of knowledge on sequences of related events or actions. Similarly, it is a symbolic representation of sequences of events into which new details (single informational elements) may be integrated. It deals with procedural knowledge, whereas a *scheme* pertains to declarative knowledge. GPSS (General Purpose Simulation System) and other simulation languages (SIMULA, SIMSCRIPT etc.) provide more detailed and more precise counterparts in electronic data processing.

**3.10 Semantic Networks.** *Semantic networks* (also called *semantic nets, lexeme nets* etc.) had been conceived as a way to represent labeled connections among entities. There is no clear distinction between semantic nets and frame systems. By assigning more structure to nodes as well as to links, however, semantic nets gradually grow into frame systems. They have become very popular in modern linguistic work.

## 4 Example of Application

As a typical example of application the concept 'organization' will be represented by a semantic network and equivalent means. It will be part of a planned computerized thesaurus of the social sciences, and is used for teaching purposes (cf. Gordesch & Gärtner 1992) as well as in computer aided translation. It reflects the tendency in lexicography to use definition techniques other than those related to the classical theory of concepts based on intension (assemblage of attributes) and extension (assemblage of objects). Cf. among sundry others Sambor & Hammerl (1991).

*Organization* is defined as "a type of collectivity for the pursuit of specific aims or goals, characterized by a formal structure of rules, authority relations, a division of labour and limited membership or admission" (*Collins Dictionary of Sociology* 1991). Starting with this or similar definitions found in published work on organization theory and after some analysis and abstraction, a skeletal

structure (in this case a semantic network) comprising categories as well as their interrelationship is obtained. The categories are listed in a table, and the whole network consisting of the categories and their interrelation is depicted by an (evaluated) graph. Then the matrix of the graph and the functional representation are put down. Matrix and graph are strictly equivalent: the graph leads to the matrix, and from the matrix the graph can be drawn. The functional representation is not unique; much interpretation enters it, and the same problems are encountered as in analytical philosophy, where propositional and predicate logic have been used extensively.

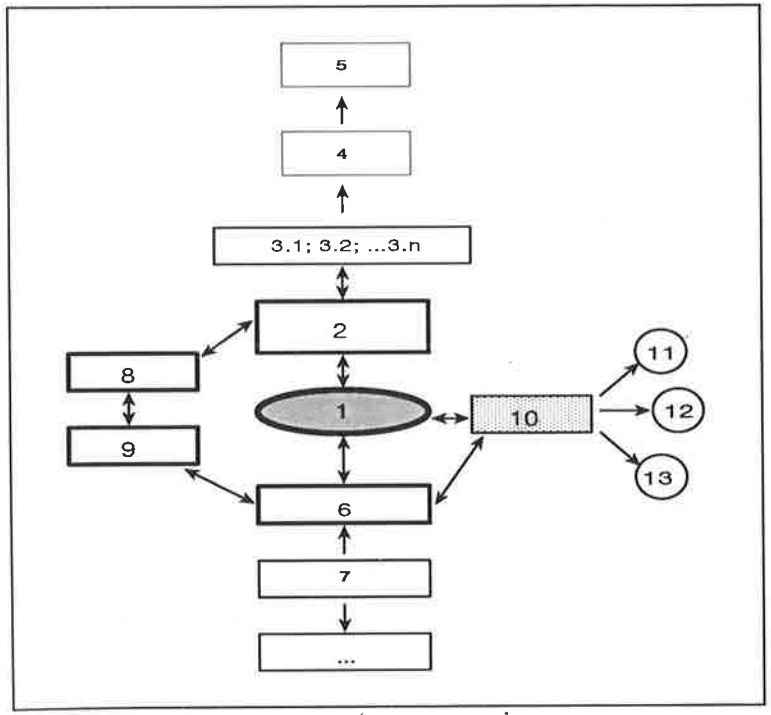
The concept defined this way provides a guideline for scientific research as well as for exchange of information (man - man, man - machine, machine - machine). The more refined the categorical system is the more efficient the exchange of information will be, and in particular interdisciplinary work will be facilitated.

Each node of the graph has to be operationalized by rules applicable to specialized corpus material on organizations. Then clustering yields the classification of possible organizations (e. g. firms, educational units, etc.).

| | |
|---|---|
| 1 | ORGANIZATION |
| 2 | STRUCTURAL ANALYSIS |
| 3 | ARRANGEMENT, CONSTRUCTION, ORDER, PATTERN, PLAN |
| 4 | UNIT COMPOSED OF INTERACTING ENTITIES |
| 5 | COLLECTIVITY ESTABLISHED FOR THE PURSUIT OF SPECIFIC AIMS OR GOALS |
| 6 | PROCESS ANALYSIS |
| 7 | ORGANIZING ETC. |
| 8 | FUNCTION |
| 9 | BEHAVIOUR |
| 10 | SEPARATION FROM ENVIRONMENT |
| 11 | LIMITATION OF SYSTEM BEHAVIOUR |
| 12 | CONSERVATION OF THE INDIVIDUALITY OF THE SYSTEM |
| 13 | CONSERVATION OF THE PERFORMANCE OF THE SYSTEM |

CATEGORIES OF SEMANTIC NETWORK 'ORGANIZATION'






GRAPH OF THE SEMANTIC NETWORK 'ORGANIZATION'

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | * | | | | * | | | | * | | | |
| 2 | * | | | | | | | * | | | | | |
| 3 | | * | | * | | | | | | | | | |
| 4 | | | | | * | | | | | | | | |
| 5 | | | | | | | | | | | | | |
| 6 | * | | | | | | * | | * | * | | | |
| 7 | | | | | | * | | | | | | | |
| 8 | | * | | | | | | | * | | | | |
| 9 | | | | | | * | | * | | | | | |
| 10 | * | | | | | * | | | | | * | * | * |
| 11 | | | | | | | | | | | | | |
| 12 | | | | | | | | | | | | | |
| 13 | | | | | | | | | | | | | |

MATRIX REPRESENTATION OF THE SEMANTIC NETWORK 'ORGANIZATION'

| | |
|---|---|
| 1 ∨ (2 ∧ 6 ∧ 10) | CENTRAL CONCEPT |
| 2 ∨ (1 ∧ 8) | STRUCTURAL ANALYSIS |
| 3 ∨ (2 ∧ 4) | |
| 4 ∨ 5 | |
| 5 | |
| 6 ∨ (1 ∧ 7 ∧ 9 ∧ 10) | PROCESS ANALYSIS |
| 7 ∨ 6 | |
| 8 ∨ (2 ∧ 10) | FUNCTION/ BEHAVIOUR |
| 9 ∨ (6 ∧ 8) | |
| 10 ∨ (1 ∧ 6 ∧ 11 ∧ 13 ∧ 14) | SEPARATION FROM ENVIRONMENT |
| 11 | LIMITATION RESP. CONSERVATION OF SYSTEM PATTERNS |
| 12 | |
| 13 | |

FUNCTIONAL REPRESENTATION OF THE SEMANTIC NETWORK 'ORGANIZATION'

Representing concepts by semantic networks can be improved in several ways. The simplest and most natural way is to replace the defining undirected graph by a directed one (as done above), by an evaluated graph where the strength of the relationship between two nodes (subconcepts) is assessed by a number say between 0 and 1, or by a multigraph where different types of relations are used (cf. the entity-relationship model of data representation). Replacing the asterisks of the table by numbers between 0 and 1, writing 0 in the empty cells if a relationship is possible though not realized and leaving it empty if it is not, the matrix representation can easily be extended to cover the more general case of an evaluated graph. The pertaining matrix of 'organization' is given below, while the graph itself resp. its functional representation are omitted here.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | 1.0 | | | | 1.0 | | | | 1.0 | | | |
| 2 | 1.0 | | | | | | | 0.8 | | | | | |
| 3 | | 0.9 | | 0.7 | | | | | | | | | |
| 4 | | | | | 0.5 | | | | | | | | |
| 5 | | | | | | | | | | | | | |
| 6 | 1.0 | | | | | | 0.8 | | 0.8 | 0.8 | | | |
| 7 | | | | | | 0.8 | | | | | | | |
| 8 | | 0.2 | | | | | | | 0.8 | | | | |
| 9 | | | | | | 0.8 | | 0.8 | | | | | |
| 10 | 1.0 | | | | | 0.8 | | | | | 0.5 | 0.5 | 0.5 |
| 11 | | | | | | | | | | | | | |
| 12 | | | | | | | | | | | | | |
| 13 | | | | | | | | | | | | | |

MATRIX REPRESENTATION OF THE EVALUATED SEMANTIC NETWORK 'ORGANIZATION'



In a similar manner the constituents (nodes) of the semantic network can also be assessed. Fuzzy set theory would provide an adequate mathematical tool. It is dealt with further below.

## 5 Statistical Modelling

*Statistics* is the assembly and mathematical analysis of numerical data. *Descriptive statistics* involves

- the reporting, the summary, and the graphical representation of data;
- the use of statistics (variates) describing association, correlation, functional relationship of data.

In addition, in *inferential statistics*, on the basis of probability theory and random sampling, inferences are drawn from one sample to another, or from a sample to the pertaining total population. Extensive material may be cut down by deliberate selection of parts of it with the object of investigating the properties of the whole (cf. Gordesch 1991a). In that instance, infinitely often repeatable phenomena are assumed to conform to an underlying probabilistic model.

The formation of concepts necessitates putting together pieces of information to make up 'chunks', 'clusters', 'constructs', 'frames', 'scripts', whichever word one may prefer. Hence various measures of association (t-scores etc.) as well as algorithms for performing these actions (cluster algorithms) are used.

Statistical methods are an indispensable tool for solving many of the problems encountered in empirical research. Nonetheless, the population 'language' does not exist; in reality, language is a collection of several more or less distinct varieties etc., and corpora are not random samples but collected for the sake of convenience. Thus statistics as used in the Church/ DeRose algorithms (Church 1988; DeRose 1988) to disambiguate word categories, or in the association ratio algorithm (Church & Hanks 1989) for the extraction of co-occurrences (Smadja 1989, Smadja & McKeown 1990) provide heuristic techniques applied to a certain corpus material. However, they do not allow generalization of the findings to English as a whole.

## 6 Fuzzy sets and fuzzy logic

**6.1** One of the most severe problems of semantics is the indeterminacy and vagueness of meaning, which, at least to some degree, may be overcome by context (cf. also confirmation). Nevertheless, special techniques to deal with context sensitivity etc. are necessary. Fuzzy set theory replaces rigid concepts and rigid logical relations by more flexible ones. In the sequel, we take the

classical view that a concept comprises extension (or denotation, the set of entities to which the concept correctly applies) as well as intension (or connotation, the set of attributes characterizing the 'meaning' of the concept), and confine ourselves to propositional and predicate logic. For the following, refer to Gordesch 1991b.

**6.2 Additivity versus subadditivity.** 'Weights' (e. g. measuring the importance) may be attributed to characteristics etc. In some cases (e. g. for relative frequencies) it may be known that these weights obey the laws of probability calculus, above all full additivity:

$$\mu(A \cup B) = \mu(A) + \mu(B),\ A \cap B = \varnothing.$$

In many cases, however, this would be mere fiction (e. g. in the case of giving the degree of rational belief), and a less restrictive condition may be imposed (subadditivity):

$$\mu(A \cup B) \leq \mu(A) + \mu(B),\ A \cap B = \varnothing.$$

Several mathematical theories may serve as a substitute for probability theory; the most successful and best known is fuzzy set theory, introduced by ZADEH a few decades ago.

**6.3 Fuzzy sets.** A fuzzy set is defined as the set of mappings $V^A$ of a set A into a set V of valuations: fuzzy subsets are specified as sets of pairs

$$\{\ \{(a,v), \ldots\} \mid a \in A,\ v \in V\}.$$

Thus instead of sets alone always subsets (and thus also elements) together with assessing values are considered. In Zadeh's original definition, V is the closed real interval [0, 1]. More generally, V is chosen to be an ordered set (so as to cover also logical systems) or even a lattice (Goghen 1967). As an example, consider the fixed collocations 'by some strange freak' (a) and 'film freak' (b), and assess how strong the collocation is by the values l (low), m (medium), and h (high). Then

$$\{l,m,h\}^{\{a,b\}}$$

denotes the following set of mappings (all possible combinations of collocations and valuations):

| | | |
|---|---|---|
| {(a,l), (b,l)}, | {(a,l), (b,m)}, | {(a,l), (b,h)}, |
| {(a,m), (b,l)}, | {(a,m), (b,m)}, | {(a,m), (b,h)}, |
| {(a,h), (b,l)}, | {(a,h), (b,m)}, | {(a,h), (b,h)}. |



The meaning of 'valuation' is open, and interpretation ranges over topics like characteristic functions in set theory ('this element belongs to the subset A'), many-valued logic ('the nuances of truth'), and probability-like calculi (e. g. probability calculus itself, infinitely-valued logic).

In order to guarantee continuity, the involved logical operators (negation, conjunction, disjunction, implication, equivalence) shall behave for V = {0,1} exactly as in traditional propositional calculus. Depending on the selection of V and the definition of the logical operators, varied logical systems can be obtained.

**6.4 Fuzzy propositional logic.** Logical operators of ordinary propositional calculus may be represented by numerical functions, as is common in computer algorithms. Negation means complementing p to 1, conjunction taking the smaller of two values, and disjunction the greater. Admitting also values between 0 and 1 may provide a more refined measure of 'being true'.

| Operation | Infix Operator | Prefix Operator | Fuzzy Operation |
|---|---|---|---|
| Negation | $\neg p$ | Np | 1-p |
| Conjunction | $p \wedge q$ | Cpq | min(p,q) |
| Disjunction | $p \vee q$ | Dpq | max(p,q) |
| Implication | $p \rightarrow q$ | Ipq | max(1-p,q) |
| Equivalence | $p \equiv q$ | Epq | min(max(p,1-q), max(q,1-p)) |

In another approach, the logical operators conjunction, disjunction, implication, and equivalence are reached determining the coefficients of the function

$$F(p,q) = \alpha p + \beta q + \gamma|p-q| + \delta$$

so that the respective surfaces pass through the points

(Cpq) = (111), (010), (001), (000),
(Dpq) = (111), (110), (101), (000),
(Ipq) = (100), (101), (111), (010),
(Epq) = (111), (010), (001), (100).

These conditions guarantee that the employed definitions do conform to ordinary propositional calculus, in particular that equivalence Epq means the implication in both directions, Ipq and Iqp. Generalization of propositional calculus to fuzzy propositional logic, then, is straightforward.

If the truth values 0 and 1 are chosen, ordinary propositional calculus results; and for V = {0,½,1}, a system similar to Kleene's three-valued logic is obtained. If these truth values are replaced by real numbers between 0 and 1, a fuzzy propositional calculus is obtained. Fuzzy sets may be thought of as a surrogate for probability measures.

| Operation | Infix Operator | Prefix Operator | Fuzzy Operation |
|---|---|---|---|
| Negation | $\neg p$ | Np | 1-p |
| Conjunction | $p \wedge q$ | Cpq | min(p,q) or ½[(p+q)-\|p-q\|] |
| Disjunction | $p \vee q$ | Dpq | max(p,q) or ½[(p+q)+\|p-q\|] |
| Implication | $p \rightarrow q$ | Ipq | 1-½[(p-q)+\|p-q\|] |
| Equivalence | $p \equiv q$ | Epq | 1-\|p-q\| |

Corresponding expressions can be deduced for more than two variables.

**6.5 General approach.** Conjunction and disjunction of fuzzy items may be defined by any appropriate function, e. g. by

$$Cp_1p_2...p_n \rightarrow \Sigma p_i/n - 2\Sigma\Sigma|p_i-p_j|/(n(n-1)),$$
$$Dp_1p_2...p_n \rightarrow \Sigma p_i/n + 2\Sigma\Sigma|p_i-p_j|/(n(n-1)).$$

$[p_1,p_2, \ldots p_n]$ attributes, propositions, etc.,
$0 \le p_i \le 1$ normalization of fuzzy values,
$i,j = 1 \ldots n, \; i<j$.

$\Sigma p_i/n$ is the arithmetic mean, and $2\Sigma\Sigma|p_i-p_j|/(n(n-1))$ equals Gini's mean distance, a measure of dispersion. The mean comprises all single values. The conjunction remains true if all propositions are true, the disjunction if at least one proposition is true. Inhomogeneity (variability of the fuzzy values) lessens the conjunction, but increases the disjunction. Certainly, other (and more complex) definitions also make sense. A single result will not be very meaningful when deciding which method to choose; the comparison of the resulting values as well as their validation in known cases, however, will facilitate a reasonable decision.

**6.6 Linguistic variables.** The concept of `linguistic' variables (a misleading term) is of special interest. Informally, a linguistic variable is a variable the values of which are ordinary words or expressions like 'simple', `foolish', 'silly', `fatuous', 'asinine'. These values are ordered, i. e. they show various degrees of 'actual or apparent deficiency in intelligence'. Intensifying adjectives often present counter-

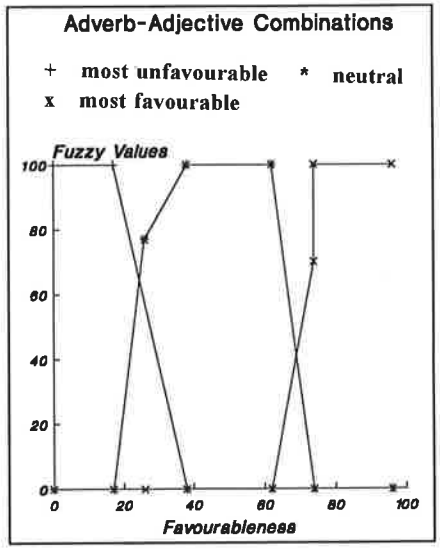


examples: 'great', 'large', 'bitter', 'big', 'definite', 'distinct', 'marked' hardly change their meaning, but their use depends on the noun (a big/ bitter/ great disappointment, a big/ definite/ distinct/ marked improvement). Adverb-adjective combinations like 'extremely disgusting' ($l_1$), 'rather evil' ($l_2$), 'slightly bad' ($l_3$), ... 'very good' ($l_n$) range from 'most unfavourable' via 'neutral' to 'most favourable'. They may serve as an example in the more formal definitions.

Formally, a linguistic variable is a finite labeled and ordered collection of fuzzy sets:

$l_1$: $v = T_1(x)$ for $x_0 \le x < x_1$
$l_2$: $v = T_2(x)$ for $x_1 \le x < x_2$
⋮
$l_n$: $v = T_n(x)$ for $x_{n-1} \le x < x_n$

$l_1 < l_2 < \ldots < l_n$ labels,
v valuations,
T, $T_i$ truth functions,
x, $x_i$ latent ordering variable
resp. fixed values of it



Linguistic Variables

Fuzzy sets pertaining to the same as well as to different variables may be combined by logical operators as defined above. If a collection $A$ of subsets of $A = \{l_1, l_2, \ldots\}$ bears the structure of a Boolean lattice, we obtain a many-valued propositional calculus, and which is denoted by the quadruple

$$(A, \mathcal{A}, V, T{:}A \rightarrow V),$$

and where

$T = T_1$ for $x_0 \le x < x_1$
$T = T_2$ for $x_1 \le x < x_2$
⋮
$T = T_n$ for $x_{n-1} \le x < x_n$.

## References


**Abelson, R. P.** (1981). The psychological status of the script concept. *American Psychologist 36, 715 - 729*

**Bandemer, Hans & Siegfried Gottwald** (1989). *Einführung in FUZZY-Methoden.* Berlin, Akademie-Verlag (2nd ed.).

**Bocheński, I. M. & Menne, A.** (1983[5]). *Grundriß der formalen Logik.* Paderborn, UTB Schöningh.

**Church, K.** (1988). A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. *Second Conference on Applied Natural Language Processing.* Austin, Texas

**Church, K. & Hanks, P.** (1989) Word Association Norms, Mutual Information, and Lexicography. *Proceedings of the 27th Annual Meeting of the ACL,* Vancouver.

**DeRose, S.** (1988). Grammatical Category Disambiguation by Statistical Optimization. *Computational Linguistics,* Vol. 14, No.1

**Dougherty, Edward R. & Giardina, Charles R.** (1988) .*Mathematical Methods for Artificial Intelligence and Autonomous Systems.* London, Prentice-Hall.

**Frege, Gottlieb** (1879). *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens.* Halle.

**Ganter, B., Wille, R. & Wolff, K.E. (eds.)** (1987). *Beiträge zur Begriffsanalyse.* Mannheim-Wien-Zürich, BI Wissenschaftsverlag.

**Goghen, J.** (1967). L-Fuzzy Sets. *Journal of Mathematical Analysis and Applications 18, 145-174.*

**Goodman, S. E. & Hedetniemi, S.T.** (1977). *Introduction to the Design and the Analysis of Algorithms.* Tokyo, McGraw-Hill Kogakusha.

**Gordesch, Johannes** (1978). On a Generalized Concept of Probability. In: J. Gordesch & P. Naeve (eds.), *COMPSTAT 1976. Proceedings in Computational Statistics.* Würzburg - Wien, 18-29.

**Gordesch, Johannes** (1991). Statistical Modelling. In: H. Best, E. Mochmann & M. Thaller (eds.), *Computers in the Humanities and the Social Sciences. Achievements of the 1980s, Prospects of the 1990s.* München, Saur etc., 356-361.

**Gordesch, Johannes** (1991a). Stichprobenverfahren bei der Auswertung von Corpora. In: Mark P. Line & Josef Wallmannsberger (eds.), *Computer und Sprache. Papiere des Workshops Universität Saarbrücken 23.-25.11.1989. Innsbrucker Beiträge zur Kulturwissenschaft, Anglistische Reihe*, Vol. 3. Innsbruck, 131 - 149.

**Gordesch, Johannes** (1991b). Fuzzy Reasoning. *Mitteilungen des Schwerpunktbereichs Methodenlehre, Institut für Soziologie, Freie Universität Berlin*, Nr. 26, Berlin. Also to be published in: *Proceedings of AHC '91, Odense, 1992.*

**Gordesch, Johannes** (1992) Probleme formaler Modelle in den historischen Wissenschaften. *Neue Methoden der Analyse historischer Daten. HSF* Vol. 23, Köln

**Gordesch, Johannes, Salzwedel, H. & Siggelkow, I.** (1990). Artificial Intelligence Techniques for Bibliographies. In: A. M. Tjoa & R. Wagner (eds.), *DEXA 90. Database and Expert Systems Applications. Proceedings of the International Conference in Vienna, Austria, 1990.* Wien, 537-541.

**Gordesch, Johannes, Salzwedel, H. & Siggelkow, I.** (1992). A Computer-Aided Theory of Concepts and its Application to Historical Research: Social Space and Historical Space. In: Josef Smets (ed.), *Histoire and Informatique. Actes du 5ième congrès international de la 'Association for History and Computing', 4-7 septembre 1990, Montpellier.* Montpellier 1992. French abridged version: Une théorie des notions et son application à la recherche automatisée: l'espace social et l'espace historique. *Montpellier Computer Conference 1990, Volume des résumés.* Montpellier 1990, 44-48

**Gordesch, Johannes & Gärtner, Frank** (1992). Eine Workbench für die Sozialwissenschaften. *Microcomputer-Forum für Bildung und Wissenschaft,* Vol. 5. Berlin-Heidelberg-New York, Springer.

**Harmon, Paul, Maus, Rex & Morrissey, William** (1988). *Expert Systems Tools and Applications.* New York, Wiley.

**Hofbauer, Dieter & Kutsche, Ralf-Detlef** (1989). *Grundlagen des maschinellen Beweisens.* Braunschweig-Wiesbaden, Vieweg.

**Hörz, Herbert et al.** (eds.) (1991). *Philosophie und Naturwissneschaft. Wörterbuch zu den philosophischen Fragen der Naturwissenschaften.* Neuausgabe 1991. 2 vols. Berlin, Dietz, (3rd ed.).

**Jary, D. & Jary J.** (eds.) (1991). *Collins Dictionary of Sociology.* Glasgow, Harper-Collins.

**Kaufmann, Arnold** (1975). *Introduction à la théorie des sous-ensembles flous. Vol. 1 Eléments théoriques de base. Vol. 2 Applications à la linguistique, à la logique et à la sémantique.* Paris, Masson.

**Khanna, Tarun** (1990). *Foundations of Neural Networks.* Reading, Mass., Addison Wesley,

**Leisi, Ernst** (1985[2]). *Praxis der englischen Semantik.* Heidelberg, Winter.

**Lorenzen, Paul** (1955[1], 1969[2]). *Einführung in die operative Logik und Methodik.* Berlin - Göttingen - Heidelberg.

**Lorenzen, Paul** (1962). *Metamathematik.* Mannheim, Bibliographisches Institut.

**Lorenzen, Paul** (1973). *Methodisches Denken.* Frankfurt/Main.

**MacCorquodale, K. & Mechl, P.E.** (1948). On a distinction between hypothetic constructs and intervening variables. *Psychological Review 55, 95 - 107.*

**Minsky, M.** (1975) A framework for representing knowledge. In: P. Winston (ed.): *The Psychology of Computer Vision.* New York, McGraw-Hill.

**Norman, D. A. & Rummelhart, D.E.** (1975). *Exploration in Cognition.* San Francisco, Freeman.

**Partee, Barbara H.** (1990). *Mathematical Methods in Linguistics.* Dordrecht, Kluwer.

**Rich, Elaine & Knight, Kevin** (1991). *Artificial Intelligence.* Auckland etc., McGraw-Hill (2nd ed.).

**Richter, Michael M.** (1989). *Prinzipien der Künstlichen Intelligenz.* Stuttgart, Teubner.

**Sambor, J. & Hammerl, R.** (eds.) (1991). *Definitionsfolgen und Lexemnetze Vol. I.* Lüdenscheid, RAM.

**Schank, R. C.** (1973). Identification of conceptualizations underlying natural language. In: R. C. Schank & K. M. Colby (eds.), *Computer Models of Thought and Language.* San Francisco, Freeman.

**Schank, R. C.** (1975). *Conceptual Information Processing.* Amsterdam, North Holland.

**Schwarz, Monika** (1992). *Einführung in die Kognitive Linguistik.* UTB 1636, Tübingen, Francke.

**Smadja, F.** (1989). Macrocoding the Lexicon with Co-Ocurrence Knowledge. *Proceedings of the First International Lexical Acquisition Workshop*, Detroit.

**Smadja, F. & K. McKeown** (1990) Automatically Extracting and Representing Collocations for Language Generation. *Proceedings of the 28th Annual Meeting of the ACL*, Pittsburgh.

**Struß, Peter** (ed.) (1991). *Wissensrepräsentation.* München, Oldenbourg.

**Zadeh, Lotfi A.** (1965). Fuzzy sets. *Information and Control 8*, 338-353.

**Zadeh, L.A. et al.** (eds.) (1975). *Fuzzy Sets and Their Applications to Cognitive and Decision Processes.* New York, Academic Press.

**Zadeh, Lotfi A.** (1987). *Fuzzy Sets and Applications. Selected Papers.* R. R. Yager et al. (eds.). New York, Wiley.

**Zimmermann, H.-J.** (1985). *Fuzzy Set Theory and its Applications.* Dordrecht, Kluwer-Nijhoff.

**Zimmermann, H.-J.** (1987). *Fuzzy Sets, Decision Making and Expert Systems.* Dordrecht, Kluwer-Nijhoff.

## Notes

In fuzzy set theory, a pioneer paper was Zadeh (1965). Kaufmann (1975), Bandemer (1989), Zimmermann (1985) and (1987) give good surveys. Special applications are found in Zadeh (1975) and Zadeh (1987). Goghen (1967) treats some generalization of fuzzy sets.

The constructivist view in logic was developed in Lorenzen (1955, 1962, 1973). Frege (1879) is the great classic, while Ganter (1987) contain modern particulars on concept analysis.

Many hints about epistemology can be found in Hörz (1991).

Among the sundry books on artificial intelligence, Goodman & Hedetniemi (1977), Harmon, Maus & Morissey (1988) Hofbauer & Kutsche (1989), Rich & Knight (1991) and Richter (1989) are to be recommended. Schwarz (1992) deals with a cognitive science approach in linguistics.

Mathematical and logical materials can be found in Bocheński & Menne (1983[5]), Dougherty & Giardina (1988), Khanna (1990), Partee (1990). Leisi 1985 treats English semantics from a more practical point of view.

In Gordesch (1978) to Gordesch & Gärtner (1992), various aspects of knowledge engineering and its epistemological background are dealt with. Pioneer articles concerning knowledge representation are by Abelson (1981), MacCorquodale & Mechl (1948), Minsky (1975), Norman & Rummelhart (1975), Schank (1973, 1975), while good surveys can be found in Rich & Knight (1975) and Struß (1991). The topic of semantic networks in linguistics is presented in Sambor & Hammerl (1991).

# Dynamic Aspects of Text Characteristics

*Reinhard Köhler & Matthias Galle, Trier*

1. The use of indices of one, two, or more variables, such as the verb-adjective-ratio (VAR) and the type-token-ratio (TTR) and many others, as measures for a given property is fundamental to a wide range of quantitative text analyses (cf. bibliography in Altmann 1988). When designed in a mathematically correct and linguistically appropriate way (Altmann 1988:18ff.) they can serve various theoretical and practical purposes, e.g. description, comparison, and classification of texts with respect to style, genre, language, etc., authorship determination, and many more.

All of these indices, however, have in common that they represent one of the features of a whole text by a single number - regardless of its length and its dynamic properties. As far as we can say, texts are neither homogeneous nor regular with respect to their (qualitative and quantitative) properties. Therefore, in many cases other measuring methods are preferred, e.g. entropy and repeat rate, which can express the degree of inhomogeneity but are single numbers too. Whenever possible, frequency or rank distributions should be considered, since they enable us to apply all the powerful tools of analytical statistics, in particular of test theory, and stimulate us to set up hypotheses.

Several attempts to investigate the dynamics of a feature in the course of a text have been made, in particular using a TTR measure repeatedly either for each word of the text or for every n-th word (Fig. 1 shows an example).

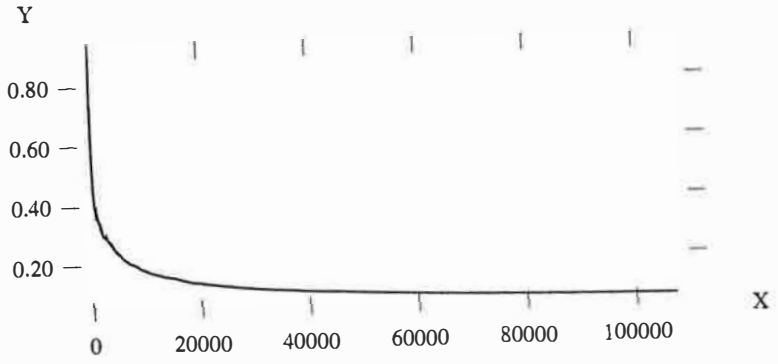



Figure 1. The TTR dynamics of "Das Schloß" (Franz Kafka)



A problem of this type of curve (Fig. 2 shows a similar example of a VAR curve) is that the mapping of the dynamic behaviour, i.e. of the increase or decrease of the number of elements of the category under investigation - e.g. new types -, is not linear.

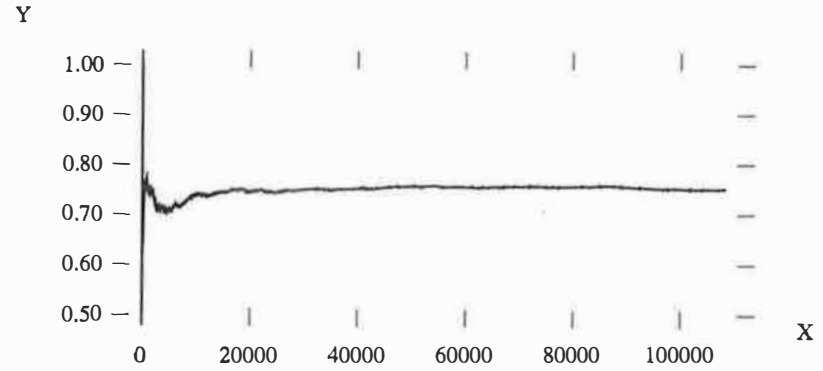



Figure 2. The verb-adjective-ratio in "Das Schloß" (Franz Kafka)

I.e., the effect of a given number of elements, say verbs, on the amplitude of the curve depends on the position in the text. An artificial "text" may illustrate this. As an example, for any index the VAR of a sequence of word blocks is displayed in Fig. 3, where 100 adjectives alternate with 100 verbs. The total length is 2000. As can be seen, the graph behaves like a damped oscillation, as a consequence of the fact that a given number of words makes a decreasing contribution to the VAR while n (the text length at the given position) increases.

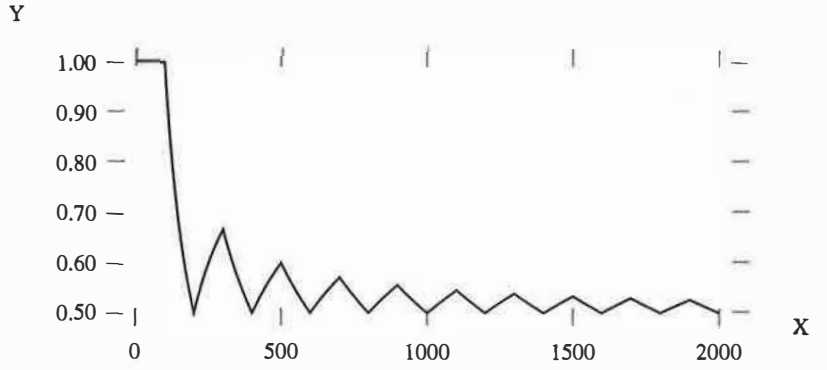



Figure 3. The VAR of an artificial "text"

A possible solution to this problem is the 'window' method, which is based on separate computation of the ratio for adjacent text sections (cf. Schach 1987, where this method has been applied to the TTR). The resulting series of TTR curves gives an improved impression of the dynamics of a text if the section size is chosen appropriately (cf. Fig. 4).

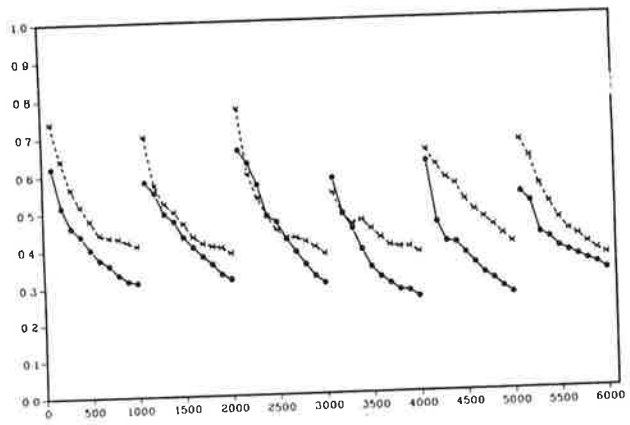



Figure 4. TTR of text sections (from Schach 1987)

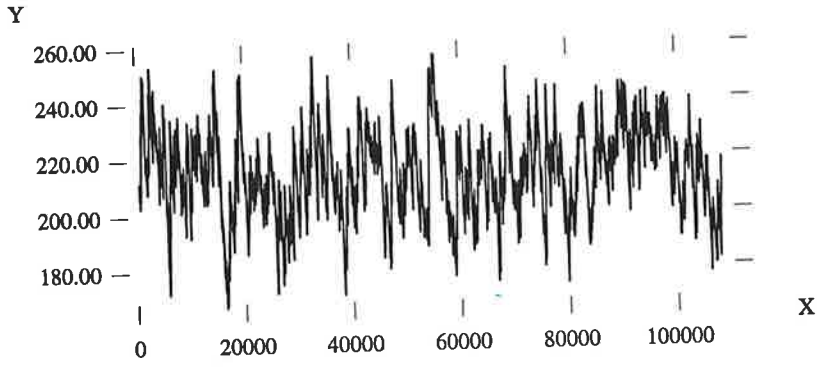



Figure 5. TTR computation based on gliding windows (of length 500)



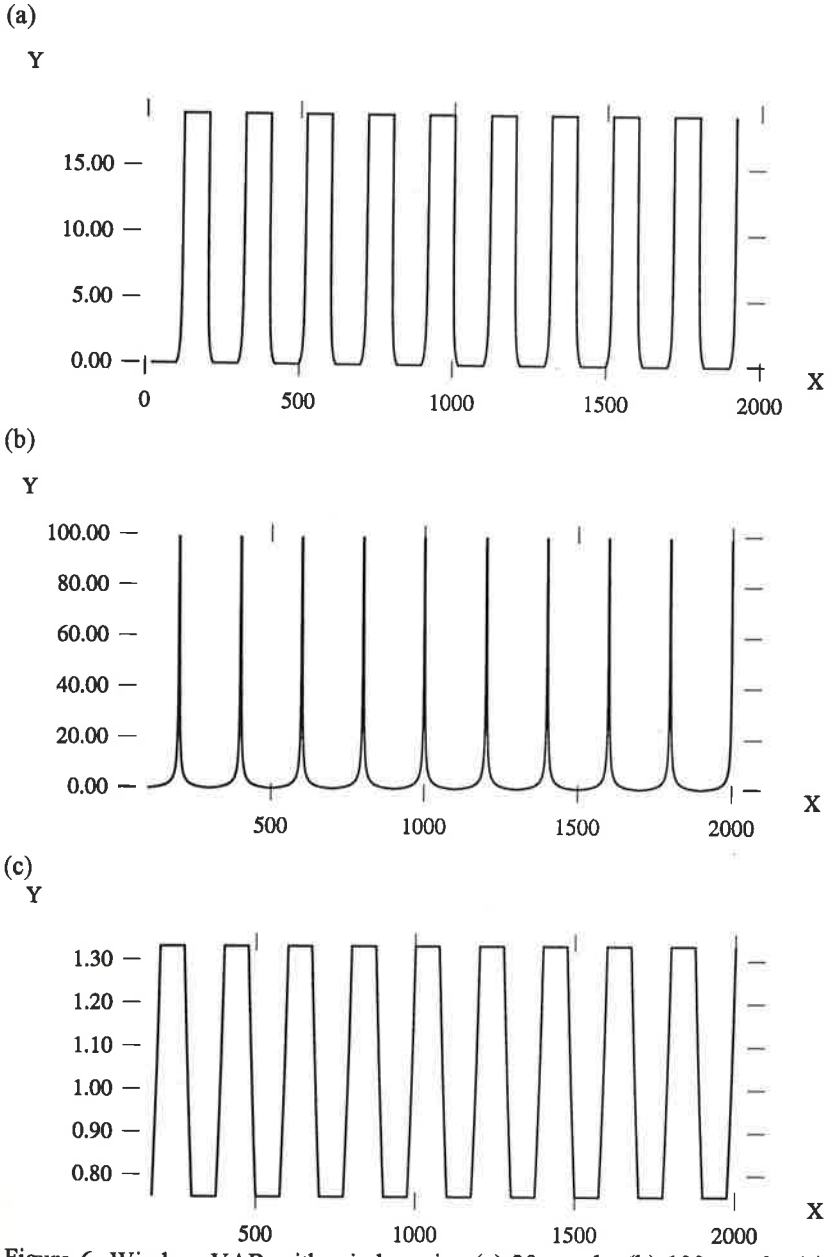



Figure 6. Window VAR with window size (a) 20 words, (b) 100 words, (c) 175 words

Even better mappings of the text dynamics can be achieved by gliding windows, in analogy to the procedures used for the Fast Fourier Transform and others (cf. Fig. 5). In this case, a value (not a curve) is computed for text sections of constant length, but the 'window' moves not more than one word to the right each time. The result is a curve of length N - w + 1 (N = text length, w = window size).

There are, however, some disadvantages of this method, among which are the dependency of the resulting curve on the window size and the effect of the window itself (analogous to the well-known problems in connection with signal processing). Fig. 6 shows VAR curves which correspond to three different window sizes applied to our artificial material.

2. Any dynamic property of a text can be investigated with respect to the laws which control the given property or with respect to the individual characteristics of a particular text. This is possible only if a theoretical model is available from which the desired law can be derived. In the case of the TTR a corresponding hypothesis follows from Altmann's proposal (1988: 88). He presents a differential equation whose solution $T = L^a$ (where T = number of types, L = number of tokens, $0 < a \leq 1$) can be used for prediction. Fig. 7 shows the number of tokens (= text position) of chapter 2 of Kafka's "Amerika" and the corresponding theoretical curve:

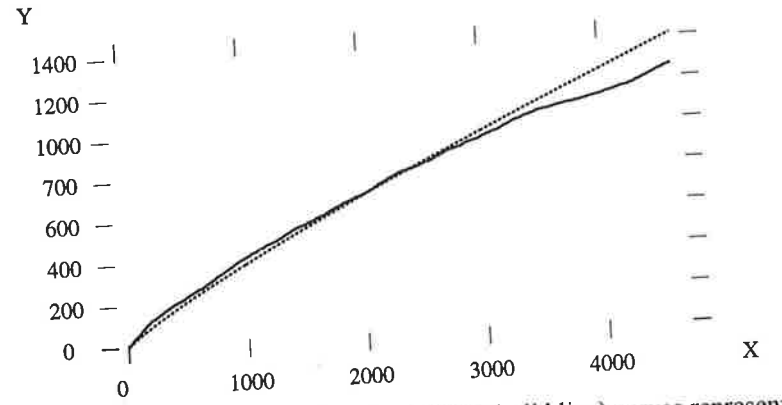



Figure 7. Theoretical (dotted line) and empirical (solid line) curves representing the number of types in the second chapter of Kafka's "Amerika".

If the model is valid,[1] then the differences between the theoretical and the empirical values can be considered as the individual characteristics of the given text (cf. Fig. 8).

[1] It is not the purpose of this paper to test this hypothesis.




Figure 8. Difference between the theoretical and the empirical curves in Fig. 7

3. In most cases, however, theoretical models are not yet available. One simply has to use empirical indices in order to study a property of a text. A simple index which avoids the above-mentioned shortcomings can be obtained by introducing a correction both in the denominator and the numerator of the formula under consideration, in order to keep the proportion on the same scale for the whole text. We choose once more the verb-adjective-ratio as an example, i.e. the ratio $v_n/a_n$, where $v_n$ is the number of verbs from the beginning of the text until the current text position $n$ and $a_n$ the corresponding number of adjectives:

$$VAR_n = \frac{v_N}{a_n}.$$

Modification of the ratio yields

$$VAR_n^* = \frac{v_n + V - n\frac{V}{N}}{a_n + A - n\frac{A}{N}} \qquad A \neq 0$$

where $A$ and $V$ are the overall sums of adjectives and verbs, respectively, $N = A + V$, and $n = a + v$. Fig. 9 presents the application of this index to our artificial material while Fig. 8 shows the result for "Das Schloß" (Franz Kafka).

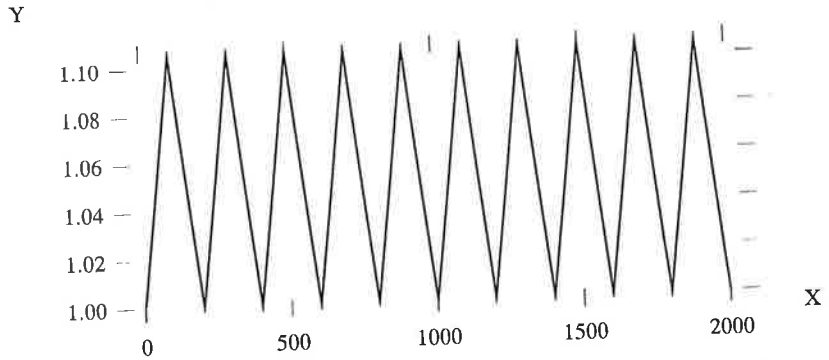



Figure 9. The modified index $VAR_n^*$ for the artificial material

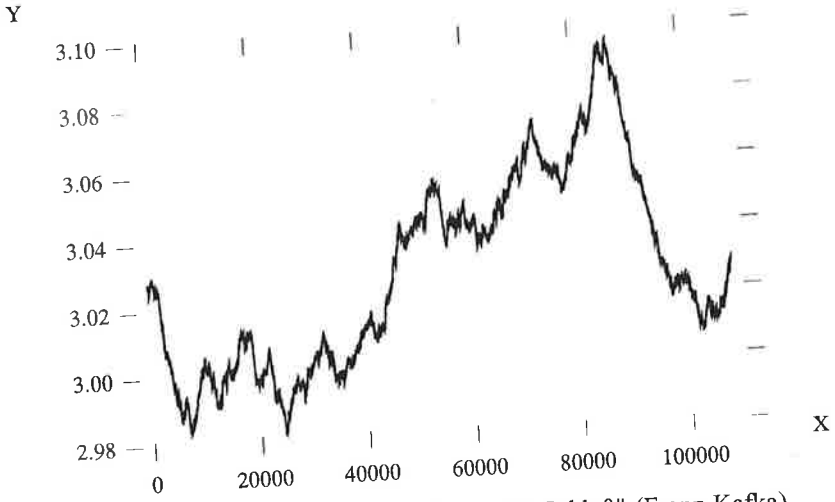



Figure 10. The graph of $VAR_n^*$ for "Das Schloß" (Franz Kafka)

Of course, norming of the index is recommended (cf. Altmann 1988: 20). For the TTR, the same principle yields, after simplification:

$$TTR_n^* = \frac{t_n + T - n\frac{T}{N}}{N}$$

where N is the text length (= number of tokens), T the number of types in the



text, $t_n$ the number of types up to the current position n. In Fig. 11 an example for this index is provided. For computational purposes, the more convenient linear function of $TTR_n^*$ can be used,[2] namely $t_n$ - nT/N.

In the same way, corrected indices can be obtained from any index by replacing each position-dependent variable $x_n$ by a term of the form $x_n$ + X - nX/N. In any case, the transformation to the (0,1) interval should be considered (cf. Altmann l.c.).

4. Since the aim of our contribution is to propose the described modification of the text indices in general, our computations are just meant as illustrative examples: In order to be useful for practical purposes, their statistical properties have to be investigated - which will be done in a future study.

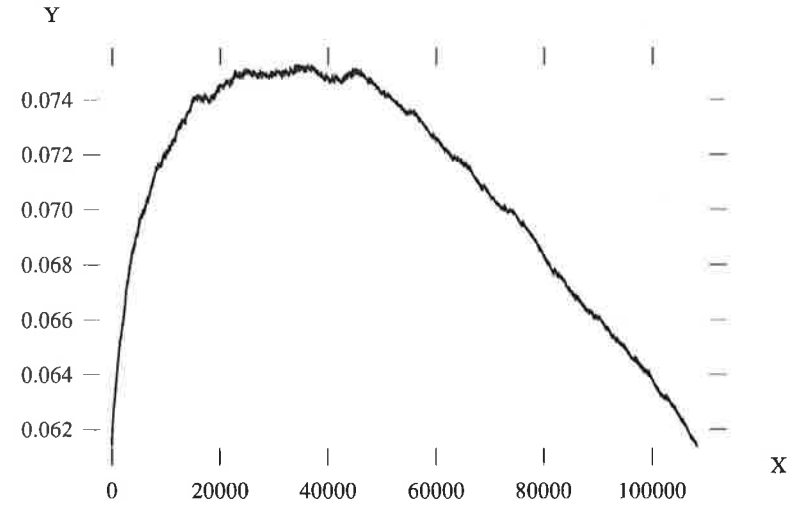



Figure 11. The graph of $TTR_n^*$ for "Das Schloß" (Franz Kafka)

## References

**Altmann, G.** (1988). *Wiederholungen in Texten.* ( = Quantitative Linguistics 36) Bochum: Studienverlag Brockmeyer.

**Schach, E.** (1987). Empirische Eigenschaften der TTR bei ausgewählten Texten. In: Wagner, Klaus R. (Hrsg.) *Wortschatz-Erwerb* (= Arbeiten zur Sprachanalyse Bd. 6). Bern-Frankfurt-New York-Paris: Lang: 102 - 114.

[2] A curve which represents the expected values for this index can be obtained from Altmann's formula: $T_n/N = n^{a-1}$

# Word Frequency Distributions

*Revas J. Chitashvili, Tbilisi*
*R. Harald Baayen, Nijmegen*

## 1. Introduction

Word frequency distributions have been studied from a variety of perspectives. In literary studies, word frequency distributions have attracted the attention of scholars interested in authorship attribution and vocabulary richness (Orlov 1983b, Muller 1977, 1979, Menard 1983, Thisted and Efron 1987, Herdan 1960, 1964). Psychologists have long been interested in word frequencies since word frequency is one of the most robust and important predictors of response time in a variety of experimental tasks addressing on-line word production and word recognition (Carroll 1969, 1970, Scarborough et al. 1977, Whaley 1978). Recently, word frequency distributions have also been exploited for the study of morphological productivity, the extent to which various word formation processes are alive in the language and may be expected to give rise to new (morphologically complex) formations (Baayen 1992, 1993a). This paper focusses on the probabilistic properties of word frequency distributions and on the statistical techniques developed for their analysis. Some attempt will be made, however, to understand the typical statistical properties of word frequency distributions of running texts in terms of the morphological structure of the constituent words and the productivity of the underlying word formation processes.

Our discussion is structured as follows. Section 2 introduces various ways of describing empirical word frequency distributions as well as a number of 'laws' that have been advanced in the literature as governing these distributions. Section 3 develops a stochastic approach to word frequency distributions. The multinomial and Poisson models are introduced as means for obtaining theoretical expressions for the expected vocabulary and the frequency spectrum as functions of the sample (text) size. The important concept of the Large Number of Rare Events Zone (LNRE ZONE) is introduced. It is shown that many empirical samples are located in this zone where relative sample frequencies are biased estimates of population probabilities. The consequences for the construction of theoretical models for word frequency distributions are considered in detail. Three such models for LNRE distributions are discussed, Carroll's (1967, 1969) lognormal 'law', Sichel's (1975, 1986) generalized inverse Gauss-Poisson 'law', and Orlov and Chitashvili's (1982a, 1982b, 1983a, 1983b) extended generalized Zipf's 'law'. In section 4 rationales for the lognormal 'law' and various extensions



of Zipf's 'law' are discussed. Section 5 outlines how statistical analyses with LNRE models may be carried out. Finally, section 6 discusses the relation between morphological productivity and the LNRE property of running texts.

Since our aim is to give a bird's eye view of the main results obtained in the study of word frequency distributions, mathematical proofs have been ommitted, many of the results reviewed here being common knowledge ever since Yule's (1944) seminal study and the important papers by Good (1953), Good and Toulmin (1956) and Kalinin (1965). For an in-depth mathematical discussion of the to our mind central notion of LNRE distributions the reader is referred to Khmaladze and Chitashvili (1989), part of which has appeared in English as Khmaladze (1987).

## 2. LNRE Features of Word Frequency Distributions

In this section we introduce some general properties of word frequency distributions. We first present some techniques for describing the frequency spectrum, and then turn to review some of the 'laws' supposedly governing word frequency distributions suggested in the literature.

### 2.1. The Frequency Spectrum

We can view a running text as an ordered sequence of word tokens

$$(w_1, w_2, \ldots, w_N).$$

Usually the observed (or empirical) vocabulary $\underline{\hat{V}}$,

$$\underline{\hat{V}} = (A_1, A_2, \ldots, A_{\hat{V}}), \tag{1}$$

the (arbitrarily ordered) set of different words (or word types) used in the text, or, alternatively,

$$\underline{\hat{V}}_o = (A_{(1)}, A_{(2)}, \ldots, A_{(\hat{V})}), \tag{2}$$

the set of word types ordered according to their (token) frequencies,

$$f_N(A_{(1)}) \geq f_N(A_{(2)}) \geq \ldots \geq f_N(A_{(\hat{V})}), \tag{3}$$

contains a much smaller number $\hat{V}$ of elements then the sample size (or text

size) N. This makes it convenient to present a text in the form of an array $\underline{A}$

$$\underline{A} = \begin{pmatrix} A_{(1)}: & \tau_1(A_{(1)}) & \tau_2(A_{(1)}) & \dots & \dots & \tau_{f(A_{(1)})}(A_{(1)}) \\ A_{(2)}: & \tau_1(A_{(2)}) & \tau_2(A_{(2)}) & \dots & \tau_{f(A_{(2)})}(A_{(2)}) & \\ & \dots & \dots & \dots & & \\ A_{(i)}: & \tau_1(A_{(i)}) & \tau_2(A_{(i)}) & & & \\ A_{(i+1)}: & \tau_1(A_{(i+1)}) & \tau_2(A_{(i+1)}) & & & \\ & \dots & & & & \\ A_{(j)}: & \tau_1(A_{(j)}) & & & & \\ A_{(j+1)}: & \tau_1(A_{(j+1)}) & & & & \\ & \dots & & & & \\ A_{(\hat{V}-1)}: & \tau_1(A_{(\hat{V}-1)}) & & & & \\ A_{(\hat{V})}: & \tau_1(A_{(\hat{V})}) & & & & \end{pmatrix} \tag{4}$$

in which $\tau_1(A_{(i)})$, $\tau_2(A_{(i)})$,... indicate the positions of word $A_{(i)}$ in the text. For instance, $\tau_7(A_{(i)}) = 137$ denotes that word $A_{(i)}$ occurred for the seventh time on the 137th stage (trial)). Note that in (4) the highest frequency type is on the first line and that the so-called hapax legomena, the types occuring once only, occupy lines $j$ down to $\hat{V}$.

Corresponding to the sample frequencies we have the sample relative frequencies $\hat{p}(A_i)$ and $\hat{p}(A_{(i)})$ for the unordered and frequentially ordered vocabulary items respectively:

$$\hat{p}(A_i) = \frac{f_N(A_i)}{N} \quad (cf.\ (1)) \tag{5}$$

$$\hat{p}(A_{(i)}) = \frac{f_N(A_{(i)})}{N} \quad (cf.\ (2)) = \hat{p}_N\{i\}\ . \tag{6}$$

The information contained in $\underline{A}$ can be used for various purposes. For instance, the transition probabilities

$$\hat{p}_N(A_i|A_j) = \frac{1}{f_N(A_j)} \sum_{n=1}^{f_N(A_j)} \sum_{k=1}^{f_N(A_i)} \mathbf{I}_{[\tau_k(A_i) = \tau_n(A_j)+1]} \tag{7}$$



can be used to study dependencies between words as they occur in some text. In this paper we focus on the analysis of the frequency distribution, i.e. the set

$$(f_N(A_1), f_N(A_2),..., f_N(A_{\hat{V}}))$$

of lengths of rows in array $\underline{A}$. To restrict attention to the frequency distribution is to use the information which is invariant with respect to permutations of elements both in the sample and in the vocabulary.

The characteristic feature of the samples (texts, morphological categories) we are to investigate in this paper is that besides the elements with high (token) frequencies (e.g. $\hat{p}_N(A_{(i)}) \approx 0.05$), in the above array the upper rows of substantial length, we observe many elements that occur only once, twice, etc. Crucially, these events constitute a significant part of the vocabulary. Often the number of elements occurring only once approximates half the observed vocabulary size. We will refer to distributions with this characteristic as Large Number of Rare Events (LNRE) distributions.

The frequency distribution can be presented in at least four equivalent forms:

**1. The frequency spectrum:**

Let $\hat{V}_N(m)$ denote the number of elements of the vocabulary which occurred $m$ times in a sample of size $N$:

$$\hat{V}_N(m) = \sum_{i\geq 1} \mathbf{I}_{[f(A_i) = m]}\ , \quad m = 1, 2, \dots, \tag{8}$$

where $\mathbf{I}_{[f(A_i) = m]}$ is the indicator of the event $[f(A_i) = m]$, i.e.

$$\mathbf{I}_{[f(A_i) = m]} = \begin{cases} 1 & \textbf{\textit{if}}\ f(A_i) = m \\ 0 & \textbf{\textit{otherwise}}. \end{cases}$$

It is easy to observe that the (empirical) vocabulary size for sample size $N$ is given by

$$\hat{V}_N = \sum_{m\geq 1} \hat{V}_N(m). \tag{9}$$

We shall often make use of the relative frequency spectrum

$$\hat{\alpha}_N(m) = \frac{\hat{V}_N(m)}{\hat{V}_N}, \quad m = 1,2,\ldots \quad . \tag{10}$$

Figure 1 illustrates these functions for the English suffix *-ness* as it appears in the Cobuild corpus (Sinclair 1987). (Here, and in all examples to follow, the frequency count is lemma-based, inflectional variants of a stem being counted as tokens of one and the same lemma type.) Note that the number of hapaxes $\hat{V}_N(m)$ is approximately $\hat{V}_N/2$.

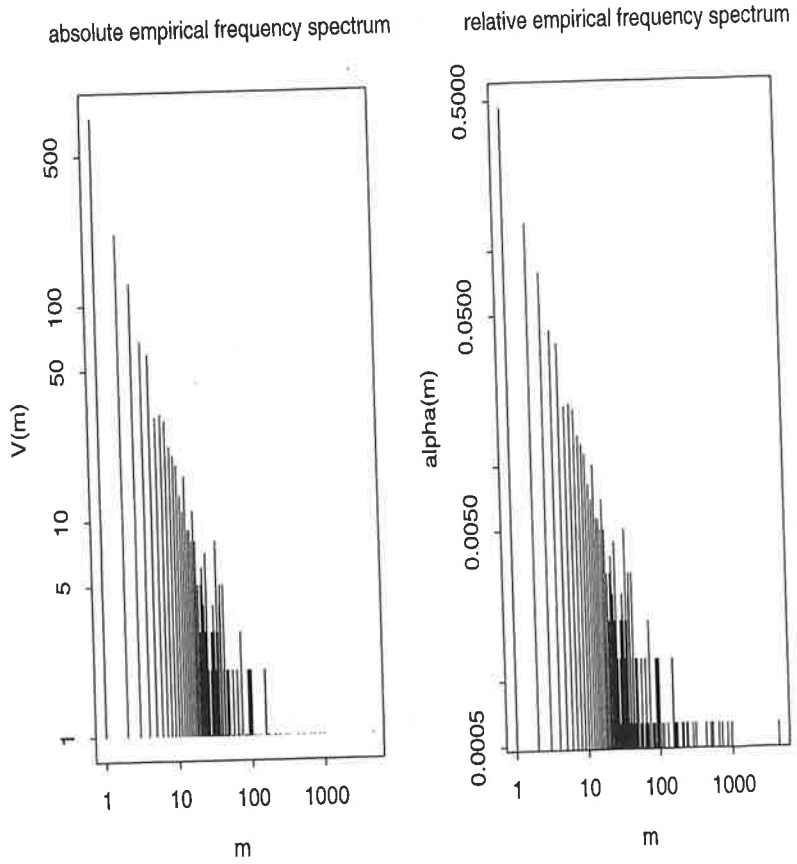



Figure 1. Absolute and relative frequency spectrum for the English suffix *-ness* as it appears in the Cobuild corpus, plotted on a double logarithmic scale.



**2. The rank frequency distribution:**

Given that the elements of the vocabulary

$$\hat{V} = (A_1, A_2, \ldots, A_{\hat{V}})$$

are ordered according to decreasing frequency as specified in (3), we can denote any word by its rank $r$ (its row number in $\underline{A}$) in the resulting list:

$$f_N\{r\} = f_N(A_{(r)}), \quad r = 1,2,\ldots \quad . \tag{11}$$

This way of representing word frequency distributions is well known from the early studies by Zipf (1935) onwards.

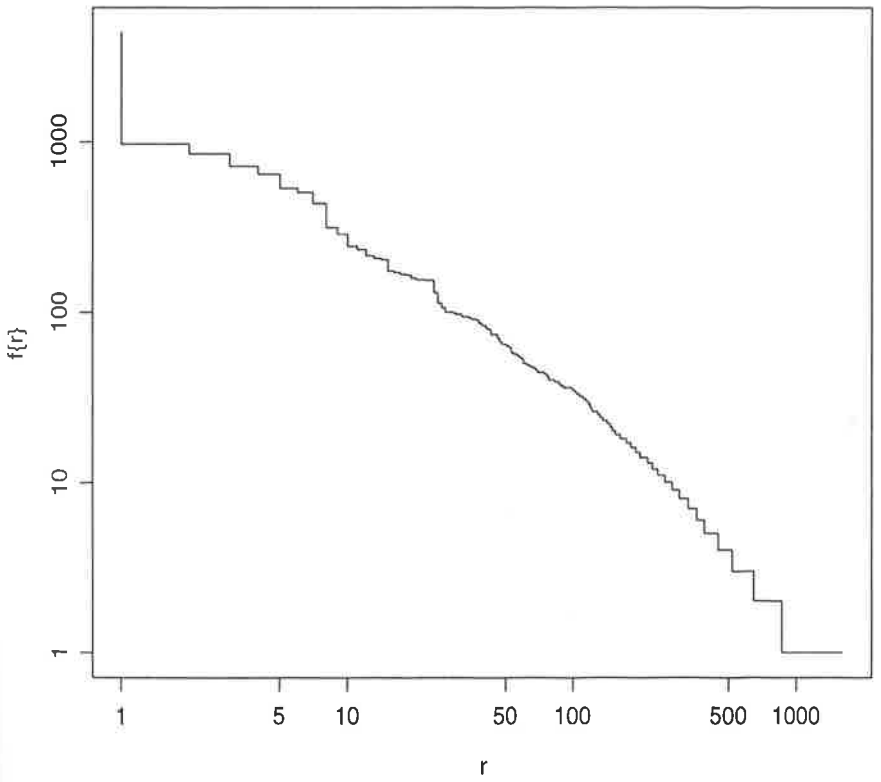



Figure 2. Rank-frequency plot for the English suffix *-ness* as it appears in the Cobuild corpus. The X-axis and the Y-axis are scaled logarithmically.

Sometimes it is more natural to consider the relative rank-frequency distribution

$$\hat{p}_N\{r\} = \frac{f(A_{(r)})}{N} .$$

Thus

$$(\hat{p}_N\{r\},\ 1 \le r \le \hat{V}_N)$$

is the ordered set of relative frequencies

$$(\hat{p}_N(A_i),\ \ 1 \le i \le \hat{V}_N).$$

Note that, as shown in figure 2, it is often convenient (and more demonstrative) to present graphs of the rank frequency distribution (or of the structural distributions to be discussed below) in a double logarithmic scale, that is, to consider the transformed step function

$$\log_a \hat{p}_N\{[a^x]\},\ \ x \ge 0$$

of a variable $x = \log r$, where we use the notation $[a^x]$ to denote the integer part of $a^x$. Usually, $e$ or 10 are chosen for the logarithmic base $a$.

### 3. The empirical structural type distribution

The cumulative type frequency or empirical structural type distribution is defined in terms of the type probability $p$ in the sample:

$$\hat{G}_N(p) = \sum_{i\ge1} \mathbf{I}_{[f_N(A_i) \ge Np]} = \sum_{i\ge1} \mathbf{I}_{[\hat{p}_N(A_i) \ge p]}. \qquad (12)$$

In (12), $\hat{G}_N(p)$ denotes the number of elements of the vocabulary which occurred at least $Np$ times in the sample (text).



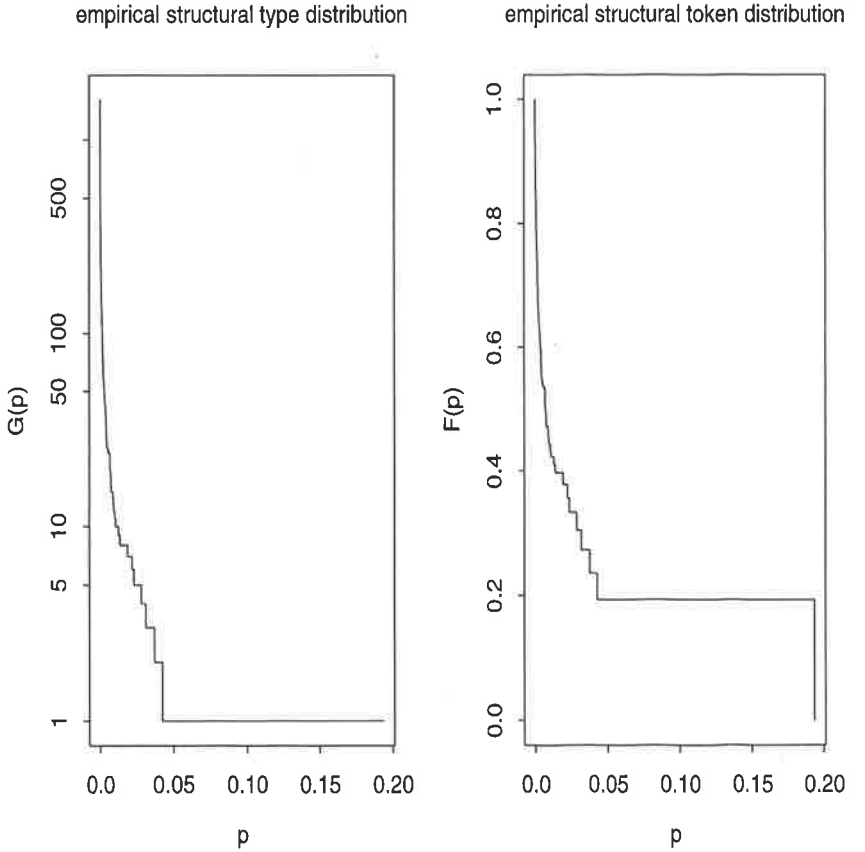



Figure 3. The empirical structural type and token distributions for the English suffix *-ness*.

### 4. The empirical structural token distribution

The cumulative token frequency or empirical structural token distribution is defined as

$$\hat{F}_N(p) = \sum_{i\ge1} \hat{p}_N(A_i)\mathbf{I}_{[\hat{p}_N(A_i) \ge p]}. \qquad (13)$$

So $\hat{F}_N(p)$ is the relative frequency of those tokens in the sample which are instances of types with a relative frequency not less then $p$. Sometimes we will re-

fer to both $\hat{G}$ and $\hat{F}$ as empirical structural distributions. Figure 3 plots these functions for the suffix *-ness*. Note how the presence of a single very high frequency type effects a sizeable difference in the shape of the two graphs.

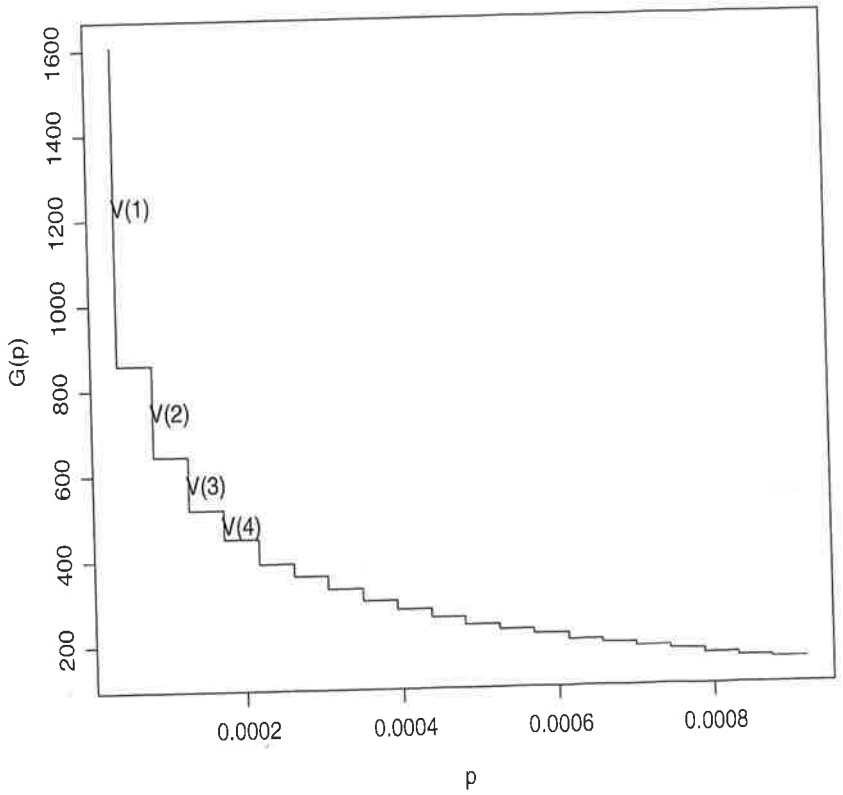



Figure 4. The relation between the frequency spectrum and the empirical structural token distribution. $\hat{G}_N(p)$ is shown for the first 23 distinct values of $p$ for the suffix *-ness*. The corresponding spectrum elements have been added for $m = 1, \ldots, 4$.

These four ways of representing the frequency distribution are fully equivalent. This becomes apparent when we make explicit the relations that hold between them:

**(a)** The terms of the frequency spectrum $(\hat{V}_N(m))_{m = 1,2,\ldots}$ can be expressed in terms of the empirical structural type distribution,



$$\hat{V}_N(m) = \hat{G}_N(\frac{m}{N}) - \hat{G}_N(\frac{m+1}{N}), \quad m = 1,2,\ldots$$

as shown in figure 4. Equivalently, we have

$$\hat{G}_N(\frac{m}{N}) = \sum_{k \geq m} \hat{V}_N(k)$$

**(b)** The empirical structural type and token distributions are related by the equality

$$\hat{G}_N(p) = \sum_{q \geq p} \frac{1}{q} \Delta \hat{F}_N(q),$$

where $\Delta \hat{F}_N(q)$ is a finite difference (i.e. the jump value) of the step function $\hat{F}_N(p)$ at a point $q$. Or, equivalently,

$$\hat{F}_N(p) = \frac{1}{N} \sum_{m \geq Np} m \hat{V}_N(m).$$

**(c)** The structural distribution $\hat{G}_N(p)$ is an inverse function to the rank-frequency distribution $f_N\{r\}$, i.e.

$$\hat{G}_N\left(\frac{f_N\{r\}}{N}\right) = \hat{G}_N(\hat{p}_n\{r\}) = r, \quad r = 1,2,\ldots . \tag{14}$$

Some probabilistic meaning can be given to the relative frequency spectrum $\hat{\alpha}_N(m)$. If the empirical vocabulary of distinct word types is conceived of as constituting the experimental population from which we are sampling a type at random, then is the probability that some word type having token frequency $m$ will be chosen, or, equivalently, $\hat{G}_N(p)/\hat{V}_N$ is the probability that some word type having a relative frequency of at least $p$ will be chosen. A stochastic interpretation of the token probability distribution $\hat{F}_N(p)$ is given in section 3.1.3.

## 2.2. Laws Proposed for Frequency Spectra

A number of simple analytical expressions have been suggested in the literature

for 'theoretical laws', either for the relative frequency spectrum or for rank-frequency distributions. In terms of the relative spectrum these 'laws' can be presented as follows:

1. Zipf (Zipf 1935)

$$\hat{\alpha}_N(m) \approx \alpha(m) = \frac{1}{m(m+1)}, \tag{15}$$

2. Yule (Yule 1924; Simon 1955)

$$\hat{\alpha}_N(m) \approx \alpha(m) = \frac{\Gamma(\beta+1)\Gamma(m)\beta}{\Gamma(m+\beta+1)}, \quad (\beta > 0), \tag{16}$$

3. Yule-Simon (Simon 1956, 1960)

$$\hat{\alpha}_N(m) \approx \alpha(m) = \frac{\beta}{(m+\beta-1)(m+\beta)}, \quad (\beta > 0), \tag{17}$$

4. Waring-Herdan-Muller (Herdan 1960, 1964; Muller 1979)

$$\hat{\alpha}_N(m) \approx \alpha(m) = \frac{\Gamma(\beta+1)\alpha}{\Gamma(\beta+1-\alpha)} \cdot \frac{\Gamma(m+\beta-\alpha)}{\Gamma(m+\beta+1)}, \quad (0<\alpha<1,\ \beta>\alpha), \tag{18}$$

5. Karlin-Rouault (Rouault 1978)

$$\hat{\alpha}_N(m) \approx \alpha(m) = \frac{\alpha\Gamma(m-\alpha)}{\Gamma(1-\alpha)\Gamma(m+1)}, \quad (0 < \alpha < 1), \tag{19}$$

6. Zipf-Mandelbrot (Mandelbrot 1962)

$$\hat{\alpha}_N(m) \approx \alpha(m) = \frac{1}{m^\gamma} - \frac{1}{(m+1)^\gamma}, \quad (\gamma > 0). \tag{20}$$

We will refer to these 'laws' as the Zipfian family of models.

Graphs of these 'laws' for varying parameter values are shown in figure 5 and figure 6. Note that in the case of the Waring-Herdan 'law' increasing the value of $\alpha$ leads to an increase in type richness, as evidenced by the values of $\alpha_N(1)$ and the ratio $r_V = \beta/(\beta - \alpha)$ by which $\hat{V}_N$ has to be multiplied to obtain the theoretical vocabulary $V$. Decreasing $\beta$ similarly leads to higher values of $V$. Also observe that especially high values of $V$ are obtained when the difference



between $\beta$ and $\alpha$ is small. Finally, note that for the Karlin-Rouault 'law', the parameter $\alpha$ equals $\alpha_N(1)$.

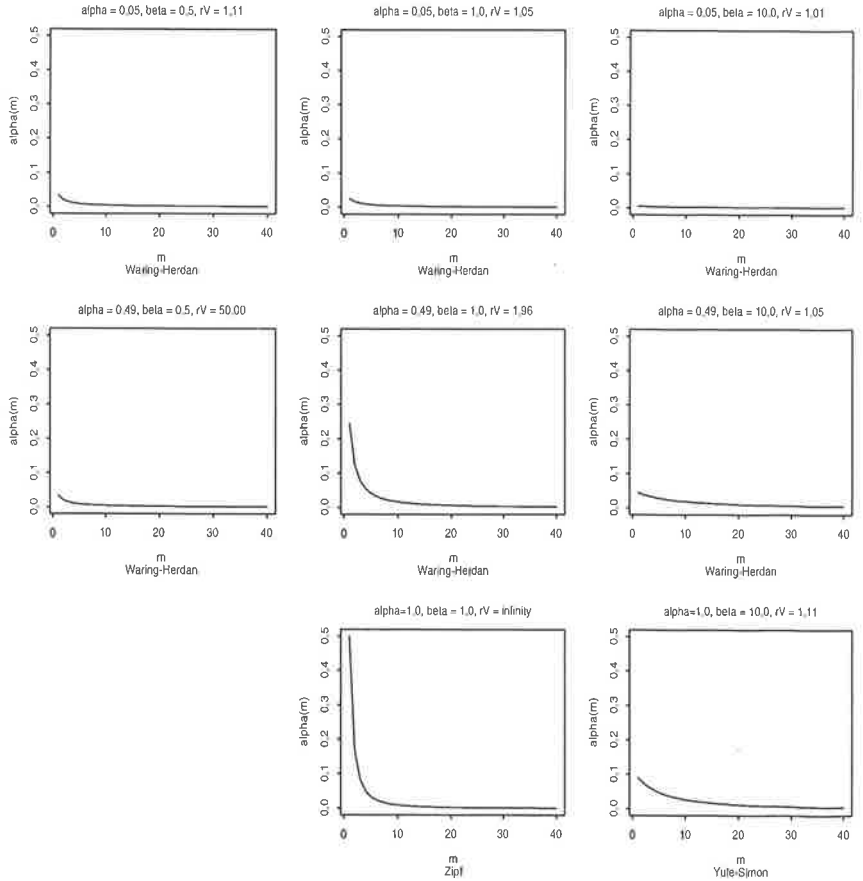



Figure 5. The Zipfian family of 'laws' for selected parameter values: Waring-Herdan, Yule-Simon and Zipf.

The corresponding 'laws' for the rank-frequency distribution may be obtained using the relation (14) between the cumulative structural distribution $G$ and the rank-frequency distribution $p\{r\}$. In fact, in the case of Zipf 's 'law', for instance, the corresponding model for the structural distribution $\hat{G}(p)$ should be any function $\hat{G}(r)$ with the property

$$\frac{\hat{G}(\frac{m}{N}) - \hat{G}(\frac{m+1}{N})}{\hat{G}(\frac{1}{N})} = \frac{1}{m(m+1)}.$$

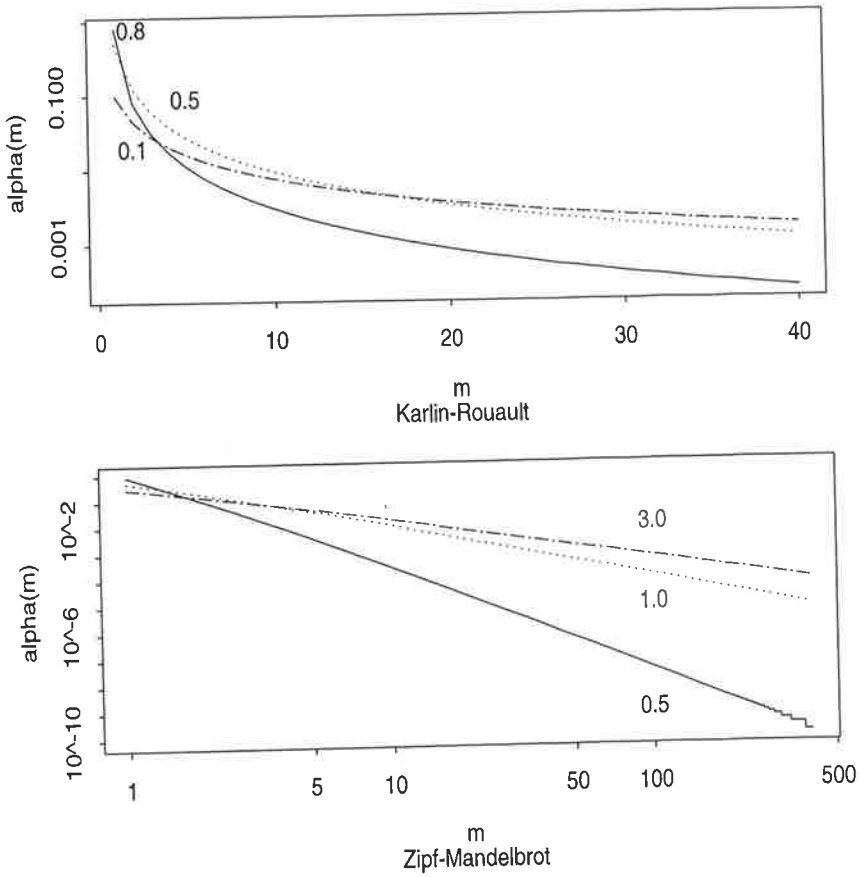



Figure 6. The Zipfian family of 'laws' for selected parameter values: Karlin-Rouault and Zipf-Mandelbrot.

The solution is simply

$$\hat{G}(\frac{m}{N}) = \frac{NC}{m} + B, \tag{21}$$

with some parameters $(C,B)$. Again using relation (14) we find that the corresponding rank-frequency distribution should have the form

$$\hat{p}_N\{r\} \approx p\{r\} = \frac{C}{r - B}$$

as a solution for the equation

$$\frac{C}{\hat{p}_N\{r\}} + B = r.$$

In the case of the Zipf-Mandelbrot 'law' we similarly have

$$\overline{G}\left(\frac{m}{N}\right) = \frac{C}{\left(\frac{m}{N}\right)^{\gamma}} + B \tag{22}$$

and

$$\hat{p}_N\{r\} \approx p\{r\} = \left(\frac{C}{r - B}\right)^{1/\gamma} \tag{23}$$

for some parameters $C$ and $B$. Graphs of these distributions for varying choices of the parameters are shown in figure 7. Note that small values of $\gamma$ effect a downward curvature for the lower ranks $r$ without influencing the shape of the curve for the higher ranks. We will return to the independence of the head and tail of the frequency distribution in section 3.3.3.

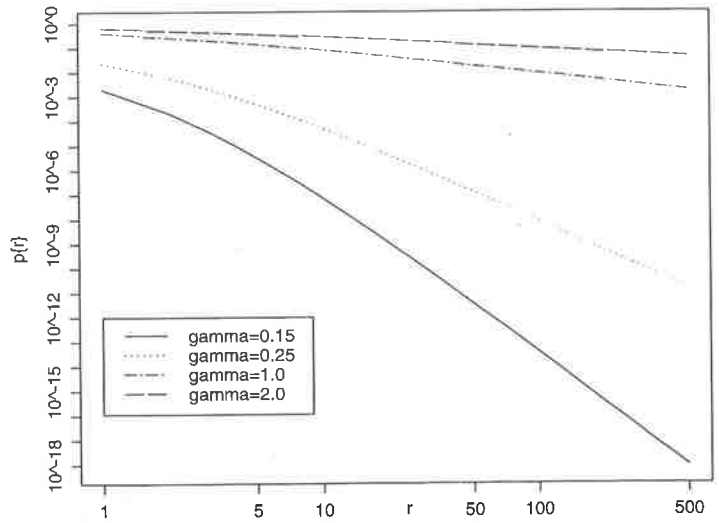



Figure 7. Rank-frequency distributions: Zipf and Zipf-Mandelbrot distributions for $B$ = -1.5, $C^{1/\gamma}$ = 0.9 and varying $\gamma$.

# 3. Stochastic Modelling of LNRE

In this section we first introduce expressions for the structural distributions using the multinomial and Poisson models. We then define the concepts of the LNRE ZONE and the generalized structural distribution. Finally, we consider the rationale for generalized structural distributions using an asymptotic approach.

## 3.1. The Structural Distribution in the Classical Scheme

### 3.1.1. The Multinomial Model

Even though the four forms in which we may represent the frequency distribution are fully equivalent in that they summarize exactly the same information, we will focus on the structural distributions since it is the structural distributions which contain explicit expressions for the relevant probabilistic characteristics.

Assuming the classical scheme of independent identically distributed trials, let

$$(\mathrm{P}(A_i),\ 1 \le i \le V)$$

be the probability distribution over the set

$$\underline{V} = (A_1, A_2, \ldots, A_V)$$

of elements of the theoretical vocabulary.

As direct analogues for the empirical structural distributions we consider the following expressions:

$$G(p) = \sum_{i=1}^{V} \mathbf{I}_{[p(A_i) \ge p]},\quad p \ge 0, \tag{24}$$

$$F(p) = \sum_{i=1}^{V} p(A_i)\mathbf{I}_{[p(A_i) \ge p]},\quad p \ge 0\ . \tag{25}$$

The functions $G(p)$ and $F(p)$ can be interpreted in the same way as their empirical analogues $\hat{G}_N(p)$ and $\hat{F}_N(p)$, be it that the general population of words is considered instead of the experimental sample population. We shall refer to these functions as the theoretical structural type and token distributions, or, alternatively, as the theoretical cumulative type and token probability distributions.



Let's now consider how the theoretical and empirical distributions are related. It is a well known fact that the vector of frequencies

$$(f_N(A_1),\ f_N(A_2),\ \ldots,\ f_N(A_V)) \tag{26}$$

is multinomially distributed:

$$\boldsymbol{Pr}(f_N(A_i) = n_i\ ,\quad 1 \le i \le V) = \binom{N}{n_1,\ n_2,\ \ldots,\ n_V} \prod_{i=1}^{V} p(A_i). \tag{27}$$

For the important special case of binomial probabilities and the corresponding upper-cumulative probabilities we will use the notations

$$B(N,\ m,\ p) = \binom{N}{m} p^m (1\ -\ p)^{N-m} \tag{28}$$

$$B^{+}(N,\ m,\ p) = \sum_{k \ge m} \binom{N}{k} p^k (1\ -\ p)^{N-k}\ . \tag{29}$$

Similarly, trinomial probabilities will be referred to as

$$T(N,\ m,\ l,\ p,\ q) = \binom{N}{m,\ k} p^m q^k (1\ -\ p\ -\ q)^{N-m-k}\ .$$

For the expected values and covariances of the indicators

$$\mathbf{I}_{[f_N(A_i)\ =\ m]},\quad m \ge 1$$

we have

$$\boldsymbol{E}\ \mathbf{I}_{[f_N(A_i)\ =\ m]} = \binom{N}{m} p(A_i)^m (1 - p(A_i))^{N-m} = B(N,\ m,\ p(A_i)) \tag{30}$$

and

$$\begin{aligned} \boldsymbol{COV}(\ \mathbf{I}_{[f_N(A_i)\ =\ m]},\ \mathbf{I}_{[f_N(A_j)\ =\ k]}) \ &= \delta_{ij}\delta_{mk} B(n,\ m,\ p(A_j)) \\ &+ (1 - \delta_{ij}) T(N,\ m,\ k,\ p(A_i),\ p(A_j)) \\ &- B(N,\ m,\ p(A_i)) B(N,\ k,\ p(A_j)), \end{aligned} \tag{31}$$

where

$$\delta_{ij} = \begin{cases} 1 & \textit{if } i = j \\ 0 & \textit{otherwise.} \end{cases}$$

Similarly,

$$\boldsymbol{E}\mathbf{I}_{[V_N(A_i) \geq m]} = B^+(N, m, p(A_i))$$

and

$$\begin{aligned} \boldsymbol{COV}(\mathbf{I}_{[V_N(A_i) \geq m]}, \mathbf{I}_{[V_N(A_j) \geq k]}) &= \delta_{ij} B^+(N, \max(m, k), p(A_i)) \\ &+ (1 - \delta_{ij}) \sum_{l \geq m, r \geq k} T(N, l, r, p(A_i), p(A_j)) \\ &- B^+(N, m, p(A_i)) B^+(N, k, p(A_j)). \end{aligned} \tag{32}$$

Now the expected values for the frequency spectrum and the empirical distributions $\hat{G}_N(p)$ and $\hat{F}_N(p)$ can be expressed as

$$\begin{aligned} V_N(m) = \boldsymbol{E}\hat{V}_N(m) &= \boldsymbol{E}\sum_i \mathbf{I}_{[V_N(A_i) = m]} \\ &= \sum_i \binom{N}{m} p(A_i)^m (1 - p(A_i))^{N-m} \end{aligned} \tag{33}$$

$$\begin{aligned} V_N = \boldsymbol{E}\hat{V}_N &= \boldsymbol{E}\sum_{m \geq 1} \hat{V}_N(m) \\ &= \boldsymbol{E}\sum_m \sum_i \mathbf{I}_{[V_N(A_i) = m]} \\ &= \sum_{m \geq 1} \sum_i \binom{N}{m} p(A_i)^m (1 - p(A_i))^{N-m} \\ &= \sum_i (1 - (1 - p(A_i))^N) \end{aligned} \tag{34}$$

$$\begin{aligned} G_N(p) = \boldsymbol{E}\hat{G}_N(p) &= \sum_i \sum_{m \geq Np} \binom{N}{m} p(A_i)^m (1 - p(A_i))^{N-m} \\ &= \sum_i B^+(N, Np, p(A_i)) \end{aligned} \tag{35}$$



$$\begin{aligned} F_N(p) = \boldsymbol{E}\hat{F}_N(p) &= E\frac{1}{N}\sum_{m \geq Np} m\hat{V}_N(m) \\ &= \sum_i \sum_{m \geq Np} \frac{m}{N} \binom{N}{m} p(A_i)^m (1 - p(A_i))^{N-m} \\ &= \sum_i p(A_i) B^+(N - 1, Np - 1, p(A_i)) \end{aligned} \tag{36}$$

and

$$\begin{aligned} &\boldsymbol{COV}(\hat{V}_N(m), \hat{V}_N(k)) = \\ &= \sum_i \binom{N}{m} (p(A_i))^m (1 - p(A_i))^{N-m} \\ &+ \sum_{ij} \binom{N}{m, k} (p(A_i))^m (p(A_j))^k (1 - p(A_i) - p(A_j))^{N-m-k} \\ &- \sum_i \binom{N}{m, k} (p(A_i))^{m+k} (1 - 2p(A_i))^{N-m-k} \\ &- \sum_i \binom{N}{m} (p(A_i))^m (1 - p(A_i))^{N-m} \sum_i \binom{N}{k} (p(A_i))^k (1 - p(A_i))^{N-k}. \end{aligned} \tag{37}$$

Similar expressions can be obtained for the covariances of other characteristics of the frequency distribution ($\boldsymbol{COV}(\hat{F}_N(p), \hat{F}_N(p'))$ for instance) as linear combinations of the spectrum, but we omit them as we will be using simpler versions based on the Poisson model.

We shall now make a rather formal step to rewrite these expressions in integral form to show that (any) probabilistic characteristics of frequency distributions can be expressed in terms of the corresponding theoretical structural distributions. This will also allow us to express further generalizations in a natural way.

Since the theoretical structural type distribution $G(p)$ is a (nonincreasing) step function defined on the interval [0,1] with jumps at the points $(p_1, p_2, \ldots, p_V)$,

$$\Delta G(p_i) = G(p_i) - G(p_i+), \tag{38}$$

where $p_i+ = \lim_{p \downarrow p_i} G(p)$, and similarly for the structural token distribution,

$$\Delta F(p_i) = p_i(G(p_i) - G(p_i+)), \tag{39}$$

sums of the form

$$S = \sum_{i=1}^{V} h(p_i),$$

with $h$ some function of $p$, can be written as Stieltjes integrals:

$$S = \sum_{p} h(p)\Delta G(p)$$
$$= \int_0^1 h(p)dG(p);$$
$$S = \sum_{p} h(p)\frac{\Delta F(p)}{p}$$
$$= \int_0^1 h(p)\frac{dF(p)}{p}.$$

We can now rewrite the expected frequency distribution as

$$V_N(m) = E\hat{V}_N(m) = \int_0^1 \binom{N}{m} p^m(1-p)^{N-m}dG(p)$$
$$= \int_0^1 \binom{N}{m} p^m(1-p)^{N-m}\frac{dF(p)}{p} \quad (40)$$

$$V_N = E\hat{V}_N = \int_0^1 (1-(1-p)^N)dG(p) \quad (41)$$

$$G_N(p) = E\hat{G}_N = \int_0^1 B^*(N, Np, q)dG(q) \quad (42)$$

$$F_N(p) = E\hat{F}_N(p) = \int_0^1 qB^*(N-1, Np-1, q)dG(q). \quad (43)$$

In the same way the covariances can be presented as



$$COV(\hat{V}_N(m), \hat{V}_N(k)) = \delta_{ij}E\hat{V}_N(m)$$
$$+ \int_0^1\int_0^1 T(N, m, k, p, q)dG(p)dG(q)$$
$$- \int_0^1 T(N, m, k, p, p)dG(p)$$
$$- E\hat{V}_N(m)E\hat{V}_N(k). \quad (44)$$

Note that we do not exclude the possibility that the theoretical vocabulary may be infinite, i.e. $V = \infty$. This is the reason that we consider the upper cumulative distributions $G$ and $F$, using for brevity the notation $dG$, $dF$ instead of $(-dG)$, $(-dF)$.

### 3.1.2. The Poisson Model

Generally, we may consider the multinomial model within the framework of the Poisson model of the (sampling) experiment. In fact, if we assume that the frequencies of the vocabulary elements

$$(f_t(A_1), f_t(A_2), ..., f_t(A_V)),$$

are independent Poisson processes in continuous time $t \geq 0$ with parameters (intensities)

$$(\lambda(A_1), \lambda(A_2), ..., \lambda(A_V)),$$

then the vector of frequencies

$$(f_{T_N}(A_1), f_{T_N}(A_2),..., f_{T_N}(A_V))$$

observed at moments in time when the number of observed tokens is increasing,

$$T_N = \min\{t: \sum_{i=1}^{V} f_t(A_i) = N\},$$

is multinomially distributed according to the probability distribution

$$(p(A_i) = \frac{\lambda(A_i)}{\sum_{j=1}^{V} \lambda(A_j)}, \quad 1 \le i \le V).$$

Interestingly, for LNRE samples it may be assumed that for the terms of the frequency spectrum the multinomial and the Poisson schemes are (asymptotically) equivalent. In the Poisson scheme the expressions for the expected values become simpler. In particular, we now have

$$E\hat{V}_t(m) = \int_0^\infty \frac{(\lambda t)^m}{m!} e^{-\lambda t} dG(\lambda)$$

$$= \int_0^\infty \Pi(t, m, \lambda) dG(\lambda) \tag{45}$$

$$E\hat{V}_t = \int_0^\infty (1 - e^{-\lambda t}) dG(\lambda) \tag{46}$$

$$E\hat{G}_t(\lambda) = \int_0^\infty \Pi^+(t, t\lambda, x) dG(x) \tag{47}$$

$$E\hat{F}_t(\lambda) = \int_0^\infty x\Pi^+(t, t\lambda - 1, x) dG(x), \tag{48}$$

where $(\hat{G}_T(\lambda), \hat{F}_t(\lambda))$ and $(G(\lambda), F(\lambda))$ play the role of empirical and theoretical distributions for type and token intensities in the general population. We use the notations

$$\Pi(t, m, \lambda) = \frac{(\lambda t)^m}{m!} e^{-\lambda t} \tag{49}$$

$$\Pi^+(t, m, \lambda) = \sum_{k \ge m} \Pi(t, k, \lambda) \tag{50}$$

for the Poisson probabilities and the corresponding upper sums.

The expressions for covariances are simplified significantly since the trinomial distribution is substituted formally as follows:



$$\binom{N}{m, k} (p(A_i)^m (p(A_j))^k (1 - p(A_i) - p(A_j))^{N-m-k} \approx \frac{(\lambda(A_i)t)^m}{m!} \frac{(\lambda(A_j)t)^k}{k!} e^{-\lambda(A_i)t - \lambda(A_j)t}.$$

Hence, for instance,

$$\begin{aligned} \mathbf{COV}(\hat{V}_t(m), \hat{V}_t(k)) &= \sum_i \frac{(\lambda(A_i)t)^m}{m!} e^{-\lambda(A_i)t} \\ &+ \sum_{ij} \frac{(\lambda(A_i)t)^m}{m!} \frac{(\lambda(A_j)t)^k}{k!} e^{-\lambda(A_i)t - \lambda(A_j)t} \\ &- \sum_i \frac{(\lambda(A_i)t)^{m+k}}{(m+k)!} \binom{m+k}{m} \frac{1}{2^{m+k}} e^{-2\lambda(A_i)t} \\ &- \sum_i \frac{(\lambda(A_i)t)^m}{m!} e^{-\lambda(A_i)t} \sum_j \frac{(\lambda(A_j)t)^k}{k!} e^{-\lambda(A_j)t} \\ &= E\hat{V}_t(m) - \binom{m+k}{m} \frac{1}{2^{m+k}} E\hat{V}_{2t}(m+k). \end{aligned} \tag{51}$$

For reasons of expositional clarity we will henceforth use the more traditional notation $p$ (probability) instead of $\lambda$ (intensity) in expressions making use of the Poisson model, even though $p$ may now range over the whole interval $[0,\infty)$ rather than $[0,1]$. Similarly $pN$ will replace $\lambda t$.

### 3.1.3. Stochastic Interpretation of the Token Probability Distribution

Further insight into the structural token distribution can be gained by investigating how the text may be generated stochastically. Consider associating with any word token $w_n$, $1 \le n \le N$ in the running text both its relative frequency and its (theoretical) probability, such that the text is viewed as a series of triplets (word token, relative frequency, probability):

$$\begin{array}{llll} w_1, & w_2, & \ldots, & w_N \\ \hat{p}_N(w_1), & \hat{p}_N(w_2), & \ldots, & \hat{p}_N(w_N) \\ p(w_1), & p(w_2), & \ldots, & p(w_N). \end{array}$$

The probabilities on the second row are sampled from a population with dis-

tribution $\hat{F}$ (without replacement). The probabilities on the third row are sampled from a general population with distribution $F$ (with replacement).

Now the following scheme for the stochastic generation of texts can be suggested. At each $n$-th stage of the experiment, first generate the (random) probability $p_n$ according to the distribution $F(p)$, then choose some appropriate word type $A_i$ from all words in the vocabulary for which $p(A_i) = p(w_n) = p_n$. To define the last step of this experimental scheme somewhat more precisely, let

$$V(p, q) = (A_i: p \le p(A_i) \le q)$$

be the part of the vocabulary $\underline{V}$ consisting of words with probability falling in the interval $[p, q]$. Obviously, the number of elements $V(p, q)$ in this 'subvocabulary' is just

$$\begin{aligned} V(p, q) &= \sum_i \mathbf{I}_{[p < p(A_i) \le q]} \\ &= \int_p^q dG(x) \\ &= \int_p^q \frac{1}{x} dF(x). \end{aligned} \tag{52}$$

Now it is easy to see that the initial multinomial scheme of experiment is equivalent to that described above if only word $w_n$ is supposed to be chosen by chance (i.e., according to the uniform distribution) from the subvocabulary $V(p_n, q_n)$. In that case the variables $(w_1, w_2, \ldots)$ are independently and identically distributed. In addition,

$$\begin{aligned} \boldsymbol{Pr}(w_n = A_i) &= \frac{1}{V(p(A_i), p(A_i))} \int_{p(A_i)}^{p(A_i)} dF(p) \\ &= \frac{\Delta F(p(A_i))}{\Delta G(p(A_i))} = p(A_i). \end{aligned}$$

Thus the running text can be viewed as the realization of an experiment governed by two stochastic mechanisms:

1. the token probability distribution $F(p)$ to generate the probabilities $p_n$ at each stage, and
2. the (conditional) distribution $W(A \mid p)$ to generate words $w_n$ from subvocabularies corresponding to the probability $p_n$ occurring at this stage.



Moreover, in as far as we are restricting ourselves to the analysis of frequency distributions, and since the particular character of the second mechanism (notably the assumption of independence) does not affect the conclusions made on the basis of the frequency distribution data, we can accept far more general hypotheses concerning the nature of the word distribution scheme, the only requirement being that for any interval $[p, q]$ the elements of the subvocabulary $V(p, q)$ should be uniformly distributed over the set

$$\tau_1(p, q), \tau_2(p, q), \tau_3(p, q), \ldots$$

of positions through the running text at which the occurring probabilities $p_n$ fell in the interval $[p, q]$.

### 3.1.4. Interpolation

We sometimes need expressions for the vocabulary or the frequency spectrum for sample sizes smaller than $N$. More precisely, if $(\hat{V}_n(m), m = 1,2,\ldots)$ is a frequency spectrum observed on a sample of size $N$, the question is how to estimate the frequency spectrum $(\hat{V}_n(m), m = 1,2,\ldots)$ for a subsample of the size $n$. The formula

$$\hat{V}_{N,n}(m) = \sum_{j \ge m} \hat{V}_N(j) \binom{j}{m} \left(\frac{n}{N}\right)^m \left(1 - \frac{n}{N}\right)^{j-m} \tag{53}$$

gives the best solution to this problem: $\hat{V}_{N,n}(m)$ is a conditional expectation of the spectrum $\hat{V}_n(m)$ given the observed spectrum $(\hat{V}_N(k), k \ge 1)$:

$$\hat{V}_{N,n}(m) = \boldsymbol{E}(\hat{V}_n(m) \| \hat{V}_N(k), k \ge 1),$$

that is optimal in the mean squares sense.

To see this, consider a finite population of size $N$ consisting of $\hat{V}_N$ types of elements

$$(A_1, A_2, \ldots, A_{\hat{V}_N})$$

with corresponding frequencies

$$(f_N(A_1), \ldots, f_N(A_{\hat{V}_N})).$$

Let some sample of size $n$ be taken from this population without replacement.

Denote by $\hat{V}_{N,n}(k, l)$ the number of elements with a frequency $k$ in the population which occur $l$ times in the sample, i.e.

$$\hat{V}_{N,n}(k, l) = \sum_i \mathbf{I}_{[f_N(A_i) = k,\, f_n(A_i) = l]}. \tag{54}$$

Evidently the spectrum terms in the sample can be presented as sums

$$\hat{V}_n(l) = \sum_i \mathbf{I}_{[f_n(A_i) = l]} \tag{55}$$

$$= \sum_{k \geq l} \hat{V}_{N,n}(k, l). \tag{56}$$

It can be shown that the (matrix) statistic $\hat{V}_{N,n}(k, l)$, $l \leq k$, $1 \leq k$, is distributed by the compound hypergeometric law, i.e.

$$\boldsymbol{Pr}(\hat{V}_{N,n}(k, l) = m_{k,l},\ l \leq k;\ 1 \leq k \leq N) = \binom{N}{n}^{-1} \prod_{k=1}^{N} \frac{\hat{V}_N(k)!}{m_{k,1}! \ \ldots \ m_{k,k}!} \prod_{l=1}^{k} \binom{k}{l}^{m_{k,l}} \tag{57}$$

on the domain

$$(m_{k,l}: \sum_{l=1}^{k} m_{k,l} = \hat{V}_N(k),\ \sum_k \sum_l m_{k,l} = n).$$

For sufficiently large sample sizes $(n, N)$ the vectors

$$(\hat{V}_{N,n}(k, l),\ l \leq k),\ k = 1, 2, \ldots$$

are independent in $k$ and multinomially distributed. As a result formula (53) can be derived as well as an expression for the expected vocabulary,

$$\hat{V}_{N,n} = \sum_{j \geq 1} \hat{V}_N(j) \left(1 - \left(1 - \frac{n}{N}\right)^j\right). \tag{58}$$



### 3.2. The LNRE ZONE

According to the law of large numbers, we have that for any probability distribution

$$(p(A_i),\ 1 \leq i \leq V)$$

with a finite vocabulary $V$ the relative sample frequencies will converge to the population probabilities for ever increasing sample size $N$:

$$\hat{P}_N(A_i) \rightarrow Pr(A_i)$$

in probability as $N \rightarrow \infty$. As a simple consequence,

$$\hat{V}_N(m) \rightarrow 0$$

for all $m$. If so, the relative expected spectrum

$$\alpha_N(m) = \frac{\boldsymbol{E}\hat{V}_N(m)}{\boldsymbol{E}\hat{V}_N} = \frac{V_N(m)}{V_N}$$

may coinside with most of the 'laws' (15-20) only for finite samples ($N < \infty$). If one of these 'laws' appears for $N = \infty$, then the general population must be necessarily infinite too. In qualitative terms, a sample in the LNRE ZONE can be defined as a sample for which (a) the sample size is large enough to allow the estimation of the first terms of the probability rank distribution (the big probabilities), but where (b) the first terms of the relative frequency spectrum take on significant values. The questions we shall try to give an answer to by applying the analytical expressions for the expected spectrum can be formulated as follows:

1. What is the empirical criterion for the LNRE ZONE? In other words, how can we ascertian whether a sample is located in the LNRE ZONE?
2. What is the theoretical definition for the LNRE ZONE? In particular, does a theoretical distribution exist which realizes some given 'law' either on finite or on infinite samples, such that the coincidence

   $$\alpha_N(m) = \alpha(m),\ m = 1, 2, \ldots$$

   takes place for 'laws' $\alpha(m)$ such as (15-20)?

### 3.2.1. Locating Samples with Respect to the LNRE ZONE

In this section we address the first question, proposing two methods for ascertaining whether a sample is located in the LNRE ZONE. The first method makes use of the way the frequency spectrum develops through sampling time, the second method gauges the extent to which the sample relative frequencies are biased estimates of the population probabilities.

The following reasoning corresponds to the intuitive understanding that the LNRE ZONE must be located in the neighborhood of sample sizes where the (relative) expected spectrum terms achieve their peaks. Using the Poisson model for the terms of the expected spectrum and differentiating in the variable $N$ we find that

$$\frac{d}{dN}V_N = \frac{1}{N}V_N(1) \tag{59}$$

$$\frac{d}{dN}V_N(m) = \frac{1}{N}[mV_N(m) - (m+1)V_N(m+1)],\ m = 1,2,... \tag{60}$$

$$\frac{d}{dN}\alpha_N(m) = \alpha_N(m)[m - \alpha_N(1)] - (m+1)\alpha_N(m+1),\ m = 1,2,... \tag{61}$$

Denote by $N_m^*$, $m = 1, 2, ...$ and $\overline{N}_m$, $m = 1, 2, ...$ the values of those sample sizes where the terms of the absolute ($V_N(m)$) or relative ($\alpha_N(m)$) expected spectrum are achieving their maximums respectively. Interestingly, the dynamic behavior of these functions is characterized by the following property. At the time moment $N = N_1^*$ at which the expected number of hapax legomena (the words occuring only once) $V_N(1)$ reaches its maximum the number of hapaxes is exactly twice that of the dislegomena $V_N(2)$: from

$$\frac{d}{dN}V_N(1) = 0$$

we have by (60) that

$$\frac{V_N(1)}{N} - \frac{2V_N(2)}{N} = 0,$$

hence $V_N(1) = 2V_N(2)$ at $N_1^*$. Similarly, the number of dislegomena increases until at time moment $N = N_2^*$ it becomes 3/2 of the expected number of words occurring three times, $V_N(3)$, and so on. The moments of peaks of the relative expected spectrum $\alpha_N(m)$ are arranged in the same way but with some (often substantial) anticipation,



$$\overline{N}_1 \le N_1^*,\ \overline{N}_2 \le N_2^*,...$$

Now suppose that Zipf 's law is realized for some sample size $Z$. When we substitute the expression 1/[m(m+1)] in the right hand side of (61) and take

$$\frac{d}{dN}\alpha_N(m) = 0,$$

then it is easy to see that

$$\frac{d}{dN}\alpha_N(1)\Big|_{N=Z} < 0$$

and that

$$\frac{d}{dN}\alpha(2)\Big|_{N=Z} = 0.$$

Hence we have that

$$\overline{N}_1 \le Z = \overline{N}_2 \le N_1^*.$$

Thus $Z$ appears as the sampling time at which the relative number of expected dislegomena E $\alpha_N(2)$ achieves its maximum.

Given some observed frequency spectrum ($\hat{V}_N(1)$, $\hat{V}_N(2)$,...), we may test whether a sample is located in the LNRE zone at sample size $N$ by inquiring whether the number of hapaxes and dislegomena are still increasing. If $N \le N_1^*$, that is,

$$\frac{1}{N}(\hat{V}_N(1) - 2\hat{V}_N(2)) > 0, \tag{62}$$

we know that the number of hapaxes is still increasing; if $N_1^* \le N \le N_2^*$, that is,

$$\begin{cases} \frac{1}{N}(\hat{V}_N(1) - 2\hat{V}_N(2)) < 0 \\ \frac{1}{N}(2\hat{V}_N(2) - 3\hat{V}_N(3)) > 0 \end{cases} \tag{63}$$

we know that the number of hapaxes has passed its maximum while the number of dislegomena is still increasing. We will refer to the 'time' interval $(0, N_1^*]$ as the central LNRE ZONE and to the interval $(N_1^*, N_k^*]$ for small $k$ as the late LNRE ZONE.

To see how this test can be applied to actual data, consider figures 8 and 9.

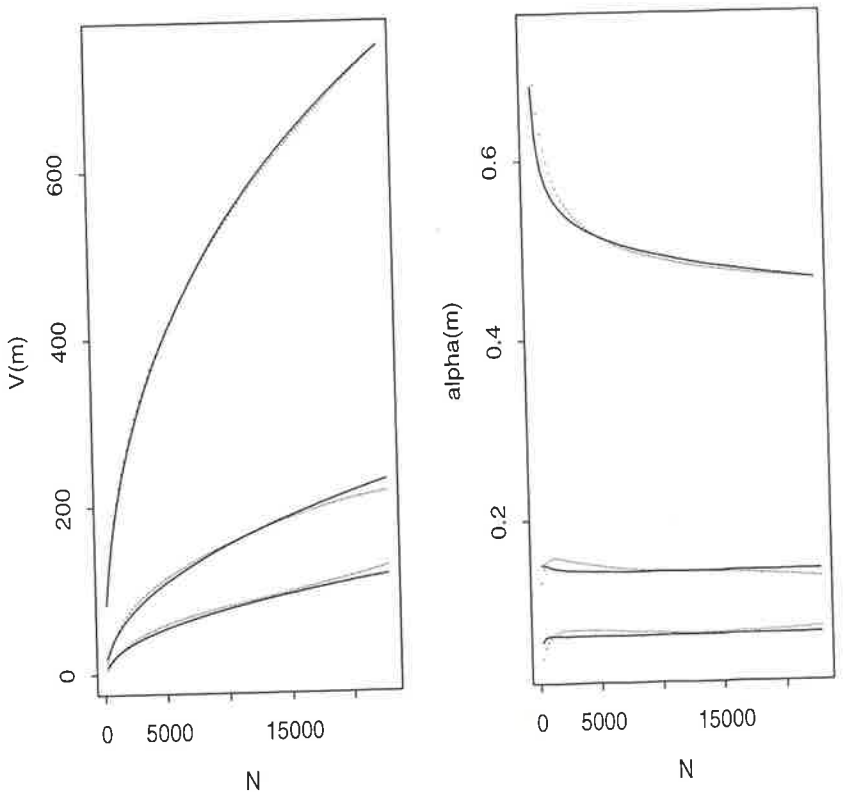



Figure 8. The development of the absolute and relative frequency spectra ($m$ = 1, 2, 3) through sampling time: the productive English suffix *-ness*. The continuous lines are calculated on the basis of the inverse Gauss-Poisson 'law', the dotted lines are obtained by hypergeometric interpolation.

Figure 8 shows how the first three spectrum elements develop through sampling time for the productive English suffix *-ness*. Since $V_N(1)$ and $V_N(2)$ are still increasing at the observed sample size, we may conclude that this sample is located in the central LNRE ZONE. Next consider the corresponding graphs for



the unproductive English prefix *en-* (figure 9). According to the Gauss-Poisson model (see sections 3.2.2 and 5.1), the sample is at a position far beyond the late LNRE ZONE. The hypergeometric interpolation curves appear to be less useful here, due to the presence of extra maxima which are brought about by the combined presence of a substantial number of very high frequency words and a smallish number of low frequency words. In this case, the early maxima observed for small $N$ are indicative of the sample's location outside the late LNRE ZONE. Note that the relative spectrum elements reach their maxima far earlier than the absolute spectrum elements, which is the reason why the test is formulated in terms of $V_N(m)$ rather than in terms of $\alpha_N(m)$.

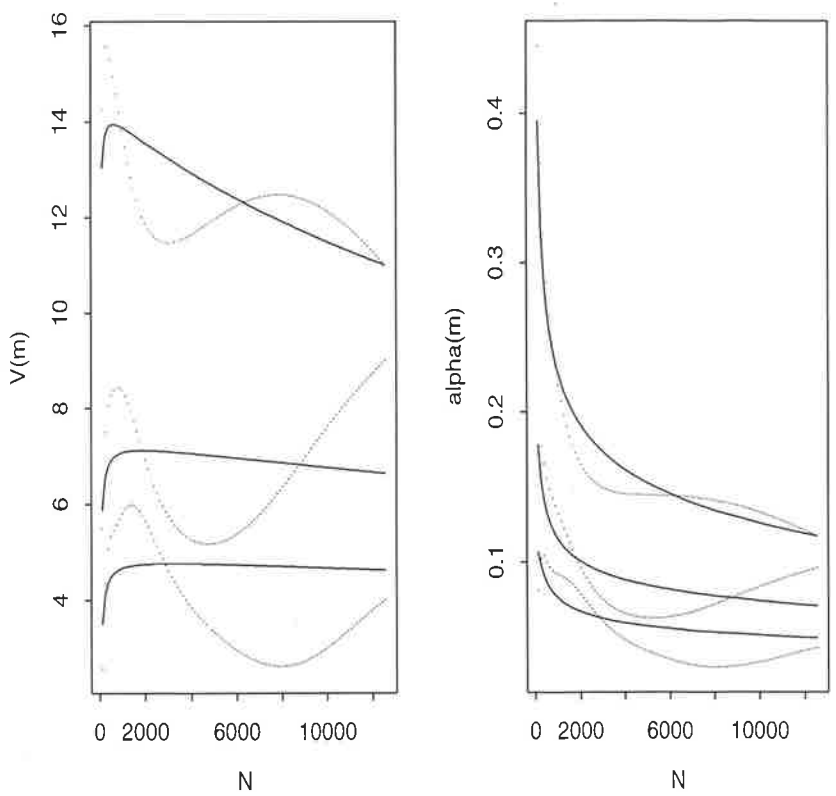



Figure 9. The development of the absolute and relative frequency spectra ($m$ = 1, 2, 3) through sampling time: the unproductive English prefix *en-*. The continuous lines are obtained using the inverse Gauss-Poisson 'law'. The dotted lines are obtained by hypergeometric interpolation.

For not too small samples located beyond the LNRE ZONE the absolute and relative spectrum elements are (globally) decreasing functions of $N$. Note that the extreme case when

$$\overline{N}_1 = \overline{N}_2 = \ldots$$

can take place only if

$$\overline{N}_1 = \overline{N}_2 = \ldots = \infty,$$

in which case the relative expected spectrum terms are increasingly converging as the sample size $N \to \infty$. If some limiting 'law' $\alpha(m)$ is to hold for $N \to \infty$,

$$\alpha(m) = \lim_{N\to\infty} \alpha_N(m)$$

then the (stationarity) condition

$$\lim_{N\to\infty} \frac{d}{dN}\alpha_N(m) = \alpha(m)(m - \alpha(1)) - (m + 1)\alpha(m + 1), \qquad (64)$$

should necessarily be satisfied. In other words, the growth rates of the spectrum elements should no longer vary with $N$. The unique solution to (64) is

$$\alpha(m) = \frac{\alpha\Gamma(m - \alpha)}{\Gamma(1 - \alpha)\Gamma(m + 1)}.$$

Thus Karlin-Rouault's 'law' appears as the only parametric family of limiting 'laws'. In section 4.2 we shall give a description of the theoretical structural distributions which can realize this law. First, however, we consider an alternative method for establishing whether a sample is located in the LNRE ZONE.

From a standard asymptotic point of view we may consider ourselves as situated outside the LNRE ZONE when, roughly speaking, we can convince ourselves that the empirical distribution $(\hat{p}_N(A_i))$ is so close to the theoretical distribution that we can allow ourselves to replace theoretical expectations by empirical ones. If the sample is located outside of the LNRE ZONE, the expected spectrum elements can be approximated by the expressions

$$\hat{\boldsymbol{E}}\hat{V}_N(m) = \sum_i \frac{(\hat{p}_N(A_i)N)^m}{m!} e^{-\hat{p}_N(A_i)N}$$



$$= \int_0^\infty \frac{(\hat{p}_N N)^m}{m!} e^{-\hat{p}_N N} dG_N(\hat{p}_N)$$

$$\hat{\boldsymbol{E}}\hat{V}_N = \sum (1 - e^{-\hat{p}_N(A_i)N})$$

$$= \int_0^\infty (1 - e^{-\hat{p}_N N}) dG_N(\hat{p}).$$

We can use the differences between the expected values

$$\boldsymbol{E}\hat{V}_N(m) - \boldsymbol{E}\hat{\boldsymbol{E}}\hat{V}_M(m)$$

$$\boldsymbol{E}\hat{V}_N - \boldsymbol{E}\hat{\boldsymbol{E}}\hat{V}_N$$

to evaluate the accuracy of the approximation and the extent of the bias introduced by estimating population probabilities by sample relative frequencies. Focussing on $\boldsymbol{E}\hat{\boldsymbol{E}}\hat{V}_N$, the expected vocabulary at the sample size $N$ if instead of the theoretical probabilities the empirical distribution $(\hat{p}_N(A_i),\ 1 \le i \le \hat{V})$ is used to simulate the experiment on the same sample size, we find that

$$\boldsymbol{E}\hat{\boldsymbol{E}}\hat{V}_N = \boldsymbol{E}\sum_{m\ge 1} (1 - e^{-m})\hat{V}_N(m)$$

$$= \sum_{m\ge 1} (1 - e^{-m})\boldsymbol{E}\hat{V}_N(m)$$

$$= \boldsymbol{E}\hat{V}_{(1-e^{-1})N}.$$

In other words, if the expected vocabulary at the smaller sample size $0.63N$ is approximately the same as for the sample size $N$, the sample is not located in the LNRE ZONE. This state of affairs obtains only when $\frac{d}{dM} V_{M|M=0.63\,N} \approx 0$.

A computationally convenient test is to consider the ratio

$$C_L = \frac{\hat{V}_N - \sum_m (1 - e^{-m})\hat{V}_N(m)}{\hat{V}_N} = \sum_m \hat{\alpha}_N(m)e^{-m}, \qquad (65)$$

large values of which can be used to identify the LNRE zone. By way of example, suppose Zipf 's law is valid in the zone we are situated in. We then obtain

$$\begin{aligned} \boldsymbol{E}\hat{\boldsymbol{E}}V(N) &= \sum_{m\geq 1}(1 - e^{-m})\boldsymbol{E}\left[V(N)\frac{1}{m(m+1)}\right] \\ &= \boldsymbol{E}[V(N)(e - 1)(1 - \ln(e - 1))] \\ &\sim 0.5\boldsymbol{E}V(N), \end{aligned}$$

so that in this case we are loosing about half of the vocabulary on the assumption that we are not positioned in the LNRE ZONE. Some typical empirical examples are presented in table 1. The sample of words prefixed with *en-* appears with the lowest score for $C_L$. This accords well with our previous findings concerning the very early stage at which $V_N(1)$ achieves its maximum for *en-* (see figure 9). Although low values of $C_L$ are typical for unproductive affixes, they are rarely observed for texts. However, we would not be surprised to find that the very large corpora that are at present being compiled ($N$ » 300,000,000) will be located outside the central LNRE ZONE and probably outside the late LNRE ZONE as well.

Table 1. Some typical $C_L$ values for various kinds of word frequency distributions

| sample type | $C_L$ | $V$ | $n_1$ |
|---|---|---|---|
| English *-ness (Cobuild)* | 0.195 | 1607 | 749 |
| English *en-* | 0.059 | 94 | 11 |
| Dutch *-heid* | 0.228 | 466 | 256 |
| Durch *-ing* | 0.147 | 942 | 302 |
| Carroll's *Alice in Wonderland* | 0.163 | 1930 | 721 |
| Bronte's *Wuthering Heights* | 0.165 | 6420 | 2427 |
| Pushkin's *Captain's Daughter* | 0.213 | 4783 | 2384 |

### 3.2.2. Generalized Structural Distributions

We have already discussed the fact that Rouault's 'law' appears as the only limiting 'law' for $N \to \infty$. We now turn to consider the question whether theoretical structural distributions can be found that realize some 'law' for a finite, specific sample size $Z$. The unique solution for $G$ as the (unknown) structural distribution appearing in the formula for the relative expected spectrum using the Poisson model,



$$\alpha(m) = \frac{\int_0^\infty \frac{(pZ)^m}{m!}e^{-pZ}dG(p)}{\int_0^\infty (1 - e^{-pZ})dG(p)}, \tag{66}$$

given that one of the Zipfian laws (15-20) is substituted for $\alpha(m)$ in (66), takes the parameterized form

$$G(p) = C\int_p^\infty e^{-Zpx}\frac{(\log(1+x))^{\gamma-1}x^{\alpha-1}}{(1+x)^{\beta+1}}dx, \tag{67}$$

with some constant $C$ (cf. Orlov and Chitashvili 1983b). In fact, if we substitute $G(p)$ in (67) into (66), the relative expected spectrum $\alpha(m)$ can be expressed as

$$\alpha(m, \alpha, \beta, \gamma) = \frac{\int_0^\infty \frac{(\log(1+x))^{\gamma-1}x^{\alpha}}{(1+x)^{m+\beta+1}}dx}{\int_0^\infty \frac{(\log(1+x))^{\gamma-1}x^{\alpha-1}}{(1+x)^{\beta+1}}dx}. \tag{68}$$

All laws (15-20) appear as special submodels for particular choices of the parameters $\alpha$, $\beta$ and $\gamma$ (Zipf: $\alpha = \beta = \gamma = 1$, Yule: $\alpha = \beta$, $\gamma = 1$, Yule-Simon: $\alpha = 1$, $\gamma = 1$, Waring-Herdan: $\gamma = 1$, Karlin-Rouault: $\beta = 0$, $\gamma = 1$, Zipf-Mandelbrot: $\alpha = \beta = 1$). Unfortunately, expression (67) does not represent any real structural distribution because

1. $G(p)$ is not a step (or step-wise constant) function,
2. the distribution
$$F(p) = \int_p^\infty x\,dG(x)$$
may not be a normalized distribution, and
3. the theoretical vocabulary $V = G(0)$ may be infinite.

Nevertheless, the reasoning presented above at least makes it natural to admit generalized forms for structural distributions so long as they allow us to formulate expressions for the expected spectrum at prescribed sample sizes.

In addition to the Zipfian family (15-20) defined by (67), to which we shall refer as the generalized Zipf 's structural distribution, two other structural distri-

butions, i.e. decreasing functions $G(p)$ of a general nature, should be mentioned. These distributions, the lognormal distribution (Herdan 1960, Carroll 1967, 1969)

$$G(p) = \frac{1}{\sigma\sqrt{2\pi}} \int_p^\infty \frac{1}{x^2} e^{-\frac{1}{2}\left(\frac{\log(x)-\mu}{\sigma}\right)^2} dx \tag{69}$$

and the generalized inverse Gauss-Poisson distribution (Sichel 1976,1986)

$$G(p) = \frac{2^\gamma}{(bc)^{\gamma+1} K_{\gamma+1}(b)} \int_p^\infty x^{\gamma-1} e^{-\frac{x}{c} - \frac{b^2 c}{4x}} dx, \tag{70}$$

where $K_\gamma(b)$ is the modified Bessel function of the second kind of order $\gamma$ and argument $b$, allow the expected spectrum to be defined as

$$V_N(m) = \int_0^\infty \frac{(pN)^m}{m!} e^{-pN} dG(p). \tag{71}$$

In both cases the structural distributions may be presented in the form

$$G(p) = \frac{Z}{\sigma\sqrt{2\pi}} \int_{pZ}^\infty \frac{1}{y^2} e^{-\frac{(\log y)^2}{2\sigma^2}} dy = ZG^*(pZ) \tag{72}$$

and

$$G(p) = \frac{2^\gamma}{cb^{\gamma+1} K_{\gamma+1}(b)} \int_{pZ}^\infty y^{\gamma-1} e^{-y-\frac{b^2}{4y}} dy = ZG^o(pZ) \tag{73}$$

with the parameters $Z = e^{-\mu}$ and $Z = 1/c$ playing the role of the sample-locator defining the sample's position with respect to the LNRE ZONE.

### 3.2.3. Simulating Generalized Laws

We now present an algorithm by which an experiment (in the framework of the multinomial model) could be simulated (approximately) corresponding to some generalized structural distribution. In other words, given the generalized probability type distribution $G^o(p)$ defined by the relation



$$\alpha(m) = \int_0^\infty \frac{p^m}{m!} e^{-p} dG^o(p), \tag{74}$$

for some relative spectrum 'law' $\alpha(m)$, we want to construct (for a sufficiently large $N$) the set of probabilities

$$(p_{i,N},\ 1 \le i \le V) \tag{75}$$

such that the corresponding structural distribution

$$G(p) = \sum_{i=1} \mathbf{I}_{[p_{i,N} > p]} \tag{76}$$

realizes on a sample of size $N$ the relative expected spectrum

$$\alpha_N(m) = \frac{V_N(m)}{V_N} = \frac{\sum_{i=1}^V \frac{(p_{i,N}N)^m}{m!} e^{-p_{i,N}N}}{\sum_{i=1}^V (1 - e^{-p_{i,N}N})}$$

$$= \frac{\int_0^\infty \frac{p^m}{m!} e^{-p} dG(\frac{p}{N})}{\int_0^\infty (1 - e^{-p}) dG(\frac{p}{N})} \approx \int_0^\infty \frac{p^m}{m!} e^{-p} dG^o(p) = \alpha(m), \tag{77}$$

with $G^0$ the standardized correlate of $G$. To do this, construct for $\varepsilon > 0$ the sequence

$$(\lambda_i(\varepsilon),\ 1 \le i \le V)$$

from the relations

$$G^0(\lambda_1(\varepsilon)) = \varepsilon$$
$$G^0(\lambda_i(\varepsilon)) = G^0(\lambda_{i-1}(\varepsilon)) + \varepsilon,\ i \ge 2, \tag{78}$$

where

$$V = V_\varepsilon = \min(k: \sum_{i=1}^{k}(1 - e^{-\lambda_i(\varepsilon)}) \geq \frac{1}{\varepsilon}). \tag{79}$$

Now define

$$n_\varepsilon = [\sum_{i=1}^{V} \lambda_i(\varepsilon)]$$

and construct the probabilities by the formula

$$p_{i,N} = \frac{\lambda_i(\varepsilon_N)}{n_{\varepsilon_N}}, \quad 1 \leq i \leq n_{\varepsilon_N} \tag{80}$$

with $\varepsilon_N$ chosen so to satisfy

$$\varepsilon_N = \max(\varepsilon: n_\varepsilon \geq N). \tag{81}$$

### 3.3. Asymptotic Approach

Under what conditions can the use of generalized structural distributions be justified? This question is discussed in section 3.3.1. Section 3.3.2 considers the accuracy of the theoretical models, and section 3.3.3 calls attention to the independence of the high and low frequency 'tails' of LNRE distributions.

#### 3.3.1. The Triangle Scheme of Experiment

We may justify generalized structural distributions (or generalized population probability distributions) by using the asymptotic approach argument. Although the LNRE ZONE is usually located at rather early stages of (imaginable) experiments, the samples in which the characteristic features of LNRE distributions are present are often large enough to apply the asymptotic analysis.

Within the framework of the classical scheme of experiment, the only way to justify generalized distributions is to admit the so-called triangle scheme of experiment, i.e. to consider the asymptotic scheme when (i) the normalized theoretical structural distribution



$$\frac{G^Z\left(\frac{p}{Z}\right)}{\int_0^\infty (1 - e^{-x}) dG^Z\left(\frac{x}{Z}\right)},$$

indexed by some parameter $Z$, approaches some generalized structural distribution $G^0$, and when (ii) the sample size $N$ is taken in the neighbourhood of $Z$ (considered as the center of the LNRE ZONE), i.e. $N \approx Z$.

We assume that the token probabilities $p_n$ are independent and identically distributed. Informally speaking, we find that at first sight there appears to be no distinction between the following suggestions:

A. The author selects some structural distribution $F(p)$ and then generates (creates) a text of some sufficiently large size $N$ according to this distribution;

B. The author determines some sample size $Z$ (the desired horizon) and chooses the structural distribution intending to get some (desired) frequency distribution 'law' on a sample of size $Z$ and then generates a text of a size $N \asymp Z$.

But the distinction becomes obvious when we set the problem in asymptotic scheme. In fact, let some 'law' $\alpha(m)$, $m \geq 1$ be fixed. Now let the problem of the existence of a structural distribution realizing this law be stated in an asymptotic form, i.e., does a sequence $G^Z(p)$, $Z \gg 1$ of structural distributions exist such that the relative expected spectrum

$$\alpha_Z(m) \approx \alpha(m), \quad m \geq 1. \tag{82}$$

But this approximation takes place if and only if the normalized structural distribution is approximated (for $Z \gg 1$) by some (generalized and normalized) distribution $G^0(p)$,

$$G^0(p) \approx \frac{G^Z(\frac{p}{Z})}{\int_0^\infty (1 - e^{-x}) dG^Z(\frac{x}{Z})}, \tag{83}$$

where $G^0(p)$ is uniquely determined from the equation

$$\alpha(m) = \int_0^\infty \frac{p^m}{m!} e^{-p} dG^0(p), \quad m = 1, 2, \ldots \tag{84}$$

In other words, $G^Z(p)$ with the property (82) for sufficiently large $Z$ can be represented as

$$G^Z(p) \approx V_Z G^0(pZ), \tag{85}$$

where

$$V_Z = \int_0^\infty (1 - e^{-x}) dG^Z\left(\frac{x}{Z}\right)$$

is the expected vocabulary on the sample size $Z$ and with the generalized structural distribution $G^0$ defined by (83). Thus to use the generalized structural distribution is equivalent to accept the hypothesis:

$$G(p) = G^Z(p)$$

with $G^Z(p)$ satisfying (82) for some 'law' $\alpha(m)$. Note that for the transition from a discrete step function for the structural distribution to a continuous function $G^0$ to be justified in the triangle scheme, the parameter $Z$, where $Z = e^{-\mu}$ for the lognormal model and $Z = 1/c$ for the generalized inverse Gauss-Poisson model, should assume a value not too different from $N$ - the ratio $t = N/Z$ should not be too small or too large.

Thus for samples of size $N \gg 1$ we have two possibilities for the asymptotics of the relative expected spectrum. If $G$ is fixed, that is, if we drop the index $Z$ from $G^Z$, we again have Rouault's 'law'

$$\alpha(m) = \frac{\alpha\Gamma(m - \alpha)}{\Gamma(1 - \alpha)\Gamma(m + 1)}$$

as the only limiting distribution. If we allow $G$ to be parameterized for $Z$ such that (83) is satisfied, the triangle scheme leads to the following expression for $\alpha_N(m)$:

$$\alpha_N(m) \approx \frac{\int_0^\infty \frac{(pt)^m}{m!} e^{-pt} dG^0(p)}{\int_0^\infty (1 - e^{-pt}) dG^0(p)}, \tag{86}$$

a parametric family of 'laws' extended in sampling time and parametrized by $Z$, the 'Zipf' size, or equivalently by the parameter $t = N/Z$. For the generalized Zipf 's 'law' (68) the extended version takes the form



$$\alpha_N(m, \alpha, \beta, \gamma, t) = t^{m-1} \frac{\int_0^\infty \frac{[\ln(1 + y)]^{\gamma-1} y^\alpha}{(t + y)^{m+1}(1 + y)^{\beta+1}} dy}{\int_0^\infty \frac{[\ln(1 + y)]^{\gamma-1} y^{\alpha-1}}{(t + y)(1 + y)^\beta} dy}. \tag{87}$$

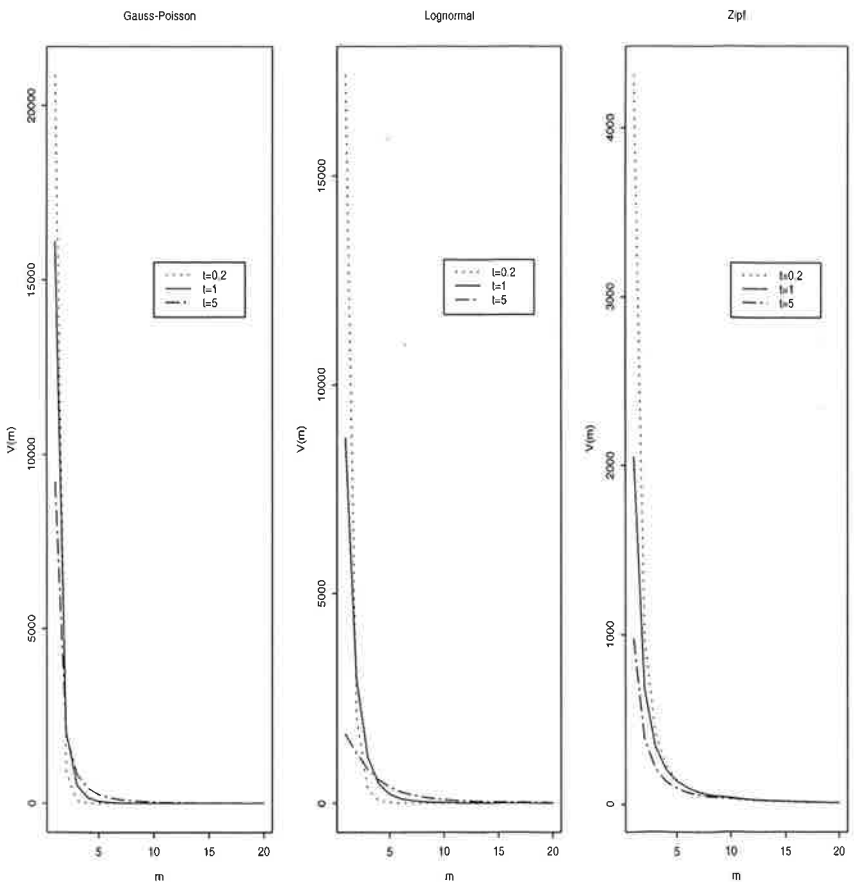



Figure 10. The role of the parameter $t$ in the extended generalized Zipf 's 'law', the lognormal 'law' and the generalized inverse Gauss-Poisson 'law'.

For Sichel's Gauss-Poisson 'law', $\alpha_N(m)$ can be expressed as

$$\alpha_N(m,\gamma,b,t) = \frac{1}{(1-\frac{t}{t+1})^{-\gamma/2}K_\gamma\left(b\sqrt{(1+t)(1-\frac{t}{1+t})}\right) - K_\gamma(b\sqrt{1+t})}$$

$$\frac{(0.5b\sqrt{1+t}\,\frac{t}{1+t})^m}{m!}K_{\gamma+m}(b\sqrt{1+t}).$$

For Carroll's lognormal model we finally have that

$$\alpha_N(m,\sigma,t) = \frac{\int_0^\infty \frac{1}{p^2}\frac{(pN)^m}{m!}e^{-pN-\frac{1}{2}\left(\frac{\log(pN/t)}{\sigma}\right)^2}dp}{\int_0^\infty \frac{1}{p^2}(1-e^{-pN})e^{-\frac{1}{2}\left(\frac{\log(pN/t)}{\sigma}\right)^2}dp}. \quad (88)$$

Figure (10) illustrates the role of the parameter $t$ for these three 'laws'. The Gauss-Poisson 'law' is shown for $\gamma = -0.5$ and $b = 0.01$, the lognormal 'law' for $\sigma = 1$ and the Waring-Herdan 'law' for $\alpha = \beta = 1$ (Zipf). For all models, increasing $t$ leads to theoretical distributions in which the lowest frequency types play less prominent roles, as expected for samples moving away from the LNRE ZONE.

### 3.3.2. The Accuracy of Theoretical Models

In the LNRE ZONE the accuracy of theoretical models for frequency distributions can be treated in the gaussian framework. Formally, if

$$N \gg 1,\ V_N \gg 1,\ \frac{\boldsymbol{E}\hat{V}_N(m)}{V_N} \asymp \alpha(m) > 0,$$

then

$$\left(\frac{\hat{V}_N(m) - \boldsymbol{E}\hat{V}_N(m)}{\sqrt{V_N}},\ m \geq 1\right) \overset{D}{\sim} N(0,R), \quad (89)$$



by which we mean that the normalized spectrum is approximated by the gaussian vector with 0 mean and covariance matrix

$$R_{m,k} = \delta_{m,k}\alpha_N(m) - \binom{m+k}{k}\frac{1}{2^{m+k}}\alpha_{2N}(m+k). \quad (90)$$

With respect to the sequence $(\hat{V}_N,\ 1 \leq n \leq N)$ of observed values of the empirical vocabulary volumes through sampling time we have

$$\left(\frac{\hat{V}_n - V_n}{\sqrt{V_n}},\ 1 \leq n \leq N\right) \overset{D}{\sim} N(0,\overline{R}), \quad (91)$$

where $n$ denotes the current sample size and where the covariance matrix

$$\overline{R}_{n,k} = \frac{\boldsymbol{COV}(\hat{V}_n - V_n, \hat{V}_k - V_k)}{V_N},\quad 1 \leq n, k \leq N$$

can be given in the form

$$\overline{R}_{n,k} = V_{n+k} - V_{\max(n,k)},\quad 1 \leq n, k \leq N. \quad (92)$$

If we use the interpolation formula

$$\hat{V}_{N,n} = \sum_{j\geq 1}\hat{V}_N(j)\left(1-(1-\frac{n}{N})^j\right)$$

for the vocabulary growth curve to estimate the accuracy of the model, we may use the approximation

$$\left(\frac{\hat{V}_{n,N} - V_n}{\sqrt{V_N}},\ 1 \leq n \leq N\right) \overset{D}{\sim} N(0,\tilde{R}) \quad (93)$$

with covariance matrix

$$\tilde{R}_{n,k} = V_{n+k} - V_{n+k-\frac{nk}{N}},\quad 1 \leq n, k \leq N. \quad (94)$$

### 3.3.3. The Distribution of the Frequency Spectrum

Even though the lower elements of the frequency spectrum are the most important for LNRE samples, more global analyses of frequency distributions including the highest frequency terms are by no means devoid of interest. Speaking in terms of the structural type frequency distribution $\hat{G}_N(p)$, the theoretical models considered above were intended to give satisfactory approximations for the left hand tails, i.e.

$$G(\frac{p}{N}) \asymp E\hat{G}_N(\frac{p}{N}) .$$

To test some theoretical model for the structural distribution $G(p)$ on the whole range of values $0 \le p \le 1$ it is useful to know that the differences

$$\Delta_N^+(p) = (\hat{G}_N(p) - E\hat{G}_N(p)) \tag{95}$$

$$\Delta_N^-(p) = (\hat{G}_N(\frac{p}{N}) - E\hat{G}_N(\frac{p}{N})) \tag{96}$$

are asymptotically gaussian with variances

$$\sigma^2\Delta_N^+(p) \sim \frac{1}{N} \tag{97}$$

$$\sigma^2\Delta_N^-(p) \sim EV_N , \tag{98}$$

and that, significantly, these differences are not correlated so that

$$COV(\sqrt{N}\Delta_N^+(p), \frac{1}{\sqrt{EV_N}}\Delta_N(p)) \approx \sqrt{\frac{EV_N}{N}} \int_0^\infty x(1 - \Pi^+(p,x))dG^0(x). \tag{99}$$

The important conclusion from this fact, which might be expected intuitively, is that mathematical models for tail and high frequency zones can be suggested independently. In particular, if some analytical expression $(\alpha(m), m = 1, 2, ...)$ is suggested for the relative expected spectrum, then we may write

$$E\hat{G}_N(p) = E\hat{V}_N(\sum_{m \ge pN} \alpha_N(m)) + \Delta_N(p), \tag{100}$$



where $\Delta_N(p)$ is intended to improve the fit for not small values of $p$, with the only property that

$$\frac{1}{E\hat{V}_N}\Delta_N(\frac{p}{N}) \to 0, \ N \to \infty. \tag{101}$$

Note that $\Delta_N(p)$ is exactly the parameter $B$ in the models (21) and (22) discussed in section 2.2. This extra parameter can be used, in particular, to improve the fit with the theoretical rank probability distribution at the left hand tail (i.e. the high probability region) without affecting the low frequency zone.

## 4. LNRE Models and their Rationales

In this section we shall present and try to systematize different mathematical models intended as analytical tools for LNRE samples. Since the empirical frequency distribution is the main object for mathematical modelling, the practical output of any such mathematical model is to suggest some analytical expression such as Zipf 's law for the frequency distribution as described by the rank frequency distribution, the frequency spectrum, or the cumulative type or token frequency distributions. By the interpretation of the corresponding analytical expressions these models can be divided into three essentially different classes:

1. Models which consider the analytical expressions used to approximate the rank-frequency distribution as structural probability distributions of a general population. Typically, such models focus on developing stochastic schemes generating such populations.

2. Models which consider the analytical expressions used to approximate the frequency spectrum as limiting distributions that characterize the equilibrium state. Typically, these models focus on stochastic schemes leading to the desired steady (equilibrium) state.

3. Models which consider the analytical expressions used to approximate the frequency spectrum as expected values for finite samples. These models focus on general population models realizing these laws on finite samples.

### 4.1. Mandelbrot and Miller

Mandelbrot's rank-probability distribution

$$p\{r\} = \frac{C^{1/\gamma}}{(r + B)^{1/\gamma}} \tag{102}$$

or the corresponding structural type distribution

$$G(p) = \frac{C}{p^{\gamma}} + B$$

has proved to be a good enough approximation for a number of observed distributions $\hat{p}_N\{r\}$ or $\hat{G}_N(p)$. Hence, from the point of view of traditional probabilistic modelling, it seems natural to be interested in general populations with a rank-probability distribution of this form. Two approaches in this direction should be mentioned. Mandelbrot (1953,1962) has shown that the significance of the distribution $p\{r\}$ can be explained by its optimality property of maximizing the information contained in a message constructed of words as sequences of letters indexed by different costs. Miller (1957) presented a pure probabilistic model where $p\{r\}$ appears as a rank probability distribution of words viewed as sequences of letters chosen by chance at each stage of an experiment (as if a monkey were typing text), and where the parameters $(B, C, \gamma)$ depend on the probability of a blank space and the number of letters.

But the significant bias between theoretical and empirical distributions symptomatic for LNRE renders such interpretations unconvincing, at least in the framework of the classical scheme of experiment. In fact, if the Zipf-Mandelbrot 'law' is taken as the theoretical probability distribution, then the relative frequency spectrum terms are converging to the expressions

$$\frac{\hat{V}_N(m)}{\hat{V}_N} = \hat{\alpha}_N(m) \to \frac{\alpha\Gamma(m - \alpha)}{\Gamma(1 - \alpha)\Gamma(m + 1)}, \tag{103}$$

i.e. to the Karlin-Rouault 'law', instead of to the expression

$$\alpha(m) = \frac{1}{m^{\gamma}} - \frac{1}{(m + 1)^{\gamma}},$$

the Zipf-Mandelbrot 'law' in terms of the spectrum which might be expected.



### 4.2. Rouault

As mentioned above, the law (103) is the only limiting expression for the relative expected spectrum when sampling from a general population with a fixed structural probability distribution. It can be shown (Rouault 1978; Khmaladze and Chitashvili 1989) that the necessary and sufficient condition on the structural distribution $G(p)$ when the limit (103) exists is the property of tails

$$G(p) = p^{-\alpha}\mathfrak{L}(p) \tag{104}$$

with some $0 < \alpha < 1$, and some at $p = 0$ slowly increasing function $\mathfrak{L}$, i.e.

$$\frac{\mathfrak{L}(cp)}{\mathfrak{L}(p)} \to 1, \; p \to 0$$

for each $c > 0$, as e.g. in the case that (102) takes place. In Khmaladze and Chitashvili (1989) it is shown that condition (104) is even necessary to have positive limits

$$\lim_{N\to\infty} \alpha_N(m) > 0, \quad r = 1,2,\ldots .$$

The aim of the mathematical models to be considered here is to suggest some natural stochastic scheme which provides the general population with property (104). The most complete is the markovian model of word generation considered by Rouault (1978), who generalized Miller's stochastic scheme. Let

$$\mathfrak{L} = \{L_0, L_1, L_2, \ldots\}$$

be the set of elements (letters) including the blank space $L_0$ which occur according to some transition probability $p_{i,j}$. A particular word $A$ can be viewed as a (finite) sequence of letters limited by two blanks:

$$A = [L_0 L_{i_1} L_{i_2}, \ldots, L_{i_t} L_0] .$$

Such a word has probability

$$p(A) = p_{0,i_1} p_{i_1,i_2} p_{i_2,i_3} \ldots p_{i_t,0} .$$

To form an idea of the structure of token probabilities $p_n$, $1 \le n \le N$ over a sample of size $N$, let $x_t$ be the Markov chain realization of the procedure generating the running text as a sequence of letters. Let

$\tau_1, \tau_2, \ldots, \tau_N$

be the successive moments when blanks occur in this sample. We can present the sample of token words and the corresponding token probabilities as

$$w_1 = (x_1, x_2, \ldots, x_{\tau_1})$$

$$p(w_1) = \prod_{t=1}^{\tau_1 - 1} p(x_t, x_{t+1})$$

$$w_2 = (x_{\tau_1}, x_{\tau_1+1}, \ldots, x_{\tau_2})$$

$$p(w_2) = \prod_{t=\tau_1}^{\tau_2 - 1} p(x_t, x_{t+1}).$$

Rouault (1978) shows that the collection of such probabilities over all words (of any length) possesses (through its structural distribution) the property (104). Note that Miller's (1957) model is a special case of Rouault's scheme with $p_{0i}$ and $p_{ij}$ not depending on $i, j$.

Thus, if the dynamics of a population are governed by the simplest multiplication rule, according to which a particular element enters the population or is re-used with a constant probability not depending on the current state of the system, then (103) expresses the only possible equilibrium state distribution. However, the class of equilibrium distributions can be essentially enlarged if more general birth and death stochastic schemes of population dynamics are considered.

### 4.3. Dynamic Models: Simon, Waring-Herdan

We have seen that within the classical scheme of independent and identically distributed observations the majority of Zipfian frequency distribution laws, with as the only exception Rouault's 'law', can be considered as analytical expressions for the relative expected spectrum on finite sample sizes in the asymptotic setting of the triangle scheme. In other words, we are dealing with 'laws' that do not have the property of stability with respect to changing sample sizes. In order to justify these 'laws' as laws expressing the equilibrium (or steady state) property of well organized systems, we therefore need more general stochastic processes as models for the formation of populations with large numbers of different elements. In this section we consider a number of such processes which can be viewed as providing rationales for a number of 'laws' of the Zipfian family.

The dynamic modelling idea becomes very natural if we look on the frequency



dynamics in the classical scheme. In fact, the frequency of some word $A_i$ can be calculated recursively

$$f_n(A_i) = f_{n-1}(A_i) + \boldsymbol{\varepsilon}_N^i, \quad n = 1, 2, \ldots, \tag{105}$$

with the increments $\boldsymbol{\varepsilon}_n^i$ taking only two values (0, 1) independently of the frequency structure

$$\mathcal{F}_n = (f_n(A_i),\ 1 \le i)$$

of the sample at the current stage $n$. In other words, the conditional probability coincides with the unconditional one, and is constant:

$$\boldsymbol{Pr}(\boldsymbol{\varepsilon}_n^i = 1 | \mathcal{F}_n) = \boldsymbol{Pr}(\boldsymbol{\varepsilon}_n^i = 1) = \boldsymbol{Pr}(A_i). \tag{106}$$

The specifics of the kind of dynamic modelling considered here lies in the assumption that this probability may depend on the current state of the 'system', and in particular on the frequency of the element $A_i$:

$$\boldsymbol{Pr}(\boldsymbol{\varepsilon}_n^i = 1 | \mathcal{F}_n) = \boldsymbol{Pr}(A_i | \mathcal{F}_n). \tag{107}$$

One of the versions of Simon's (1955, 1960) models can be presented as the simplest (but nevertheless very natural) example. The conditional probability of the 'birth' of the element $A_i$ is defined as

$$\boldsymbol{Pr}(A_i | \mathcal{F}_n) = q \mathbf{I}_{[f_n(A_i) = 0]} \frac{p(A_i)}{\sum_j \mathbf{I}_{[f_n(A_j) = 0]} p(A_j)} + (1 - q) \mathbf{I}_{[f_n(A_i) > 0]} \frac{f_n(A_i)}{n}, \tag{108}$$

where $0 \le q \le 1$ stands for the probability that some new element from the vocabulary with the frequency $f_n(A_i) = 0$ can be included into population, and where with probability $1 - q$ some already present element can be re-used proportional to its frequency in the sample. From the simple relation

$$\mathbf{I}_{[f_{n+1}(A_i) = m]} = \mathbf{I}_{[f_n(A_i) = m]} (1 - \mathbf{I}_{[\boldsymbol{\varepsilon}_n^i = 1]}) + \mathbf{I}_{[f_n(A_i) = m-1]} \mathbf{I}_{[\boldsymbol{\varepsilon}_n^i = 1]} \tag{109}$$

the recursive relations for the spectrum terms can be derived,

$$\hat{V}_{n+1}(m) = \hat{V}_n(m) + \hat{V}_n(m-1)\frac{(1-q)(m-1)}{n} - \hat{V}_n(m)\frac{(1-q)m}{n} + \varepsilon_n(m), \quad (110)$$

where the additive term $\varepsilon_n(m)$ is conditionally centered

$$E(\varepsilon_n(m)|\mathcal{F}_n) = 0. \quad (111)$$

For the expected spectrum we have the recursive equations

$$V_{n+1}(m) = V_n(m) + V_n(m-1)\frac{(1-q)(m-1)}{n} - V_n(m)\frac{(1-q)m}{n} \quad (112)$$

$$V_{n+1}(1) = V_n(1) + 1 - V_n(1)\frac{1-q}{n} \quad (113)$$

$$V_{n+1} = V_n + q. \quad (114)$$

Now recursive equations for the relative expected spectrum can be obtained

$$\alpha_{n+1}(m) = \alpha_n(m) + \frac{1}{n}[(1-q)(m-1)\alpha_n(m-1) - ((1-q)m + 1)\alpha_n(m)] \quad (115)$$

which tells us that the limiting law

$$\alpha_n(m) \to \alpha(m), \; n \to \infty \quad (116)$$

- in fact it can be shown that this convergence and moreover the law of large numbers

$$\frac{\hat{V}_n(m)}{\hat{V}_n} \to \alpha(m), \; n \to \infty \quad (117)$$

takes place here - should be the solution of the equilibrium or steady state equation

$$(1-q)(m-1)\alpha(m-1) = ((1-q)m + 1)\alpha(m). \quad (118)$$

The solution to (118),

$$\alpha(m) = \frac{\Gamma(1 + \frac{1}{1-q})}{1-q} \cdot \frac{\Gamma(m)}{\Gamma(m + \frac{1}{1-q} + 1)}, \quad (119)$$

is a particular instantiation of Simon's (1960) model.



For a slightly more general model where an element with frequency $m$, $m > 0$ can be re-used proportionally to its relative frequency but in which the probability that a new word occurs on the n-th stage is proportional to the total probability of unused words, we have

$$Pr(A_i|\mathcal{F}_n) = \frac{rp(A_i)\mathbf{I}_{[f_n(A_i)=0]} + \frac{f_n(A_i)}{n}\mathbf{I}_{[f_n(A_i)>0]}}{r\sum_j p(A_j)\mathbf{I}_{[f_n(A_j)=0]} + 1}, \quad (120)$$

Note that the probability of generating new words will now eventually decrease to 0. If the structural distribution

$$G(p) = \sum_{i \geq 1} \mathbf{I}_{[p(A_i) \geq p]}$$

of the general population satisfies the condition (104), then the steady state law is

$$\alpha(m) = \frac{\Gamma(m)\Gamma(1 + \alpha r)}{\Gamma(m + 1 + \alpha r)}, \quad (121)$$

the beta function of Yule (1924).

Both the Yule distribution (119) and what we have called the Yule-Simon 'law'

$$\alpha(m) = \frac{\beta}{(m + \beta - 1)(m + \beta)},$$

which we have found to be the more useful expression for the analysis of texts, are special cases of the 'law' advanced by Simon (1960:69) on the basis of a birth and death process model for the population dynamics,

$$\alpha(m) = A\lambda^m B(m + c, d - c + 1),$$

with $B(.,.)$ the Beta-function and with parameters $c$, $d$, $\lambda$ defining the birth and death probabilities and with normalizing constant $A$. For specially chosen parameters and $\lambda$ fixed at unity both models can be derived. Interestingly, the Waring-Herdan-Muller model, which includes the two above versions of Simon's model, can be obtained along similar lines when the probability of re-using some word is a linear function of the frequency of that word (see Khmaladze and Chitashvili, 1989).

These dynamic (birth and death) equilibrium models are satisfactory for the LNRE analysis of biological (distribution of species), social (distribution of incomes), psychological (distribution of responses to some stimulus), and a number of other living and technical systems. Unfortunately, lexical samples generally do not show the tendency to equilibrium state, and though the LNRE features may be displaid clearly, the frequency distributions change considerably with changing sample size.[1] This makes it natural to view such samples as being located in the LNRE zone and to apply models for general populations realizing appropriate frequency distributions on finite samples. In the following sections, we discuss three such models.

### 4.4. The Lognormal Model (Herdan, Carroll)

In this model a word token is chosen on each stage of the experiment as a result of some recursive procedure. Unlike the Markovian scheme underlying Rouault's 'law', this procedure is interpreted as a choice between words rather than as a word generation procedure. The token probabilities are now expressed as products of transition probabilities which are themselves random.

In the lognormal model as considered by Carroll (1967, 1969) the structure of the vocabulary is described in terms of a decision tree. (More general schemes can be considered than that proposed by Carroll, here we will limit ourselves to Carroll's approach.) Assume that we have a binary decision tree where the paths leading to the leaves of the tree, the elements of the vocabulary, may have different lengths. Let $\vec{y} = y_s$, $(s = 0, 1, 2, ...)$ denote some path from the root of the tree to some leaf, i.e. the sequence of decisions made at the different levels of the tree, with $y_s \in \{0, 1\}$ indicating the possible decisions on each stage. For each path $\vec{y}$ we define some stopping moment $\tau(\vec{y})$ indicating the length of the path. Each path uniquely determines some word $A(y)$ as a result of the decision procedure.

To define probabilities of words let $(\vec{\pi} = \pi_s, s = 0,1,2,...)$ be decision probabilities corresponding to the stages $s = 0, 1, \ldots$ . Also assume that these probabilities are randomly distributed according to some distribution function $\phi(\pi)$, $(0 \le \pi \le 1)$ on the interval [0, 1]. Suppose finally that $y_s$ as well as $\pi_s$ are independent. Then the probability of word $w = w(\vec{y})$, given the probabilities $\vec{\pi}$ equals

[1] For a dynamic model which combines the Mandelbrot/Miller/Rouault approach with that of Simon without imposing the equilibrium constraint, the reader is referred to Baayen (1991), a simulation study that focusses on the similarity relations between words in the lexicon.



$$p(w) = \prod_{s=0}^{\tau_y} (\pi_s)^{y_s}(1 - \pi_s)^{1 - y_s}. \tag{122}$$

The token probability distribution can now be expressed as

$$F(p) = \boldsymbol{Pr}(p(w) \ge p) = \boldsymbol{Pr}\left(\sum_{s=0}^{\tau_y} [y_s \log(\pi_s) + (1 - y_s)\log(1 - \pi_s)] \ge \log p\right)$$

$$= \boldsymbol{Pr}\left(\sum_{s=0}^{\tau_y} [\hat{\theta}(s) \ge \log p\right). \tag{123}$$

The mean and the variance of the random variable $\hat{\theta}$ are easily calculated:

$$\theta = \boldsymbol{E}\hat{\theta} = \boldsymbol{E}[\pi \log \pi + (1 - \pi)\log(1 - \pi)]$$

$$= \int_0^1 [x \log x + (1 - x)\log(1 - x)]\, d\Phi(x) \tag{124}$$

$$\sigma_\theta^2 = \boldsymbol{VAR}\hat{\theta} = \int_0^1 [x(\log x)^2 + (1 - x)(\log(1 - x))^2]\, d\Phi(x) - \theta^2 \tag{125}$$

We need some conditions on the stopping moment $\tau$ and on the variance of the distribution $\Phi$. Suppose that (i) $\tau$ is a Markov moment. This is a conventional assumption in the theory of stochastic processes when sums of random numbers of random variables are investigated. In the present case, the assumption that $\tau$ is a Markov moment implies that for any path $\vec{y}$ the event $[\tau = k]$, given the path $(y_s, 0 \le s \le k)$, does not depend on the future values $(y_s, k + 1 \le s \le N)$. This condition is met when, for instance, $\tau$ does not depend on the path $\vec{y}$ at all, as is assumed in Carroll (1969). Also suppose that (ii) the expected path-length, the time needed to come to one of the leaves of the decision tree, is sufficiently large $(E\tau \gg 1)$ and that the variance is relatively small, as in the case of $\tau$ having a Poisson distribution:

$$\frac{\boldsymbol{VAR}(\tau)}{\boldsymbol{E}\tau^2} \ll 1.$$

Finally, suppose that

$$VAR\hat{\theta} \rightarrow \frac{VAR(\tau)}{E\tau}.$$

Under these assumptions the token probability distribution indexed by the parameter $Z$ can asymptotically be expressed as

$$F(p) = F^Z(p) \rightarrow \boldsymbol{Pr}(\sigma N^Z - \log(Z)) \geq \log(p)$$

where we introduce the notation $Z = e^{-\theta E\tau}$, and where $N^Z$ is an asymptotically standard gaussian variable. Equivalently,

$$\begin{aligned} F(p) &= F^Z(p) \\ &= \frac{1}{\sqrt{2\pi}\,\sigma}\int_p^\infty \frac{1}{x} e^{-\frac{(\log x - \log Z)^2}{2\sigma^2}} dx \\ &= \frac{1}{\sqrt{2\pi}\,\sigma}\int_{\frac{p}{Z}}^\infty \frac{1}{x} e^{-\frac{(\log x)^2}{2\sigma^2}} dx. \end{aligned} \tag{126}$$

Thus the lognormal model is justified asymptotically in the framework of the triangle scheme discussed above.

### 4.5. The Generalized Inverse Gauss-Poisson 'Law'

The motivation for the generalized inverse Gaussian-Poisson distribution (Sichel, 1986), which is presented by the structural distribution

$$G(p) = \frac{(2/bc)^\gamma}{2K_\gamma(b)}\int_\lambda^\infty x^{\gamma-1}\exp\left(-\frac{x}{c} - \frac{\sigma^2 c}{4x}\right)dx, \tag{127}$$

seems formal, though as special cases it includes e.g. the $\Gamma$-distribution ($b = 0$) and the distribution of an inverse of a Gaussian random variable ($c \rightarrow \infty$, $\sigma^2 \rightarrow 0$, $\sigma^2 c = \text{const}$).

### 4.6. The Generalized Zipf's 'Law'

The basis for the rationale of the generalized Zipf 's model, presented by the structural distribution



$$G(p) = C\int_0^\infty e^{-Zpx}\frac{(\log(1+x))^{\gamma-1}x^{\alpha-1}}{(1+x)^{\beta+1}}dx,$$

is clear enough: this is the unique parametric family of structural distributions which can realize on a finite sample of a particular size Z (Z being one of its parameters) a desired representative of the Zipfian family of 'laws' in terms of the relative expected spectrum.

## 5. Statistical Analysis with LNRE Models

In this section the information needed for the application of parametric LNRE models to statistical data analysis is presented. In section 5.1 we discuss the expressions for the various theoretical characteristics in which we are interested. We present some expressions for covariances in section 5.2. Section 5.3 outlines a number of ways in which the parameters of theoretical models may be estimated. Section 5.4 briefly discusses how to estimate confidence regions for estimated parameters. Goodness-of-fit tests for theoretical models are given in section 5.5. Section 5.6 contains some suggestions how to compare LNRE samples. Finally, the software known to us for the modelling of LNRE distributions is discussed in section 5.7 and applied to a number of empirical distributions.

To make these sections as independent of the other parts as possible and thus more convenient for application, we give some expressions in detail even though they can be found in previous sections. For ease of presentation, we will phrase the discussion in terms of a general three-parameter model with the structural distributions

$$\begin{aligned} G(p) &= G(p;\ \alpha,\ \beta,\ \gamma) \\ F(p) &= F(p;\ \alpha,\ \beta,\ \gamma) \\ \phi(p) &= \phi(p;\ \alpha,\ \beta,\ \gamma), \end{aligned} \tag{128}$$

where $\phi(p)$ is the density function of the token probability distribution $F(p)$,

$$\phi(p) = \frac{d}{dp}F(p).$$

## 5.1. Expressions for the Theoretical Spectrum

### 5.1.1. Nonparametric Expressions

General expressions for the expected frequency spectrum and the expected empirical vocabulary for three parameter models can be presented in integral form:

$$\begin{aligned} V_N(m) &= V_N(m;\ \alpha,\ \beta,\ \gamma) = \boldsymbol{E}\hat{V}_N(m) \\ &= \int_0^\infty \frac{(pN)^m}{m!} e^{-pN}\, dG(p) \\ &= \int_0^\infty \frac{(pN)^m}{m!} e^{-pN} \frac{1}{p}\, dF(p) \\ &= \int_0^\infty \frac{(pN)^m}{m!} e^{-pN} \frac{1}{p} \phi(p)\, dp. \end{aligned} \tag{129}$$

for the expected frequency spectrum, and

$$\begin{aligned} V_N(0) &= V_N(0;\ \alpha,\ \beta,\ \gamma) = \boldsymbol{E}\hat{V}_N \\ &= \int_0^\infty (1 - e^{-pN})\, dG(p) \\ &= \int_0^\infty (1 - e^{-pN}) \frac{1}{p}\, dF(p) \\ &= \int_0^\infty (1 - e^{-pN}) \frac{1}{p} \phi(p)\, dp. \end{aligned} \tag{130}$$

for the expected empirical vocabulary. Note that for notational convenience the expected empirical vocabulary is denoted by $V_N(0)$. Thus the vector $(V_N^{(m)},\ m = 0, 1, 2, \ldots, M)$ denotes the first $M$ elements of the theoretical frequency spectrum and the expected empirical vocabulary jointly. The same holds for its empirical analogue, $(\hat{V}_N^{(m)},\ m = 0, 1, 2, \ldots, M)$.



### 5.1.2. Parametric Expressions

We now present explicit expressions for the three parametric families of structural distribution models, the lognormal model, the inverse generalized Gauss-Poisson model, and the generalized Zipf model.

**The lognormal model**. Carroll's (1967, 1969) lognormal model is defined by the structural token probability distribution

$$F(p) = \frac{1}{\sigma\sqrt{2\pi}} \int_p^\infty \frac{1}{x} e^{-\frac{1}{2}\left(\frac{\log(x) - \mu}{\sigma}\right)^2} dx. \tag{131}$$

The expected spectrum and vocabulary for the sample size $N$ can be expressed in integral form:

$$\begin{aligned} V_N(m) &= \boldsymbol{E}\hat{V}_N(m) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_0^\infty \frac{(xN)^m}{x^2 m!} e^{-xN - \frac{1}{2}\left(\frac{\log(x) - \mu}{\sigma}\right)^2} dx \end{aligned} \tag{132}$$

$$V_N = \boldsymbol{E}\hat{V}_N \tag{133}$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_0^\infty (1 - e^{-xN}) \frac{1}{x^2} e^{-\frac{1}{2}\left(\frac{\log(x) - \mu}{\sigma}\right)^2} dx. \tag{134}$$

The theoretical vocabulary (number of types) is obtained by considering $V_N$ in the limit for $N \to \infty$:

$$V = \lim_{N \to \infty} V_N = e^{\frac{\sigma^2}{2} - \mu}.$$

Note that the parameter $\mu$, the mean value of $\log p$ in the general population, typically is a negative number < -1.

**The generalized inverse Gauss-Poisson model**. Sichel's (1975, 1986) generalized inverse Gauss-Poisson 'law' is based on the structural type distribution

$$G(p) = V \frac{(2/bc)^\gamma}{2K_\gamma(b)} \int_p^\infty x^{\gamma - 1} e^{\left(-\frac{x}{c} - \frac{b^2 c}{4x}\right)} dx\ , \tag{135}$$

where $K_\gamma(b)$ is the modified Bessel function of the second kind of order $\gamma$ and argument $b$. The theoretical vocabulary $V$, the number of types in the population, can be determined on the basis of the normalizing argument. In fact, since

$$\int_0^\infty dF(p) = \int_0^\infty p dG(p) = 1,$$

we can easily find the expression for $V$:

$$V = \frac{2}{bc}\frac{K_\gamma(b)}{K_{\gamma+1}(b)}. \tag{136}$$

For the expected vocabulary and the (relative) expected spectrum for arbitrary sample size $N$ explicit formulas in terms of the Bessel function can be found:

$$V_N(0) = \frac{2}{bc}\frac{K_\gamma(b)}{K_{\gamma+1}(b)}\left[1 - \frac{K_\gamma(b\sqrt{1+cN})}{(1+cN)^{\gamma/2}K_\gamma(b)}\right].$$

$$V_N(m) = \frac{V_N(0)}{(1-\theta_N)^{-\gamma/2}K_\gamma(\alpha_N(1-\theta_N)^{1/2}) - K_\gamma(\alpha_N)}\frac{(0.5\alpha_N\theta_N)^m}{m!}K_{\gamma+m}(\alpha_N), \tag{137}$$

$$\alpha_N(m) = \frac{1}{(1-\theta_N)^{-\gamma/2}K_\gamma(\alpha_N(1-\theta_N)^{1/2}) - K_\gamma(\alpha_N)}\frac{(0.5\alpha_N\theta_N)^m}{m!}K_{\gamma+m}(\alpha_N),$$

where the parameters $\alpha_N = b(1+cN)^{1/2}$ and $\theta_N = cN/(1+cN)$ are introduced for notational simplicity. Note that the parameters $\alpha_N$ and $\theta_N$ are functions of the sample size $N$, while the parameters $b$, $c$ and $\gamma$ are population invariants.

**The extended generalized Zipf's 'law'**. Orlov and Chitashvili (1982a,b, 1983 a,b) develop a model that is a generalization of Zipf's law. For this model the structural probability type distribution

$$G(p) = C\int_p^\infty e^{-Zpx}\frac{(\ln(1+x))^{\gamma-1}x^{\alpha-1}}{(1+x)^{\beta+1}}dx,$$

where $C$ is a normalizing coefficient (defined below), is characterized by the property that it realizes on the sample size $Z$ the relative expected spectrum



$$\alpha_Z(m) = \frac{\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1}y^\alpha}{(1+y)^{m+1}(1+y)^\beta}dy}{\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1}y^{\alpha-1}}{(1+y)^{\beta+1}}dy}.$$

A number of known 'laws' are included as special cases. The following expressions for the expected vocabulary $V_N$ and frequency spectrum terms $V_N(m)$ can be obtained:

$$V_N(m) = \boldsymbol{E}\hat{V}_N(m) \tag{138}$$

$$= C(Z, \alpha, \beta, \gamma)t^m\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1}y^\alpha}{(t+y)^{m+1}(1+y)^{\beta+1}}dy \tag{139}$$

$$V_N = \boldsymbol{E}\hat{V}_N \tag{140}$$

$$= C(Z, \alpha, \beta, \gamma)t\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1}y^{\alpha-1}}{(t+y)(1+y)^\beta}dy \tag{141}$$

where $t = N/Z$ and where the coefficient $C$ is defined by

$$C(Z, \alpha, \beta, \gamma) = \frac{V_Z}{\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1}y^{\alpha-1}}{(1+y)^{\beta+1}}dy}. \tag{142}$$

The expected number of types $V_Z$ for the sample size $Z$ can be determined by the normalizing argument, namely from the relation

$$N = \sum_{m=1}^{Np_N\{1\}} m\hat{V}_N(m),$$

where $p_N\{1\}$ is the maximal observed relative frequency. Application of this relation to the expected frequency spectrum for $Z = N$ leads to the following expression

$$V_Z = Z\frac{\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1}y^{\alpha-1}}{(1+y)^{\beta+1}}dy}{\int_0^\infty \frac{[\ln(1+y)]^{\gamma-1}y^{\alpha-2}}{(1+y)^{\beta+Z\hat{p}_N\{1\}}}[(1+y)^{Z\hat{p}_N\{1\}} - 1 - \frac{Z\hat{p}_N\{1\}y}{1+y}]dy}. \quad (143)$$

For the important case of the extended Waring-Herdan-Muller law ($\gamma = 1$), all formulas are significantly simplified:

$$V_N(m) = C(Z, \alpha, \beta)t^m \int_0^\infty \frac{y^\alpha}{(t+y)^{m+1}(1+y)^{\beta+1}}dy \quad (144)$$

$$V_N = C(Z, \alpha, \beta)t \int_0^\infty \frac{y^{\alpha-1}}{(t+y)(1+y)^\beta}dy \quad (145)$$

with

$$C(Z, \alpha, \beta) = \frac{V_Z}{\int_0^\infty \frac{y^{\alpha-1}}{(1+y)^{\beta+1}}dy}.$$

The theoretical vocabulary $V$ is finite if $\beta > \alpha$ and can be expressed as

$$V = \frac{V_Z\beta}{\beta - \alpha}.$$

If, furthermore, $\alpha = 1$ (the extended Yule-Simon law), then the expression for $V_Z$ can be approximated by

$$V_Z \approx \frac{Z}{\beta \ln(Z\hat{p}_N\{1\})}. \quad (146)$$

### 5.2. Expressions for Covariances

The covariances between the terms $(\hat{V}_N, \hat{V}_N(m), m = 1, 2, ...)$, i.e. the autocovariances and crosscovariances for varying sample sizes, can be presented in terms of the expected values of the spectrum terms, as shown in sections 3.1 and 3.3.2. When convenient, they can be stated in integral form, applying the parametric and non-parametric representations discussed above.



Given the vector statistic $(\hat{V}_N^{(m)} = 0, 1, 2, ..., M)$ and the expressions for $V_N(m)$, $m = 0, 1, 2, ...$, the corresponding covariance matrix $R_{m,k}(N, M)$ is easily calculated:

$$R_{m,k}(N, M) = (\mathbf{COV}(\hat{V}_N(m), \hat{V}_N(k)))_{k,m=0,1,...,M} =$$

$$= \begin{cases} \delta_{m,k}V_N(m) - \binom{m+k}{m}\frac{1}{2^{m+k}}V_{2N}(m+k) & \text{for } m, k = 1,2,...,M. \\ -\frac{1}{2^m}V_{2N}(m) & \text{for } m = 0,\ k = 1,2,...,M. \\ V_N - V_{2N} & \text{for } m = 0,\ k = 0. \end{cases} \quad (147)$$

For two different samples with size $N$ and $n$, $N \le n$ we have

$$\mathbf{COV}(\hat{V}_n(m), \hat{V}_N(k)) = V_n(m)\binom{m}{k}\left(\frac{N}{n}\right)^k\left(1 - \frac{N}{n}\right)^{m-k} - V_{N+n}(m+k)\binom{m+k}{m}\left(\frac{N}{N+n}\right)^k\left(1 - \frac{N}{N+n}\right)^m$$

$$\mathbf{COV}(\hat{V}_n, \hat{V}_N(k)) = \left(\frac{N}{n+N}\right)^k V_{n+N}(k)$$

$$\mathbf{COV}(\hat{V}_n(m), \hat{V}_N) = \left(\frac{n}{n+N}\right)^m V_{n+N}(m) - \left(1 - \frac{N}{n}\right)^m V_n(m)$$

$$\mathbf{COV}(\hat{V}_n, \hat{V}_N) = V_{n+N} - V_{\min(N,n)}.$$

In section 3.1.4 we considered the interpolation problem. The results obtained there can be generalized, so that for arbitrary $n$, $N$ the recursive relations

$$V_n(m) = \sum_{j \ge m} V_N(j)\binom{j}{m}\left(\frac{n}{N}\right)^m\left(1 - \frac{n}{N}\right)^{j-m} \quad (148)$$

$$V_n = \sum_{j \ge 1} V_N(j)\left(1 - \left(1 - \frac{n}{N}\right)^j\right) \quad (149)$$

between the expected spectrum terms can be defined (Good and Toulmin 1956; Kalinin 1965). The autocovariance of $\hat{V}_{N,n}$ equals

$$\boldsymbol{COV}(\hat{V}_{N,n}, \hat{V}_{N,k}) = V_{n+k} - V_{n+k-\frac{nk}{N}}, \; 1 \le n, k \le N. \tag{150}$$

The mean square deviation of $\hat{V}_{N,n}$ from the "true" value of the vocabulary on the sample of size $n$ (the interpolation accuracy) can be presented as

$$\boldsymbol{E}(\hat{V}_{N,n} - \hat{V}_n)^2 = V_{2n-\frac{n^2}{N}} - V_N \, . \tag{151}$$

Expected spectrum elements for sample sizes $N' > N$ are often required in the formulas for variances and covariances. Unfortunately, the nonparametric expressions (148) and (149) become unstable for $n > 2N$ (see e.g. Good and Toulmin 1956), even though for instance (149) still possesses some optimality property: it gives the best linear extrapolation whereas the optimal extrapolation formula

$$\hat{V}_{N,n} = \boldsymbol{E}(\hat{V}_n | \hat{V}_N(k), \; k \ge 1)$$

is strictly nonlinear for $n \ge N$ and rather complicated for an exact calculation. Perhaps the best way to proceed is to use the simple extrapolation formulas based on some parametric model and to substitute the estimated parameters $(\hat{\alpha}_N, \hat{\beta}_N, \hat{\gamma}_N)$ for their theoretical counterparts $(\alpha, \beta, \gamma)$. In the case of extrapolated vocabulary sizes,

$$\hat{V}_{n,N} = V_n(\hat{\alpha}_N, \hat{\beta}_N, \hat{\gamma}_N) \, ,$$

the accuracy of the predicted values can be gauged by considering

$$D_{n,N} = \boldsymbol{E}(\hat{V}_n - \hat{V}_{n,N})^2 \, ,$$

where $\hat{V}_{n,N}$ can be approximated for sufficiently large $N$ by

$$\hat{V}_{n,N} \approx V_n + (\hat{\alpha}_N - \alpha)\dot{V}_n^1 + (\hat{\beta}_N - \beta)\dot{V}_n^2 + (\hat{\gamma}_N - \gamma)\dot{V}_N^3, \tag{152}$$

where



$$V_n = V_n(\alpha, \beta, \gamma)$$

$$\dot{V}_n^1 = \frac{\partial}{\partial\alpha} V_n(\alpha, \beta, \gamma)$$

$$\dot{V}_n^2 = \frac{\partial}{\partial\beta} V_n(\alpha, \beta, \gamma)$$

$$\dot{V}_n^3 = \frac{\partial}{\partial\gamma} V_n(\alpha, \beta, \gamma) .$$

Note that to use this accuracy expression, we must again replace the parameters $(\alpha, \beta, \gamma)$ in the right hand side of $D_{n,N}$ by their estimators $(\hat{\alpha}_N, \hat{\beta}_N, \hat{\gamma}_N)$.

### 5.3. Parameter Estimation

Several procedures can be suggested for estimating the parameters of a word frequency 'law'.

#### 5.3.1. Method 1

The simplest way is to require that the first (three) 'most remarkable' terms of the frequency spectrum, that is, the vector $(\hat{V}_N(m), \; m = 0, 1, 2)$, should coincide with their expected values:

$$\left.\begin{aligned} \hat{V}_N(0) &= V_N(0; \alpha, \beta, \gamma) \\ \hat{V}_N(1) &= V_N(1; \alpha, \beta, \gamma) \\ \hat{V}_N(2) &= V_N(2; \alpha, \beta, \gamma) \end{aligned}\right\} \Rightarrow (\hat{\alpha}_N, \hat{\beta}_N, \hat{\gamma}_N), \tag{153}$$

where we denoted the resulting parameter estimators by

$$(\hat{\alpha}_N, \hat{\beta}_N, \hat{\gamma}_N).$$

Note that the number of equations equals the number of parameters.

#### 5.3.2. Method 2

A more global, though rather complicated algorithm can be used which takes more terms of the spectrum into consideration. We fix some number $M \ge 3$ of terms of the vector

$(\hat{V}_N^{(m)},\ m = 0,1,2,...,M)$

and construct the chi-square statistic

$$\chi^2_{(M-3)} = \sum_{0\le m,k\le M} (\hat{V}_N(m) - V_N(m))R^{-1}_{m,k}(N, M)(\hat{V}_N(k) - V_N(k)) \qquad (154)$$

where $R^{-1}_{m,k}(N,M)$, $0 \le m, k \le M$ is the inverse of $R_{m,k}(N,M)$. We then search for the estimators

$(\alpha^*_N, \beta^*_N, \gamma^*_N)$

for which $\chi^2_{(M-3)}$ is minimal.

### 5.3.3. Method 3

A method that we have found to be especially useful is to fix one parameter, say $\gamma$, and to choose the other two parameters such that

$$\left.\begin{aligned}\hat{V}_N(0) &= V_N(0;\ \alpha,\ \beta,\ \gamma)\\ \hat{V}_N(1) &= V_N(1;\ \alpha,\ \beta,\ \gamma)\end{aligned}\right\} \qquad (155)$$

is satisfied. Following this, $\gamma$ is varied (and $\alpha$ and $\beta$ adjusted to satisfy (155) such that the value of $\chi^2_{(M-3)}$ is minimal.

### 5.3.4. Method 4

As a modification of method 2, the estimator

$(\alpha^{**}_N, \beta^{**}_N, \gamma^{**}_N)$

can be constructed so as to minimize the chi-square statistic for the differences between the (nonparametric) interpolated vocabulary growth curve

$$\hat{V}_{N,n} = \sum_{m\ge 1} \hat{V}_N(m)\left(1 - (1 - \frac{n}{N})^m\right)$$



and its expectation with respect to the parametric model. Thus the estimators

$(\alpha^{**}_N, \beta^{**}_N, \gamma^{**}_N)$

are chosen such that

$$\tilde{\chi}^2_{(M-3)} = \sum_{1\le i,j\le M} (\hat{V}_{n_i N} - V_{n_i})\tilde{R}^{-1}_{n_i n_j}(N,M)(\hat{V}_{n_j N} - V_{n_j}) \qquad (156)$$

is minimal, where $\tilde{R}^{-1}_{n_i n_j}(N, M)$, $1 \le n_i, n_j \le N$ is the inverse of the covariance matrix

$$\tilde{R}_{n_i n_j}(N,M) = \boldsymbol{COV}(\hat{V}_{n_i N}, \hat{V}_{n_j N}) = V_{n_i+n_j} - V_{n_i+n_j-\frac{n_i n_j}{N}},\ 1 \le n_i, n_j \le N.$$

### 5.4. Confidence Intervals

For completeness, we briefly discuss how confidence regions for the estimators can be constructed. To do so, we need the matrix $\dot{V}(M,3) = \dot{V}_m(M,3)_{m=0,1,2,...,M}$ of partial derivatives of the expected spectrum with respect to the parameters:

$$\dot{V}(M,\ 3) = \begin{Bmatrix} \frac{\partial}{\partial\alpha}V_N(0;\ \alpha,\ \beta,\ \gamma) & \frac{\partial}{\partial\beta}V_N(0;\ \alpha,\ \beta,\ \gamma) & \frac{\partial}{\partial\gamma}V_N(0;\ \alpha,\ \beta,\ \gamma)\\ \frac{\partial}{\partial\alpha}V_N(1;\ \alpha,\ \beta,\ \gamma) & \frac{\partial}{\partial\beta}V_N(1;\ \alpha,\ \beta,\ \gamma) & \frac{\partial}{\partial\gamma}V_N(1;\ \alpha,\ \beta,\ \gamma)\\ \vdots & \vdots & \vdots\\ \frac{\partial}{\partial\alpha}V_N(M;\ \alpha,\ \beta,\ \gamma) & \frac{\partial}{\partial\beta}V_N(M;\ \alpha,\ \beta,\ \gamma) & \frac{\partial}{\partial\gamma}V_N(M;\ \alpha,\ \beta,\ \gamma)\end{Bmatrix}. \qquad (157)$$

Then for the parameter estimators

$(\hat{\alpha}_N, \hat{\beta}_N, \hat{\gamma}_N)$

the normal distribution can be assumed

$$\begin{pmatrix} \hat{\alpha}_N - \alpha \\ \hat{\beta}_N - \beta \\ \hat{\gamma}_N - \gamma \end{pmatrix} \xrightarrow{D} N(0, \hat{C}) , \tag{158}$$

with the covariance matrix

$$\hat{C} = (\hat{C}_{ij})_{1 \le i,j \le 3} = (\dot{V}(3,3))^{-1} R(N,3) (\dot{V}(3,3))^{-1}. \tag{159}$$

For the estimators

$$(\alpha_N^*, \beta_N^*, \gamma_N^*)$$

the normal distribution

$$\begin{pmatrix} \hat{\alpha}_N^* - \alpha \\ \hat{\beta}_N^* - \beta \\ \hat{\gamma}_N^* - \gamma \end{pmatrix} \xrightarrow{D} N(0, \hat{C}^*) , \tag{160}$$

can be used with the covariance matrix

$$\hat{C}^* = (\hat{C}_{ij}^*)_{1 \le i,j \le 3} = [\dot{V}(M,3) R^{-1}(N,M) \dot{V}(M,3)]^{-1}. \tag{161}$$

With $M = 3$ these covariances obviously coincide, but if $M > 3$ then the estimators

$$(\alpha_N^*, \beta_N^*, \gamma_N^*)$$

are characterized by the narrower confidence region.

### 5.5. Goodness-of-fit Test for Models

The minimal values of the $\chi^2$ statistics can be used to test whether the chosen parametric model fits the data. For instance, if estimation method 2 is used, then the minimal value of $\chi^2_{(M-3)}$ obtained when the parameters $(\alpha, \beta, \gamma)$ are substituted by their estimators $(\hat{\alpha}_N^*, \hat{\beta}_N^*, \hat{\gamma}_N^*)$ should be less then the desired significance level of the $\chi^2$ distribution with $M-3$ degrees of freedom.



Note also that some particular parametric model, satisfactory for the first $M$ terms of the spectrum statistics

$$(\hat{V}_N, \hat{V}_N(m), 1 \le m \le M - 1)$$

may not be acceptable in a global sense for the whole vector

$$(\hat{V}_N(m), 1 \le m).$$

### 5.6. Comparing Samples

Two samples can be compared to establish the identity of the (theoretical) probability distributions of the corresponding general populations, for instance for the purpose of authorship determination.

Let two samples of sizes $N^1$ and $N^2$ be given with, generally speaking, different vocabularies, as in the case that texts written in different languages are compared:

$$\underline{V}^i = (A_1^i, A_2^i, \ldots, A_V^i), \; i = 1,2.$$

The corresponding frequencies, rank frequency distributions and frequency spectra are, for $i = 1,2$:

$$f^i_{N^i}(A_1^i), f^i_{N^i}(A_2^i), f^i_{N^i}(A_{V^i})$$

$$f^i_{N^i}(A_1^i) \ge f^i_{N^i}(A_2^i) \ge \ldots \ge f^i_{N^i}(A_{V^i}),$$

$$\hat{V}^i_{N^i}(m) = \sum_{j \ge 1} \mathbf{I}_{[f(A_j^i) = m]}, \; m = 1, 2, \ldots$$

$$\hat{V}^i_{N^i} = \sum_{m \ge 1} \hat{V}^i_{N^i}(m).$$

We must distinguish several ways in which the comparison problem can be stated in terms of the corresponding theoretical models expressed in the form of probability distributions

$$(P^i(A_j^i), 1 \le j \le V), \; i = 1,2$$

or structural probability distributions

$$G^i(p) = \sum_{j=1}^{V^i} \mathbf{I}_{[p^i(A_j^i) \geq p]}, \; p \geq 0, \; i = 1,2.$$

First consider the case for which the vocabularies are identical,

$$A_j^1 = A_j^2 = A_j, \; j = 1,2,\ldots ,$$

the two texts being written in one and the same language. It is natural to construct this comparison problem in terms of the hypothesis that the (individual) probabilities coincide:

$$\boldsymbol{P}^1(A_j) = \boldsymbol{P}^2(A_j), \; 1 \leq j \leq V.$$

Since high and low frequencies can be considered as independent for LNRE samples (see section 3.3.3), we can focus on the left hand (high probabilities) or on the right hand (low probabilities) tails of a rank probability distribution. We can apply e.g. the standard $\chi^2$ test to check the coincidence of high probabilities. The testing of the right hand tails is quite nontrivial, however.

To do this, consider the united sample of size $N = N^1 + N^2$ with frequencies

$$f_N(A_1) = f_{N^1}^1(A_1) + f_{N^2}^2(A_2),$$

$$f_N(A_2) = f_{N^1}^1(A_2) + f_{N^2}^2(A_2), \; \ldots ,$$

$$f_N(A_{\hat{V}}) = f_{N^1}^1(A_{\hat{V}}) + f_{N^2}^2(A_{\hat{V}}).$$

Introduce the joint frequency spectrum

$$\hat{V}_N(m, k, l) = \sum_{j \geq 1} \mathbf{I}_{[f_N(A_j) = m, \, f_{N^1}^1(A_j) = k, \, f_{N^2}^2(A_j) = l]}, \; m = 1, 2, \ldots, k+l = m,$$

the number of elements which appear $k$ times in the first and $l$ times in the second sample, and let $\hat{V}_N(m)$ be a frequency spectrum on the united sample:

$$\hat{V}_N(m) = \sum_{k+l=m} \hat{V}_N(m, k, l).$$

Applying the scheme of sampling without replacement presented in section 3.1.4, according to which, for large enough $N$, the vector



$$\hat{V}_N(m, k, l), \; 0 \leq k \leq m, \; k + l = m$$

is multinomially distributed given $\hat{V}_N(m)$, the following series of $\chi^2$ statistics can be suggested for the test of comparison:

$$\chi^2(m) = \sum_{k=0}^{m} \frac{(\hat{V}_N(m, k, l) - B(m, k, \frac{N^1}{N})\hat{V}_N(m))^2}{B(m, k, \frac{N^1}{N})\hat{V}_N(m)}$$

in particular,

$$\chi^2(1) = \frac{1}{N^1N^2} \frac{(N^1\hat{V}_N(1, 0, 1) - N^2\hat{V}_N(1, 1, 0))^2}{\hat{V}_N(1, 0, 1) + \hat{V}_N(1, 0, 1)}.$$

For equal sample sizes ($N^1 = N^2$), the distance between two samples measured in terms of the hapaxes, the number of elements that appeared in exactly one of the samples only, is expressed by the ratio

$$\chi^2(1) = \frac{(\hat{V}_N(1, 0, 1) - \hat{V}_N(1, 1, 0))^2}{\hat{V}_N(1, 0, 1) + \hat{V}_N(1, 1, 0)}.$$

By successively checking the admissibility of the values

$$\chi^2(1), \; \chi^2(1) + \chi^2(2), \; \chi^2(1) + \chi^2(2) + \chi^2(3), \ldots$$

with respect to the critical levels of the $\chi^2$-distribution with 1, 3, 6, ... degrees of freedom respectively, we are able to accept with increasing accuracy the hypothesis of coincidence on the tails of the probability distributions studied.

Next consider the case that the samples have been obtained from general populations with different vocabularies (texts written in different languages). It is reasonable to analyse this problem in terms of structural probability distributions. Of course, we can do this even when the vocabularies are the same, in which case we accept the identity of the theoretical models and state that although the individual probabilities may be different, the rank probability distributions coincide.

To check the coincidence of the right hand tails of rank probability distributions we must compare the components of the frequency spectra $\hat{V}_{N^1}^1(m)$ and $\hat{V}_{N^2}^2(m)$, $m \geq 1$. To construct the $\chi^2$ statistic for the difference of the spectrum

terms (even in the case that the sample sizes are the same), we need the covariance matrix $\boldsymbol{COV}(\hat{V}^i_{N^i}(m), \hat{V}^i_{N^i}(k))$, which is itself unknown. We can apply formula (147), which represents this matrix in terms of the expected spectrum components corresponding not only to the sample sizes $N^1$ and $N^2$, but to $2N^1$ and $2N^2$ as well. The natural way to proceed is to use (149) to obtain $V_{2N^i}(m)$, $m \geq 1$, substituting the observed values for the expected ones. Unfortunately, we are again confronted with the problem of the instability of (149) for $n \geq 2N$, apart from the necessity of taking into account the differences between the sample sizes.

To avoid these difficulties, the following construction scheme of a $\chi^2$-distance between the samples can be suggested. First construct the interpolated vocabulary growth curves for both samples $i = 1, 2$:

$$\hat{V}^i_{N^i,n} = \sum_{j\geq 1} \hat{V}^i_{N^i}(j)\left(1 - (1 - \frac{n}{N})^j\right), \quad 1 \leq n \leq N^i. \tag{162}$$

Next construct the estimations $\hat{Q}^i_{N^i}(n, k)$, $1 \leq n, k \leq N^i$, $i = 1,2$ for the covariances

$$\boldsymbol{COV}(\hat{V}^i_{N^i,n}, \hat{V}^i_{N^i,k}), \quad 1 \leq n, k \leq N^i$$

using

$$\hat{Q}^i_{N^i}(n, k) = \hat{V}^i_{N^i,n+k} - \hat{V}^i_{N^i,n+k-\frac{n}{N}}$$

(cf. sections 3.3.2 and 5.2). Finally, fix some sample size values on which the interpolated vocabulary growth curves for two samples should be compared,

$$1 \leq n_1 \leq n_2 \leq n_3 \leq \ldots \leq \min(N^1, N^2) ,$$

and construct the $\chi^2$-distance

$$D^{1,2}(N^1, N^2) = \sum_{j,k} (\hat{V}^1_{N^1,n_j} - \hat{V}^2_{N^2,n_j}) Q^{(-1)}(n_j, n_k) (\hat{V}^1_{N^1,n_k} - \hat{V}^2_{N^2,n_k})$$

where $Q^{(-1)}$ is the inverse matrix of the sum of the matrices

$$\hat{Q}^1_{N^1}(n, k) + \hat{Q}^2_{N^2}(n, k).$$



In other words, the identity of the theoretical structural distributions is checked by the $\chi^2$-distance between the interpolated vocabulary growth, and the above-mentioned difficulty is avoided because the maximal index $n_j + n_k$ is selected to be less than $\min(N^1, N^2)$ in the expression of $\hat{Q}^i_{N^i}(n_j, n_k)$.

Up till now we have focussed on non-parametric tests for the identity of two samples. If some parametric family of (structural) probability distribution is accepted as satisfactory for both samples, then the comparison tests can be improved (become more powerfull) if they are based on the comparison of the estimated parameters. Let

$$(\hat{\alpha}^i_{N^i}, \hat{\beta}^i_{N^i}, \hat{\gamma}^i_{N^i}, \ i = 1,2)$$

be the estimated parameters constructed by one of the estimation schemes discussed above. As the $\chi^2$-distance between the samples we can consider the quadratic form (a $\chi^2$-statistic with three degrees of freedom)

$$(X, AX) = \sum_{i,j=1}^{3} X_i A^{(-1)}_{i,j} X_j$$

where

$$X_1 = \hat{\alpha}^1_{N^1} - \hat{\alpha}^2_{N^2}$$

$$X_2 = \hat{\beta}^1_{N^1} - \hat{\beta}^2_{N^2}$$

$$X_3 = \hat{\gamma}^1_{N^1} - \hat{\gamma}^2_{N^2}$$

and where $A^{(-1)}$ is the inverse matrix of the sum of covariance matricies of the estimators.

### 5.7. Software

Software for carrying out LNRE analyses is currently being developed by various researchers. J.K.Orlov and A.J.Orlov have completed a program (STATEXT) for IBM-compatible PC that estimates the parameters for the extended Zipf and Yule-Simon 'laws'. The output of the program is a plot of *log(r)* versus *log($p_N${r})* and a short list with the main summary statistics and the estimated theoretical vocabulary *V*. Figure 11 shows the output of STATEXT when run on the frequency spectrum of the English suffix *-ness*, using the extended Zipf's 'law'. STATEXT exploits the independence of the head and tail of the distribution, plotting separate graphs for the left and right hand sides of the dis-

tribution. The plot shows that there is considerable divergence for the lowest probabilities. A similar plot using the Yule-Simon 'law' is, at least to visual inspection, quite satisfactory.

At the Institute of Mathematics of the Georgian Academy of Sciences (Tbilisi), the first author has initiated a more general project for the investigation of various aspects of LNRE distributions such as modelling of LNRE samples, comparison of samples, distribution and interaction between words, and the analysis of word frequency distributions using the generalized Zipf 's 'law'. With respect to the analysis of the frequency spectrum, a program has been developed for PC that estimates parameters for the Zipf and Yule-Simon 'laws'. Like STATEXT, it plots the empirical and expected rank frequency curves, but in addition it calculates confidence intervals for the frequency spectrum as well as the goodness-of-fit in terms of the test statistic $\chi^2_{(M-h)}$, with $h$ the number of parameters. For the data on *-ness* the fit obtained for the Yule-Simon model is shown in table 2. The parameters of the Yule-Simon ( $\hat{\beta}$ = 1.009, $\hat{t}$ = 3.342) model were estimated using the Tbilisi program ($\chi^2$ = 65.66, q « 0.001).

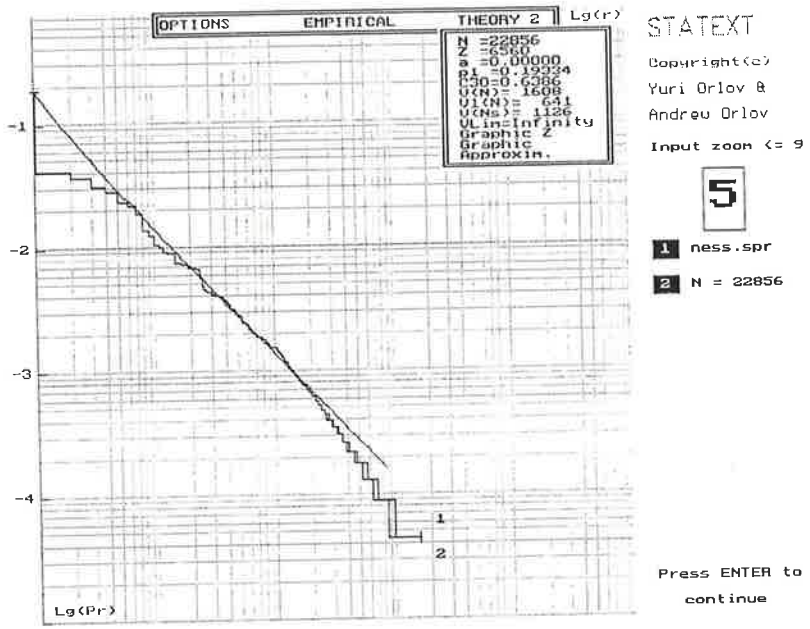



Figure 11. Rank-probability plot of the extended Zipf 's 'law': STATEXT applied to the English suffix *-ness*.

A semi-automatic program for estimating the parameters of the extended Waring-Herdan 'law' has been developed at the Max-Planck Institute for Psycho-



linguistics by the present authors. This program, which runs under UNIX, allows one to interactively search through the parameter space for parameter values satisfying $V_N(0) = \hat{V}_N(0)$ and $V_N(1) = \hat{V}_N(1)$ and minimalizing $\chi^2_{(M-h)}$. Thusfar, attempts to develop a fully automatic estimation procedure have failed, due to the for numerical calculation infelicitous expression for $V^2$ (143) and the bounded parameter ranges. The results obtained for English *-ness* are summarized in table 2. The fit, with the estimated parameters ($\hat{\alpha}$ = 0.712, $\hat{\beta}$ = 1.075, $\hat{t}$ = 0.1) is optimal in the chi-square sense, other choices of the parameters leading to higher values of the $\chi^2$ statistic. Since $\chi^2_{(4)}$ = 8.21, q = 0.084, we may be confident that a reasonable fit has been obtained.

Table 2. Observed and estimated frequency spectrum: *-ness*

| m | $\hat{V}_N(m)$ | $V_N(m)$ | | | |
|---|---|---|---|---|---|
| | | Yule-Simon | Waring-Herdan | Lognormal | Gauss-Poisson |
| 1 | 749 | 646 | 748 | 523 | 749 |
| 2 | 215 | 257 | 228 | 226 | 229 |
| 3 | 126 | 144 | 110 | 130 | 116 |
| 4 | 68 | 94 | 65 | 86 | 73 |
| 5 | 59 | 66 | 44 | 62 | 51 |
| 6 | 30 | 50 | 32 | 47 | 38 |
| 7 | 31 | 39 | 24 | 37 | 30 |
| 8 | 29 | 31 | 19 | 30 | 24 |
| 9 | 22 | 25 | 15 | 25 | 20 |
| 10 | 20 | 21 | 13 | 21 | 17 |

The authors have completed a fully automatic estimation and evaluation programs for the lognormal 'law' and the Gauss-Poisson 'law'. The results obtained for *-ness* can be found in table 2. For the lognormal 'law', the parameter values $\hat{\mu}$ = -5.0, $\hat{\sigma}$ = 2.570 lead to a minimal chi-square value ($\chi^2_{(4)}$ = 206.73) that, unfortunately, fails to meet any standards of acceptability (q = 0.000). The lowest chi-square value for the Gauss-Poisson 'law', $\chi^2_{(4)}$ = 6.292, q = 0.178, was obtained for the parameters $\hat{\gamma}$ = 0.5, $\hat{b}$ = 0.0092, $\hat{c}$ = 0.0264. Evidently, the Gauss-Poisson 'law' provides the best fit. Perhaps not surprisingly, the 'law' with the smallest number of parameters, the lognormal 'law', fails to meet the simultaneous requirements $\hat{V}_N = V_N$ and $\hat{V}_N(1) = V_N(1)$ [1].

[1] In this paper, the lognormal 'law' is fitted to the data using the expressions (132), the integrals being evaluated numerically by means of Romberg integration (see Press et al. 1988). The results obtained contrast with those reported in Baayen (1993b). Using the approximation method suggested by Carroll (1967), he obtained reasonable fits for the more

Table 3. Observed and estimated frequency spectrum: *en-*.

| m | $\hat{V}_N(m)$ | $V_N(m)$ | | |
|---|---|---|---|---|
| | | Waring-Herdan | Lognormal | Gauss-Poisson |
| 1 | 11 | 6 | 11 | 11 |
| 2 | 9 | 4 | 8 | 7 |
| 3 | 4 | 3 | 6 | 5 |
| 4 | 2 | 2 | 4 | 4 |
| 5 | 1 | 2 | 4 | 3 |

By way of comparison, consider table 3, which lists the results obtained for the unproductive prefix *en-*. The lognormal 'law' ($\hat{\mu}$ = -2.0, $\hat{\sigma}$ = 2.335, $\chi^2_{(3)}$ = 4.27, $q$ = 0.234) does much better than the extended Waring-Herdan 'law' ($\hat{\alpha}$ = 0.3, $\hat{\beta}$ = 1.0027, $\hat{I}$ = 10, $\chi^2_{(3)}$ = 34.47, $q$ = 0.000), which fails to provide a parameter set that simultaneously satisfies the equations $\hat{V}_N = V_N$ and $\hat{V}_N(1) = V_N$ (1). Given that *en-* is located outside the (late) LNRE ZONE (see table 1 and section 3.2.1), and given that the extended Waring-Herdan model is tightly linked with the LNRE ZONE, the lack of accuracy - note the large value of $\hat{I}$ - is to be expected. The most accurate fit is again provided by the Gauss--Poisson 'law' ($\hat{\gamma}$ = -0.0005, $\hat{b}$ = 0.02289, $\hat{c}$ = 0.0683, $\chi^2_{(3)}$ = 3.31, $q$ = 0.346), but even here the extremely low value of $\hat{\gamma}$ and the slightly too high values for $m$ = 3, 4, 5 suggest that this 'law' is stretched to, or perhaps beyond its limits in its attempt to model the frequency spectrum of this unproductive prefix. (For a more detailed comparison of these models with respect to goodness-of-fit for a variety of samples the reader is refered to Baayen 1993b).

## 6. Morphology and the LNRE ZONE

In the previous sections the frequency spectra of the English affixes *-ness* and *en-* have been analyzed in some detail. The suffix *-ness*, a typical example of a productive affix, is characterized by a frequency distribution that is dominated by low-frequency types. Not surprisingly, the theoretical vocabulary as estimated by the Gauss-Poisson 'law' exceeds the observed vocabulary by a factor 5

productive affixes. The more rigidly defined methods used in the present paper, however, suggest that for the more productive affixes the lognormal 'law' fails to reach the same level of accuracy as the Waring-Herdan and Gauss-Poisson 'laws'.



($V$ = 8261, $\hat{V}_N$ = 1607). In contrast, the frequency distribution of the unproductive prefix *en-* is dominated by the higher-frequency types and the theoretical vocabulary $V$ = 114, again calculated using the Gauss-Poisson 'law', is only slightly larger than the observed vocabulary $\hat{V}_N$ = 94.

Interestingly, frequency spectra of full texts resemble the spectrum of the morphological category of nouns in *-ness*, the spectrum of the prefix *en-* being quite atypical. Since the large numbers of rare types appearing in the frequency spectra of the productive morphological categories as realized in some text necessarily appear as the 'rare events' of the frequency spectrum of the text as a whole, it seems natural to explore the hypothesis that productive word formation processes anchor texts in the LNRE ZONE. We will investigate this possibility by analysing the morphological constituency of the words appearing in two 'texts', E.Bronte's 'Wuthering Heights' ($N \approx$ 120,000), the full text of which was obtained by anonymous ftp from the Online Book Initiative at obi.std.com, and the Dutch INL corpus ($N \approx$ 40,000,000), using the word frequencies as given in the CELEX lexical database (Burnage, 1990).

### 6.1. The Development of Morphology in Bronte's 'Wuthering Heights'

First consider E.Bronte's novel. According to the tests developed in section 3.2.1, we are dealing with a text that is located in the central LNRE ZONE: $C_L$ = 0.165, the number of hapaxes is increasing ($\frac{d}{dN}V_N(1) = \frac{1}{N}(V_N(1) - 2V_N(2))$= (1/119321)*(2427 - 2*973) > 0), and in addition, the theoretical vocabulary $V$ as calculated using the Gauss-Poisson 'law' encompasses some 12,150 word types, a number exceeding by roughly a factor 2 the observed vocabulary (6420).

One way of investigating the extent to which productive morphological rules may be held responsible for this novel's location in the central LNRE ZONE is to focus on how morphology contributes to the growth rate $\frac{d}{dN}V_N$ of the vocabulary. Using (59), we may estimate the growth rate by $\frac{\hat{V}_{119321}(1)}{119321}$ = 0.02. Note that the growth rate, when viewed as a function of the sample size, is completely determined by the number of hapax legomena. Thus it seems natural to approach the question of whether morphology effects a text's location in the LNRE ZONE by investigating what proportion of the hapax legomena are morphologically complex, since this will allow us to gauge the extent to which word formation gives rise to the substantial growth rate of Bronte's vocabulary as it unfolds through 'text time'. The left hand graph of figure 12 plots the fraction

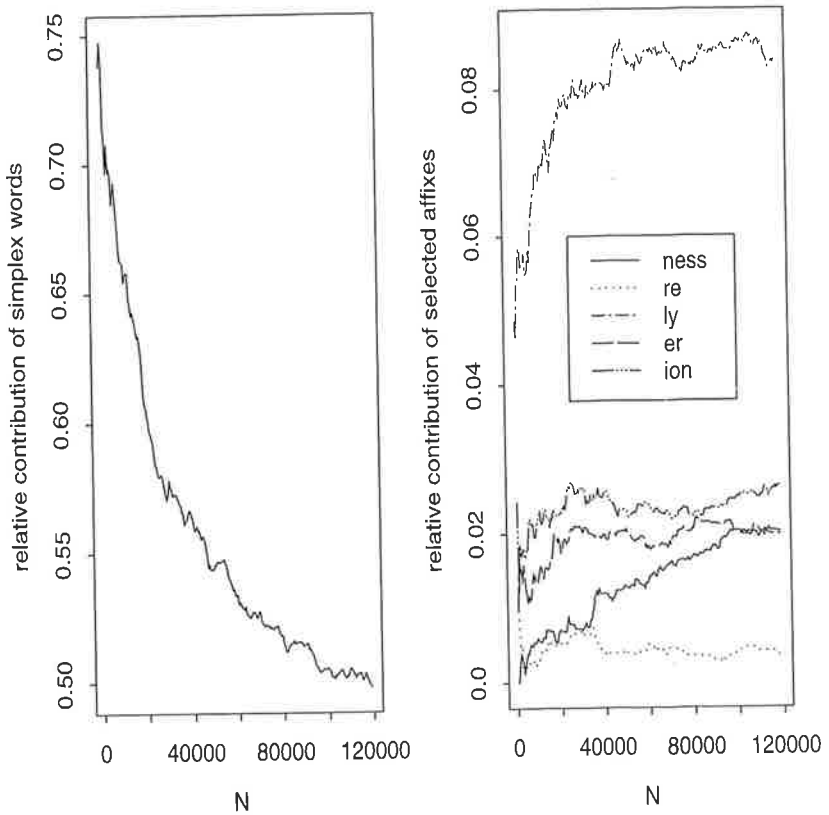



Figure 12. The relative contribution of simplex words and a selected set of affixes to the growth rate of the vocabulary in E.Bronte's 'Wuthering Heights'. The measurement interval for $N$ equals 500 word tokens.

$$\hat{H}_N^{(s)} = \frac{\hat{V}_N^{(s)}(1)}{\hat{V}_N(1)}$$

of hapax legomena that are monomorphemic or simplex *(s)* as a function of the text size $N$. The right hand graph plots the fraction

$$\hat{H}_N^{(e)} = \frac{\hat{V}_N^{(e)}(1)}{\hat{V}_N(1)}$$



of polymorphemic hapaxes for selected affixes $e$. What we find is that $\hat{H}_N^{(s)}$ is a decreasing function of $N$, whereas $\hat{H}_N^{(e)}$ increases with $N$, notably so for highly productive suffixes like *-ness* and especially *-ly*. Note that for the full novel, the relative contribution of morphology is substantial:

$$\hat{H}_N^{(\bar{s})} = \sum_e \hat{H}_N^{(e)} = 0.502.$$

Bronte's novel is too small to allow us to investigate how the relative contribution of morphology to the growth rate will develop for larger samples. The left hand graph of figure 12, however, suggests that a further increase in the relative contribution of morphology may be expected for larger texts. To gain some insight into the 'limiting' properties of $\hat{H}_N^{(s)}$ and $\hat{H}_N^{(\bar{s})}$ we therefore analyze the frequency distribution of a much larger sample, the INL corpus of written Dutch.

### 6.2. Morphology in the INL Corpus

The INL corpus, compiled by the Dutch Institute for Lexicography, contains roughly 40,000,000 wordforms. With the exception of the hapaxes, the frequencies of the words occurring in this corpus as well as detailed information on the orthographical, morphological and phonological properties of these words are available in the CELEX lexical database. The first spectrum elements of the frequency distribution of the INL corpus are presented in table 4. Even though the hapax legomena are not registered in the CELEX lexical database, the available spectrum elements allow us to ascertain that even this moderately large text corpus is located in the LNRE ZONE. For instance, using (63) we find that $\frac{d}{dN} V_N(2) > 0$. In addition, $C_L = 0.022$, which is quite high given that some 65,000 types have been registered. Inspection of the morphological constituency of the dislegomena reveals that $\hat{H}_N^{(s)} = 291/7264 = 0.04$, a substantially lower value than the corresponding value for Bronte's 'Wuthering Heights', $540/973 = 0.555$ ($p < 0.001$). This suggests informally that the asymptotic value of $\hat{H}_N^{(s)}$ for $N \to \infty$ will tend to zero.

Table 4. The tail of the frequency distribution of the lemmas registered in the CELEX lexical database.

| $m$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{V}_N(m)$ | - | 7269 | 4355 | 3433 | 2569 | 2296 | 1834 | 1646 | 1391 | 1313 |

The crucial relation between morphology and the LNRE property of texts can also be approached from a slightly different angle. Figure 13 plots the degree of morphological complexity, measured in terms of the number of morphological constituents of a word, as a function of the frequency of that word, using a non-parametric regression technique (see Haerdle 1991). Clearly, morphological complexity is a decreasing function of word frequency, an illustration on the morphological level of Zipf 's 'law of abbreviation' (Zipf 1935). Table 5 illustrates the same point in a slightly different way: a relatively small set of monomorphemic words accounts for the bulk of all tokens, morphological complexity being relatively scarce token-wise but, paradoxically, frequentially dominant type-wise.

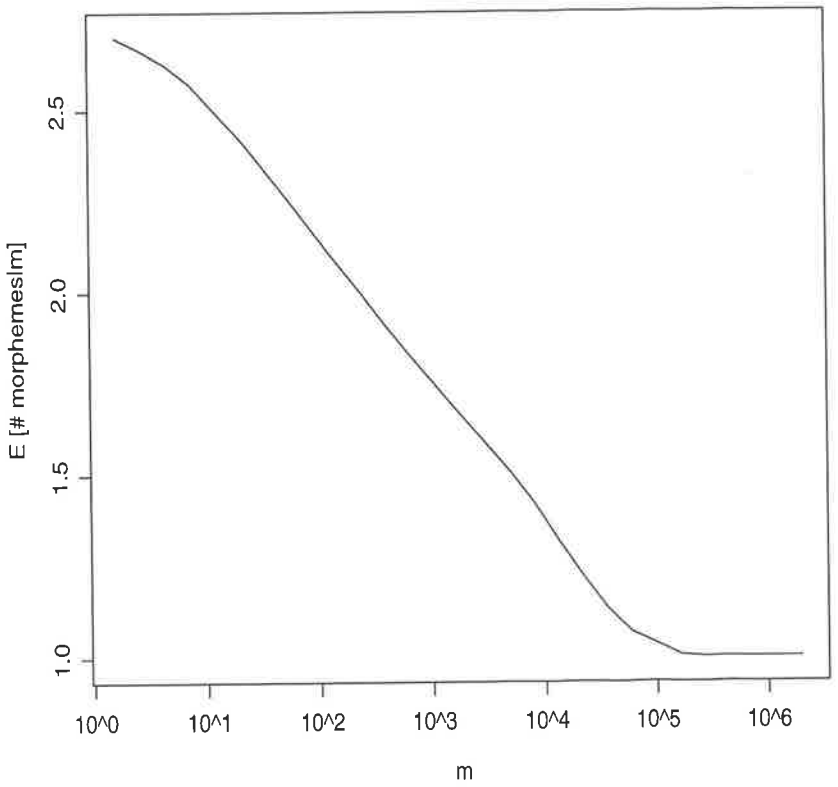



Figure 13. The number of constituent morphemes as a function of (log) word frequency (ln $m$) in the INL corpus. (WARPing approximation of the Nadaraya-Watson estimator using an Epanechnikov kernel, a bin width 0.5 and a window width 2.0).



Table 5. Statistics for the morphological constituency of the lemmas in the CELEX lexical database.

| Morphology | $N$ | $\hat{V}_N$ | $\hat{V}_N(2)$ |
|---|---|---|---|
| monomorphemic words | 30197189 | 8083 | 293 |
| compounding | 1959754 | 32622 | 5324 |
| derivation | 2713506 | 13656 | 1144 |
| synthetic compounding | 85025 | 1206 | 159 |
| undefined | 3464960 | 9795 | 679 |
| Total | 38420434 | 65362 | 7599 |

### 6.3. Productive Rules as LNRE Generators

We have seen that word formation rules play a crucial role in anchoring texts in the LNRE ZONE. This result seriously questions the validity of the rationals for the Rouault, Mandelbrot and Waring-Herdan 'laws' discussed in section 4. The main problem with these rationals is that they fail to take into account what Martinet (1965) has called the 'double articulation' of language, the fact that language is structured on two relatively autonomous planes, the phonological plane and the morphology-syntax plane. Since Markovian models in which words appear as strings of letters focus exclusively on the phonological plane, they cannot and in simulations do not give rise to lexica with realistic frequency-length characteristics, nor can the similarity relations in the lexicon be modelled adequately (see Baayen 1991 for detailed discussion).

Interestingly, the defining characteristic of monomorphemic words is, from a quantitative point of view, that the associated theoretical vocabulary is strictly finite. In contrast, productive morphological categories can be argued to be, at least in theory, infinite. To see this, first consider the simplex words registered in the Ascot version of the Longman dictionary of Contemporary English and the Oxford Advanced Learner's Dictionary. Table 6 summarizes the first ten spectrum elements. The first value of $m$ for which $\frac{d}{dN} V_N(m) > 0$ equals 6, indicating that this sample is located outside the late LNRE ZONE. This is confirmed by a comparison of the observed number of types $\hat{V}_N = 11869$ and the approximated theoretical vocabulary size $V \approx 12{,}000$, calculated on the basis of a rather bad Gauss-Poisson fit ($\chi^2_{(18)} = 836.87$, $\hat{\gamma} = -0.01$, $\hat{b} = 0.0229$, $\hat{c} = 0.00067$). The nearly identical sizes of the observed and theoretical vocabularies is reminiscent of what we have observed for the unproductive prefix en-. In both cases we are dealing with strictly finite populations.

Table 6. The tail of the frequency distribution of the monomorphemic lemmas in the Longman dictionary of Contemporary English and the Oxford Advanced Learner's Dictionary in the Cobuild corpus as registered in the CELEX lexical database.

| $m$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{V}_N(m)$ | 357 | 336 | 259 | 264 | 261 | 225 | 182 | 193 | 163 | 169 |

Next consider the Dutch diminutive *-tje*, an extremely productive derivational suffix. Using the Uit den Boogaart (1977) corpus, the Gauss-Poisson 'law' predicts a theoretical vocabulary of 1,239,156,496 types ($\chi^2_{(13)} = 19.95$, $q = 0.0965$), a value large enough to substantiate the claim that unrestricted productivity gives rise to infinite populations. From this point of view, the following formal definition for LNRE distributions (Khmaladze & Chitashvili 1989), which can be realized only for infinite $V$ (see section 3.2),

$$\lim_{N \to \infty} \alpha_N(1) > 0 \qquad (163)$$

appears to be useful as a formal definition of productivity as well. Of course, many productive categories do not meet this strict probabilistic definition. In the case of *-ness*, for instance, the theoretical vocabulary is approximately 8,000, exceeding the observed vocabulary by 'only' a factor 5. Even in the case of the highly productive English adverbial suffix *-ly* ($\hat{V}_N = 3914$) the estimated theoretical vocabulary equals a 'mere' 24,000 types (Gauss-Poisson estimation, $\chi^2_{(18)} = 166.09$). Observe, however, that *-ly* by itself potentially generates a morphological category with twice as many types as estimated for the monomorphemic English words discussed above (see table 6). This suggests that, even when (163) is not strictly met, the very large numbers of 'morphologically possible words' defined by all word formation rules of the language jointly, will anchor running text in the LNRE ZONE for substantial values of $N$. How large $N$ should be for a text to move out of the late LNRE ZONE into the 'Law of Large Numbers ZONE' is at present unclear. Perhaps the huge corpora that are at present being compiled, such as the British National Corpus and the International Corpus of English, will shed more light on this issue, that, for as yet, has to be left unresolved.



## References


**Baayen, R.H.** (1989). *A Corpus-Based Approach to Morphological Productivity. Statistical Analysis and Psycholinguistic Interpretation.* Diss. Free University, Amsterdam, 1989.

**Baayen, R.H.** (1991). A stochastic process for word frequency distributions. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics: 271-278.* Berkeley.

**Baayen, R.H.** (1992). A Quantitative Approach to Morphological Productivity. In: G.E.Booij & J.van Marle (eds.), *Yearbook of Morphology 1991: 109-149.* Dordrecht: Kluwer.

**Baayen, R.H.** (1993a). On frequency, transparency and productivity. In: G.E. Booij & J.van Marle (eds.), *Yearbook of Morphology 1992: 181-208.* Dordrecht: Kluwer.

**Baayen, R.H.** (1993b). Statistical models for word frequency distributions: a linguistic evaluation. *Computers and the Humanities 26, 331-347.*

**Baayen, R.H. & Lieber, R.** (1991). Productivity and English Derivation: A corpus bases study. *Linguistics, 29, 801-843.*

**Brunet, E.** (1978). *Le Vocabulaire de Jean Giraudoux. Structure et Évolution.* Genève: Slatkine.

**Burnage, G.** (1990). *CELEX; A guide for users.* Nijmegen: Centre for Lexical Information.

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

**Carroll, J.B.** (1969). A Rationale for an Asymptotic Lognormal Form of Word Frequency Distributions. *Research Bulletin - Educational Testing Service. Princeton, November 1969.*

**Carroll, J.B.** (1970). An alternative to Juilland's Usage Coefficient for Lexical Frequencies, and a proposal for a Standard Frequency Index (SFI). *Computer Studies in the Humanities and Verbal Behavior 3, 61-65.*

**Efron, B. & Thisted, R.** (1976). Estimating the Number of Unseen Species: How many Words did Shakespeare Know? *Biometrika 63, 435-447.*

**Good, I.J.** (1953). The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika 40, 237-264.*

**Good, I.J. & Toulmin, G.H.** (1956). The Number of New Species and the Increase in Population Coverage, when a Sample is Increased. *Biometrika 43, 45-63.*

**Guiraud, H.** (1954). *Les Caractères Statistiques du Vocabulaire.* Paris: Presses Universitaires de France.

**Haerdle, W.** (1991). *Smoothing Techniques With Implementation in S.* Berlin: Springer.

**Herdan, G.** (1960). *Type-Token Mathematics*. The Hague: Mouton.

**Herdan, G.** (1964). *Quantitative Linguistics*. London: Buttersworths.

**Kalinin, V.M.** (1965). Functionals Related to the Poisson Distribution, and Statistical Structure of a Text. In: J.V. Finnik (ed.), *Articles on Mathematical Statistics and the Theory of Probability: 202-220*. Providence, Rhode Island: American Mathematical Society.

**Khmaladze, E.V.** (1987). *The Statistical Analysis of Large Number of Rare Events*. Report MS-R8804, Dept. of Mathematical Statistics, CWI. Amsterdam: Center for Mathematics and Computer Science.

**Khmaladze, E.V. & Chitashvili, R.J.** (1989). Statistical Analysis of Large Number of Rare Events and Related Problems. *Transactions of the Tbilisi Mathematical Institute 91, 196-245.*

**Lánský, P. & Radil-Weiss, T.** (1980). A Generalization of the Yule-Simon Model, with Special Reference to Word Association Tests and Neural Cell Assembly Formation. *Journal of Mathematical Psychology, 21, 53-65.*

**Mandelbrot, B.** (1953). *An information theory of the statistical structure of language*. In: Jackson, W.E. (ed.), *Communication Theory: 503-512*. New York, Academic Press.

**Mandelbrot, B.** (1962). On the Theory of Word Frequencies and on Related Markovian Models of Discourse. In: R. Jakobson (ed.), *Structure of Language and its Mathematical Aspects: 190-219*. Proceedings of Symposia in Applied Mathematics, Vol. XII. Providence, Rhode Island: American Mathematical Society.

**Martinet, A.** (1965). *La linguistique synchronique: études et recherches*. Paris: Presses Universitaires de France.

**Menard, N.** (1983). *Mesure de la Richesse Lexicale. Théorie et vérifications expérimentales. Etudes stylométriques et sociolinguistiques*. Genève: Slatkine-Champion.

**Miller, G.A.** (1957). Some Effects of Intermittent Silence. *The American Journal of Psychology 52, 311-314.*

**Muller, C.** (1977). *Principes et Méthodes de Statistique Lexicale*. Paris: Hachette.

**Muller, C.** (1979). Du nouveau sur les distributions lexicales: la formule de Waring-Herdan. In: C. Muller. (ed.), *Langue Française et Linguistique Quantitative: 177-195*. Genève: Slatkine.

**Orlov, J.K.** (1983a). Dynamik der Häufigkeitsstrukturen. In: H. Guiter & M.V. Arapov (eds.), *Studies on Zipf's Law:* 116-153. Bochum: Brockmeyer.

**Orlov, J.K.** (1983b). Ein Model der Häufigkeitsstruktur des Vokabulars. In: H. Guiter & M.V. Arapov (eds.), *Studies on Zipf's Law: 154-233*. Bochum: Brockmeyer.

**Orlov, J.K. & Chitashvili, R.Y.** (1982a). On the Distribution of Frequency Spectrum in Small Samples from Populations with a Large Number of Events. *Bulletin of the Academy of Sciences, Georgia 108.2, 297-300.*



**Orlov, J.K. & Chitashvili, R.Y.** (1982b). On Some Problems of Statistical Estimation in Relatively Small Samples. *Bulletin of the Academy of Sciences, Georgia 108.3, 513--516.*

**Orlov, J.K. and Chitashvili, R.Y** (1983a). On the Statistical Interpretation of Zipf's Law. *Bulletin of the Academy of Sciences, Georgia 109.3, 505-508.*

**Orlov, J.K. and Chitashvili, R.Y.** (1983b). Generalized Z-Distribution Generating the Well-Known 'Rank-Distributions'. *Bulletin of the Academy of Sciences, Georgia 110.2, 269-272.*

**Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T.** (1988). *Numerical Recipes in C. The Art of Scientific Computing*. Cambridge: Cambridge University Press.

**Rouault, A.** (1978). Loi de Zipf et sources markoviennes. *Annales de l'Institute H. Poincare 14, 169-188.*

**Scarborough, D.L., Cortese, C. & Scarborough, H.S.** (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance 3.1, 1-17.*

**Sichel, H.A.** (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association 70, 542-547.*

**Sichel, H.A.** (1986). Word Frequency Distributions and Type-Token Characteristics. *Mathematical Scientist 11, 45-72.*

**Simon, H.A.** (1955). On a Class of Skew Distribution Functions. *Biometrika 42, 435-440.*

**Simon, H.A.**(1960). Some further notes on a class of skew distribution functions. *Information and Control 3, 80-88.*

**Sinclair, J.M.** (ed.) (1985). *Looking Up: An Account of the Cobuild Project in Lexical Computing*. London: Collins.

**Thisted, R. & Efron, B.** (1987). Did Shakespeare write a newly-discovered poem? *Biometrika 74, 445-455.*

**Uit den Boogaart, P.C.** (ed.) (1975). *Woordfrekwenties in gesproken en geschreven Nederlands*. Utrecht: Oosthoek, Scheltema & Holkema.

**Whaley, C.P.** (1978). Word-Nonword Classification Time. *Journal of Verbal Learning and Verbal Behavior 17, 143-154.*

**Yule, G.U.** (1924). A mathematical theory of evolution, based on the conclusions of Dr. J.C.Willis, F.R.S. *Philosophical Transactions of the Royal Society of London Ser. B, 213, 21-87.*

**Yule, G.U.** (1944). *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.

**Zipf, G.K.** (1935). *The Psycho-Biology of Language*. Boston: Houghton Mifflin.

# Text as a Strategic Process

*Luděk Hřebíček, Prague*

Any text can be considered to be an item that has originated with some intention of its producer. This concerns also the reception of text. Hence it follows that between two arbitrary points of the language phenomenon called text, for example between its beginning and end, there is a relation which can be deciphered as a functional relation of a special sort. All such relations can be treated as an intention directed to some final state of text. Thus text is a language construct, the final state of which is accessible only through certain successive inner states; each of these states is, in a certain sense, a final state and each is also of a sequential character.

In linguistics, this kind of comprehension has been labeled "strategy". This excellent idea offers clues for the solution of many problems. The idea of process-grammar should be recalled in this connection too. In some works dealing with the concept of strategy (cf. T.A. van Dijk 1980, 1985; T.A. van Dijk, W. Kintsch 1983) the idea of strategy is applied to the process of text comprehension. Strategy is presented here as a conceptual complex encompassing subordinated conceptual complexes as well as certain superior ones. The way of presenting axiomatic expressions in such systems leads to conclusions which do not make it possible to confront the conclusions with some unambiguous observational facts and guarantee an explanation.

The scientific treatment of strategy was formalized in game theory (cf. especially the works by J. von Neumann, e.g. J. von Neumann, O. Morgenstern 1953). In the present paper the concept of strategy is not related to the basic modelling idea of game theory, as there is no reason to understand the process of origination of texts as a result of antagonistic trends. Our modelling imagination is based on communication theory and the term strategy is understood here as a synonym of the neutral non-terminological 'intention'.

In quantitative linguistics the idea of text proposed in its structure by its producer as a construct having certain length was introduced and convincingly argued for by Ju.K. Orlov (1982) as well as by his co-authors (cf. also Altmann 1988:59).

Text can be observed not only as a finite phenomenon, but also as a process leading to the creation of a certain language structure. Our aim is to combine the dynamic treatment of text with the recently discovered Menzerath-Altmann law. This connection represents an analytical unification of the linear and non-linear structure of text.

The idea of strategy implies one substantial part or constituent of text,



namely, *communicator,* a useful term for signifying the integration of producer and recipient. A human being, text producer or recipient, is not only an item that interacts with text, but it is also part of it. G. Altmann (1988: 8-9) characterized this situation in the following words:

> *'Speaker/hearer is a language subsystem as much as a liver is a subsystem of an organism.'*

This aspect should not be disregarded in text analyses.

## Strategy in Analyses

The usual understanding of text strategy is semantic. "The author, the poet, etc., aims at making known his/her opinion, feelings...He wants to say that..." Texts and their intention are often characterized in this way. The semantic treatment as a starting point of analyses which are directed to semantic results, however, is hardly acceptable, as such presumptions often contain the conclusions of investigation. At any rate, semantic approaches do not enable one to control these possibilities. Any semantic analysis of text is nothing but text interpretation. Each interpretation of a text represents a creation of a new text, and this new text can be interpreted again, and so on *ad infinitum*. This interpretative regression may finish only due to fatigue or with the arrival of a new generation. And then the interpretation can begin again from an arbitrary point in the regression chain.

Communication is not only a transmission of meaning. Information is a broader concept: it exhibits a measurable property of systems. Its measure is inferred from the degree of organization proper to the respective system. Yet text represents an arrangement of different elements and subsystems moving into, and interacting in a system, and a given arrangement is a vehicle of information. The above mentioned possibility of semantic interpretation is inferred from a certain arrangement of elements and subsystems.

The concept of strategy can be understood as a means leading to a certain arrangement of elements. It cannot be denied that linguistics was set the task of investigating the laws or principles of arrangements and of the ways leading to a certain arrangement.

Information theory brings the concept of entropy into play in order to characterize the degree of organization of a system and thus measures its information. Entropy is a characteristic based on the notion of probability. A text can, however, be characterized in many ways with many other characteristics related to different units at different levels. It can be described as a unit, as a finite system of systems, and it can also be described as a process in which this system originates step by step. Consequently, the notion of text has two basic

meanings: a completed formation and a sequence of parts, mutually embedded, having a common beginning. The process of origination of a text can be stopped at arbitrary moments, and the formations produced in this way can also be called texts. One of the aims of the scientific inquiry into a text is to find what is different and what is common to both kinds of text. The concept of strategy seems to be important for this investigation.

The easiest obtainable characteristics are those defined in mathematical statistics. Different parts of texts can be characterized by mean values, variation, probability, entropy, etc. The communication strategy can be understood as the communictor's endeavor to attain a certain organization of text structure. The analysis of this strategy should describe the process of arrival at a given arrangement or organization of units. As was indicated above, text reception in a certain sense represents the production of a new text through interpretation of the text stimulating this interpretation. Any analysis using some characteristics represents a simplification of the facts observed. Only after many steps of examination and after the construction of theories presenting explanations can a general theory of text be formulated.

The simplification contained in the present paper can be based on the following understanding of the concept of strategy:

*Strategy is a set of steps aiming at a certain value of a supposed characteristic of text.*

This simplification can be interpreted as the communicator's intention to reach a certain value of a characteristic supposed by a text theory. A communicator creates a text with the purpose of reaching this value. This evidently is not in concordance with facts. There is no reason to impersonate the intention or strategy in the respective communicator. As the linguistic disclosure of the strategy cannot be reached through the imitation of the communicator's intuitive semantic procedure, other ways of scientific analyses must be found. The origination of a text is attended by many sub-processes coordinated by statistical regularities. The complicated structure of the phenomenon reveals chance. This structure has nothing in common with a personified conscious intention. Consequently, strategy cannot be identified with the communicator's wishes. The complicated system

/TEXT-COMMUNICATOR/ <=> (INTERPRETED) TEXT

is full of random phenomena which are not visible at first sight. They must be discovered laboriously and with the help of simplifying presumptions. It must be stressed, however, that chance is not a devil that destroys systems; on the contrary, it is an element that organizes dynamic systems with the help of special laws and principles. The objective of any analytical activity is the



seeking of these laws.

In the following sections an example of a strategic analysis based on one characteristic is presented. Of course, this paper has a preliminary character, and not all consequences will be mentioned here. Our discussion concerns the mean value of a characteristic, and let us recall that this quantity is sometimes called 'mathematical expectation'. And expectation doubtlessly implies some intention or even strategy.

## An Example: Sentence Length Strategy

Let us take sentence length as measured in the number of words. The aim is to demonstrate the strategic tendencies emerging from the data of this characteristic. Two different kinds of sentence length are considered:

- The mean sentence length of the finite text *(finite mean)*.
- The mean sentence length of each uninterrupted part of a text beginning with the first sentence and finishing with an arbitrary sentence, except for the last sentence of the finite text *(sequential mean)*.

The following quantities are observed:

$h,i = 1, 2,..., k$ are the rank numbers of sentences in a text, $h \leq i$; the symbol $h$ is used here only for sums of the values which do not reach the final rank number $k$;

$k$ = the number of sentences of a finite text;

$n_i$ = the length of the $i$th sentence in number of words;

$N_i = \Sigma_{h=1}^{i} n_h$ is the cumulative text length (sequential text length) in number of words;

$n = N_k = \Sigma_{i=1}^{k} n_i$ is the finite text length in number of words.

The strategic aim of a text (as far as sentence length is concerned) is expressed in the characteristic $n/k$, i.e. in the finite mean. In this particular theory we suppose that the communicator of a text behaves as if he/she would like to reach a certain value of $n/k$. The two following quantities can be considered with the purpose of characterizing the process of reaching this strategic aim:

$d_i = n_i - N_i/i$, i.e. the deviation of the $i$th sentence length from the sequential mean;

$D_i = n_i - n/k$, i.e. the deviation of the $i$th sentence length from the finite mean.

It evidently holds that

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

The difference between the two deviations equals the difference of the sequential mean from the final mean.

It can easily be proved that the sum of sequential deviations is

$$(2) \qquad \sum_{h=1}^{i} D_h = N_i - i\frac{n}{k} .$$

In the process of the origination of a text an important role is fulfilled by the mean value of this sum. From (1) and (2) it follows that

$$(3) \qquad \frac{1}{i}\sum_{h=1}^{i} D_h = D_i - d_i .$$

Formulae (1) and (3) testify to the presence of an intention in the process under consideration. The sum of the final deviations $D_i$ can be estimated as a product of the sentence rank number $i$ and the difference of the two deviations.

This treatment of strategy means that each text reaches its strategic aim. If $i = k$, from (2) it follows that the sum of all $D_i$'s is equal to zero:

$$(4) \qquad \sum_{i=1}^{k} D_i = n - k\frac{n}{k} = 0.$$

The strategic process can be characterized also with the help of the increase of the sequential mean. This variable is defined in the following way:

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

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

$$r_i = \frac{in_i - N_i}{i(i-1)}$$

$$(5) \qquad = \frac{1}{i-1}\left(n_i - \frac{N_i}{i}\right)$$

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

For the sake of simplification, let us introduce the symbol:



$$R_i = \sum_{h=2}^{i} r_h .$$

From the above definitions it follows that

$$(6) \qquad n_1 + R_i = \frac{N_i}{i}$$

and consequently also the analogical relation:

$$(7) \qquad n_1 + R_k = \frac{n}{k}$$

These relations enable certain assertions concerning the sentence length strategy. The last two formulae can be understood as expressions concerning the relation of two choices determining the process of reaching the strategic aim. The two choices concern the length of the first sentence of a text and the second one the sum of increases $R_i$ (or $R_k$). Of course, the sum of increases represents $i$ (or $k$) choices of terms forming this sum. It is more natural, therefore, to ask which rule of the choice is applied in order to obtain a given sum of increases; this rule can be formulated as follows:

*The sum of increases (sequential or final) inevitably equals the difference of the mean (sequential or final) and the length of the first sentence.*

Consequently, two constants determine the process, namely, $n_1$ and $n/k$.

Now it is natural to pose the following question: What are the theoretical possibilities for the distribution of terms forming the sum $R_i$ (or $R_k$)?

Two texts (Redhouse 1968 and Emre 1965), in Turkish and Old Ottoman, were selected at random and analyzed with respect to the sentence-length strategy. The observed data are presented in Table 1(a) and (b). Let us note that the former text is prose and the latter a poem; this fact, however, is not important from the viewpoint of the problem discussed.

The values obtained for variables defined in relations (1) - (7) indicate that there is some intention or strategy in the process of creation of the texts, but it is not clearly visible. The course of the values $D_i - d_i = (N_i/i - n/k)$ is presented in Figure 1(a) and (b). It is evident that the number of ways of reaching the strategic aim is infinite.

We are unable to make any conjectures concerning the distribution of terms forming the sum $R_k$. When the values of $r_i$ for the two analyzed texts are observed, the distributions presented in Table 2(a) and (b) are obtained. The values are normally distributed with respect to their means. It can be deduced that chance together with its laws is involved as its main principle. The language system contains a generator of chance taking part in the creation of texts. It

functions when the values of $n_i$ and $n/k$ are selected, cf. (6) and (7).

Table 1
Characteristics of sentence-length strategy
(a) Text Redhouse (1968)

| i | $n_i$ | $N_i$ | $N_i/i$ | $D_i$ | $d_i$ | $r_i$ | $R_i$ | $N_i/i-n/k$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 15 | 15 | 15.00 | 1.10 | 0.00 | - | - | 1.12 |
| 2 | 11 | 26 | 13.00 | -2.88 | -2.00 | -2.00 | -2.00 | -0.88 |
| 3 | 17 | 43 | 14.33 | 3.12 | 2.67 | 1.33 | -0.67 | 0.45 |
| 4 | 10 | 53 | 13.25 | -3.88 | -3.25 | -1.08 | -1.75 | -0.63 |
| 5 | 17 | 70 | 14.00 | 3.12 | 3.00 | 0.75 | -1.00 | 0.12 |
| 6 | 16 | 86 | 14.33 | 2.12 | 1.67 | 0.33 | -0.67 | 0.45 |
| 7 | 7 | 93 | 13.29 | -6.88 | -6.29 | -1.05 | -1.71 | -0.60 |
| 8 | 14 | 107 | 13.38 | 0.12 | 0.63 | 0.09 | -1.63 | -0.51 |
| 9 | 16 | 123 | 13.67 | 2.12 | 2.33 | 0.29 | -1.33 | -0.22 |
| 10 | 28 | 151 | 15.10 | 14.12 | 12.90 | 1.43 | 0.10 | 1.22 |
| 11 | 12 | 163 | 14.82 | -1.88 | -2.82 | -0.28 | -0.18 | 0.93 |
| 12 | 8 | 171 | 14.25 | -5.88 | -6.25 | -0.57 | -0.75 | 0.37 |
| 13 | 12 | 183 | 14.08 | -1.88 | -2.08 | -0.17 | -0.92 | 0.19 |
| 14 | 18 | 201 | 14.36 | 4.12 | 3.64 | 0.28 | -0.64 | 0.47 |
| 15 | 17 | 218 | 14.53 | 3.12 | 2.47 | 0.18 | -0.47 | 0.65 |
| 16 | 31 | 249 | 15.56 | 17.12 | 15.44 | 1.03 | 0.56 | 1.68 |
| 17 | 20 | 269 | 15.82 | 6.12 | 4.18 | 0.26 | 0.82 | 1.94 |
| 18 | 9 | 278 | 15.44 | -4.88 | -6.44 | -0.38 | 0.44 | 1.56 |
| 19 | 10 | 288 | 15.16 | -3.88 | -5.16 | -0.29 | 0.16 | 1.27 |
| 20 | 14 | 302 | 15.10 | 0.12 | -1.10 | -0.06 | 0.10 | 1.22 |
| 21 | 10 | 312 | 14.86 | -3.88 | -4.86 | -0.24 | -0.14 | 0.97 |
| 22 | 42 | 354 | 16.09 | 28.12 | 25.91 | 1.23 | 1.09 | 2.21 |
| 23 | 8 | 362 | 15.74 | -5.88 | -7.74 | -0.35 | 0.74 | 1.86 |
| 24 | 5 | 367 | 15.29 | -8.88 | -10.29 | -0.45 | 0.29 | 1.41 |
| 25 | 19 | 386 | 15.44 | 5.12 | 3.56 | 0.15 | 0.44 | 1.56 |
| 26 | 10 | 396 | 15.23 | -3.88 | -5.23 | -0.21 | 0.23 | 1.35 |
| 27 | 6 | 402 | 14.89 | -7.88 | -8.89 | -0.34 | -0.11 | 1.01 |
| 28 | 11 | 413 | 14.75 | -2.88 | -3.75 | -0.14 | -0.25 | 0.87 |
| 29 | 18 | 431 | 14.86 | 4.12 | 3.14 | 0.11 | -0.14 | 0.98 |
| 30 | 6 | 437 | 14.57 | -7.88 | -8.57 | -0.30 | -0.43 | 0.68 |
| 31 | 11 | 448 | 14.45 | -2.88 | -3.45 | -0.12 | -0.55 | 0.57 |
| 32 | 8 | 456 | 14.25 | -5.88 | -6.25 | -0.20 | -0.75 | 0.37 |
| 33 | 27 | 483 | 14.64 | 13.12 | 12.36 | 0.39 | -0.36 | 0.75 |
| 34 | 13 | 496 | 14.59 | -0.88 | -1.59 | -0.05 | -0.41 | 0.71 |
| 35 | 10 | 506 | 14.46 | -3.88 | -4.46 | -0.13 | -0.54 | 0.57 |
| 36 | 7 | 513 | 14.25 | -6.88 | -7.25 | -0.21 | -0.75 | 0.37 |
| 37 | 14 | 527 | 14.24 | 0.12 | -0.24 | -0.01 | -0.76 | 0.36 |
| 38 | 18 | 545 | 14.34 | 4.12 | 3.66 | 0.10 | -0.66 | 0.46 |
| 39 | 7 | 552 | 14.15 | -6.88 | -7.15 | -0.19 | -0.85 | 0.27 |
| 40 | 8 | 560 | 14.00 | -5.88 | -6.00 | -0.15 | -1.00 | 0.12 |
| 41 | 11 | 571 | 13.93 | -2.88 | -2.93 | -0.07 | -1.07 | 0.04 |
| 42 | 16 | 587 | 13.98 | 2.12 | 2.02 | 0.05 | -1.02 | 0.09 |
| 43 | 10 | 597 | 13.88 | -3.88 | -3.88 | -0.09 | -1.12 | 0.00 |



(b) Text Emre (1965)

| i | $n_i$ | $N_i$ | $N_i/i$ | $D_i$ | $d_i$ | $r_i$ | $R_i$ | $N_i/i-n/k$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 6 | 6 | 6.00 | 2.14 | 0.00 | - | - | 2.14 |
| 2 | 3 | 9 | 4.50 | -0.86 | -1.50 | -1.50 | -1.50 | 0.64 |
| 3 | 3 | 12 | 4.00 | -0.86 | -1.00 | -0.50 | -2.00 | 0.14 |
| 4 | 3 | 15 | 3.75 | -0.86 | -0.75 | -0.25 | -2.25 | -0.11 |
| 5 | 1 | 16 | 3.20 | -2.86 | -2.20 | -0.55 | -2.80 | -0.66 |
| 6 | 2 | 18 | 3.00 | -1.86 | -1.00 | -0.20 | -3.00 | -0.86 |
| 7 | 6 | 24 | 3.43 | 2.14 | 2.57 | 0.43 | -2.57 | -0.43 |
| 8 | 2 | 26 | 3.25 | -1.86 | -1.25 | -0.18 | -2.75 | -0.61 |
| 9 | 4 | 30 | 3.33 | 0.14 | 0.67 | 0.08 | -2.67 | -0.53 |
| 10 | 5 | 35 | 3.50 | 1.14 | 1.50 | 0.17 | -2.50 | -0.36 |
| 11 | 1 | 36 | 3.27 | -2.86 | -2.27 | -0.23 | -2.27 | -0.59 |
| 12 | 4 | 40 | 3.33 | 0.14 | 0.67 | 0.06 | -2.67 | -0.53 |
| 13 | 5 | 45 | 3.46 | 1.14 | 1.54 | 0.13 | -2.54 | -0.40 |
| 14 | 1 | 46 | 3.29 | -2.86 | -2.29 | -0.18 | -2.71 | -0.58 |
| 15 | 5 | 51 | 3.40 | 1.14 | 1.60 | 0.11 | -2.60 | -0.46 |
| 16 | 3 | 54 | 3.38 | -0.86 | -0.38 | -0.03 | 2.63 | -0.49 |
| 17 | 2 | 56 | 3.29 | -1.86 | -1.29 | -0.08 | -2.71 | -0.57 |
| 18 | 1 | 57 | 3.17 | -2.86 | -2.17 | -0.13 | -2.83 | -0.69 |
| 19 | 4 | 61 | 3.21 | 0.14 | 0.79 | 0.04 | -2.79 | -0.65 |
| 20 | 3 | 64 | 3.20 | -0.86 | -0.20 | -0.01 | -2.80 | -0.66 |
| 21 | 3 | 67 | 3.19 | -0.86 | -0.19 | -0.01 | -2.81 | -0.67 |
| 22 | 4 | 71 | 3.23 | 0.14 | 0.77 | 0.04 | -2.77 | -0.63 |
| 23 | 3 | 74 | 3.22 | -0.86 | -0.22 | -0.01 | -2.78 | -0.64 |
| 24 | 3 | 77 | 3.21 | -0.86 | -0.21 | -0.01 | -2.79 | -0.65 |
| 25 | 2 | 79 | 3.16 | -1.86 | -1.16 | -0.05 | -2.84 | -0.70 |
| 26 | 7 | 86 | 3.31 | 3.14 | 3.69 | 0.15 | -2.69 | -0.55 |
| 27 | 6 | 92 | 3.41 | 2.14 | 2.59 | 0.10 | -2.59 | -0.45 |
| 28 | 6 | 98 | 3.50 | 2.14 | 2.50 | 0.09 | -2.50 | -0.36 |
| 29 | 6 | 104 | 3.59 | 2.14 | 2.41 | 0.09 | -2.41 | -0.27 |
| 30 | 6 | 110 | 3.67 | 2.14 | 2.33 | 0.08 | -2.33 | -0.19 |
| 31 | 2 | 112 | 3.61 | -1.86 | -1.61 | -0.05 | -2.39 | -0.25 |
| 32 | 4 | 116 | 3.63 | 0.14 | 0.38 | 0.01 | -2.38 | -0.24 |
| 33 | 7 | 123 | 3.73 | 3.14 | 3.27 | 0.10 | -2.27 | -0.13 |
| 34 | 7 | 130 | 3.82 | 3.14 | 3.18 | 0.10 | -2.18 | -0.04 |
| 35 | 5 | 135 | 3.86 | 1.14 | 1.14 | 0.03 | -2.14 | 0.00 |
| 36 | 3 | 138 | 3.83 | -0.86 | -0.83 | -0.02 | -2.17 | -0.03 |
| 37 | 3 | 141 | 3.81 | -0.86 | -0.81 | -0.02 | -2.19 | -0.05 |
| 38 | 2 | 143 | 3.76 | -1.86 | -1.76 | -0.05 | -2.24 | -0.10 |
| 39 | 2 | 145 | 3.72 | -1.86 | -1.72 | -0.05 | -2.28 | -0.14 |
| 40 | 3 | 148 | 3.70 | -0.86 | -0.70 | -0.02 | -2.30 | -0.16 |
| 41 | 5 | 153 | 3.73 | 1.14 | 1.27 | 0.03 | -2.27 | -0.13 |
| 42 | 7 | 160 | 3.81 | 3.14 | 3.19 | 0.08 | -2.19 | -0.05 |
| 43 | 6 | 166 | 3.86 | 2.14 | 2.14 | 0.05 | -2.14 | 0.00 |

Table 2
The distribution of $r_i$
(a) Text Redhouse (1968)

| $r_i$ | Occurrence |
|---|---|
| 1.00 - 1.50 | 4 |
| 0.50 - 0.99 | 1 |
| 0.00 - 0.49 | 10 |
| -0.01 - -0.49 | 23 |
| -0.50 - -0.99 | 1 |
| -1.00 - -1.50 | 3 |

The actual mean is $1/(i-1)/R_k = -0.0266$

(b) Text Emre

| $r_i$ | Occurrence |
|---|---|
| 0.20 - | 1 |
| 0.10 - 0.19 | 5 |
| 0.00 - 0.09 | 11 |
| -0.01 - 0.10 | 16 |
| -0.11 - -0.20 | 4 |
| -0.21 - -0.30 | 2 |
| -0.31 - | 3 |

The actual mean is $1/(i-1)R_k = -0.0510$

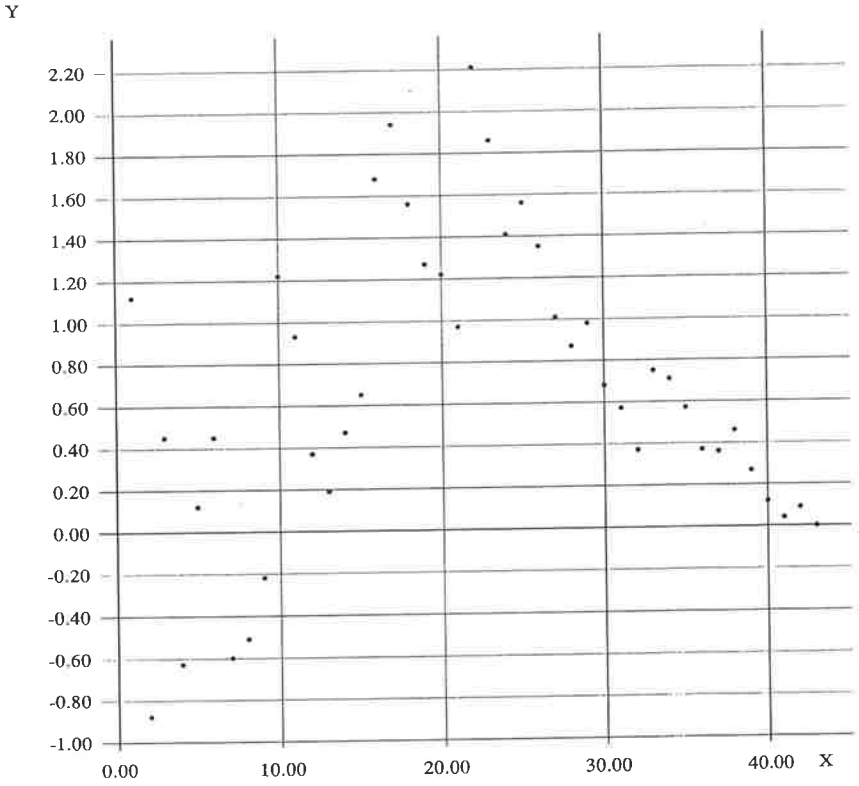



Figure 1a.



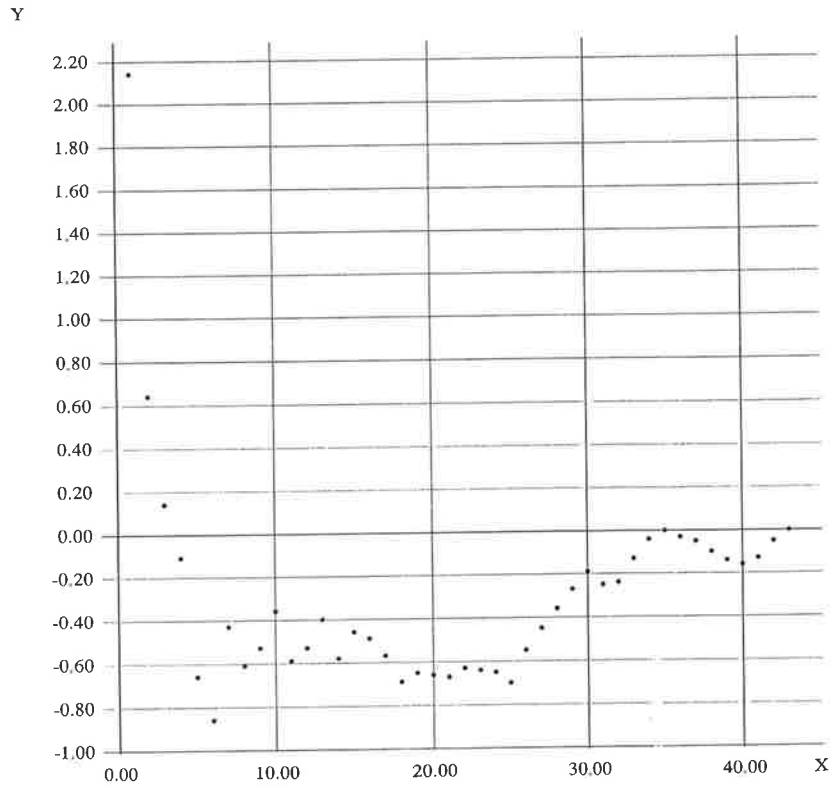



Figure 1b.

## Lexical Structure and Sentence Length

When the values of $n_i$ are treated in the way described above, arbitrary numbers (e.g. the random sampling numbers) can be substituted for $n_i$ with the same consequences. They are valid regardless of the source of observation.

This is a frequent but unconvincing objection against the quantitative approach to scientific problems. In short, it can be said that meanings are not removed from quantities when they obtain a numerical expression. And meanings cannot be arbitrarily interchanged. Construction of theories is the only scientific way to obtain latent meanings and the items bearing them. This can be demonstrated by the connection between the sentence length and the lexical structure of a text.

Lexical structure is quite a vague notion and can be considered in different ways. The following treatment is based on the Menzerath-Altmann (MA) law. This law describes the relation between language constructs and their constituents in accordance with the following equation:

(8) $y = Ax^b$ ,

where x = the length of a construct (e.g. the word-length in number of syllables)
y = the length of constituents (e.g. the length of syllables in number of phonemes),
A, b = constants.

The MA law determines the value of *b* to be negative. Expressed in words the law means that the longer the language construct, the shorter its constituents.

This law was verified for different constructs and their constituents in different languages, cf. especially G. Altmann, M.H. Schwibbe (1989). R. Köhler (1986) revealed the consequences of the law for different language levels, particularly for the lexical level.

The MA law was also applied to a newly defined language construct which can arbitrarily be called *text aggregate*. It is defined as a set of all sentences of a text containing a certain lexical unit. In a corpus of texts it was verified that text aggregate are constructs in the sense of the MA law, cf. L. Hřebíček (1992). The constituents of these constructs are sentences of a given text.

This theory is intuitively acceptable as it is evident in advance that a text cannot be a construct of sentences being its constituents.

It was also indicated that text aggregates function exactly as constituents of the text in the sense of the same law, cf. L. Hřebíček (1991).

The actual meaning of constant $A$ in (8) is obvious: if $x = 1$, then $y = A$. Consequently, in the case of text aggregates, $A$ is the mean sentence length of aggregates having length 1 (in the number of sentences), i.e. in aggregates based on *hapax legomena* of the text supposed.

As far as the second constant, i.e. $b$, is concerned, certain analyses indicate that it is a characteristic of text dimension that expresses the differentiation or level of heterogeneity of the system in question; in this instance, of the system of text aggregates.

An arbitrary text can be analyzed from the point of view of its aggregates. Thus, constructs having the length $x = 1$, $x = 2$, $x = 3$,...(in the number of sentences) are obtained. The theoretically maximum length of an aggregate is $x = k$ corresponding to an aggregate based on a unit occurring in each sentence of the text. Consequently, this aggregate contains all sentences of the text and thus its mean sentence length equals $n/k$. Therefore, (8) can be rewritten in the follwing way:



(9) $y_k = Ak^b = \frac{n}{k}$ .

With regard to formula (7), it then holds that

(10) $Ak^b = n_1 + R_k$

and

(11) $Ak^b = n_1 + \sum_{i=2}^{k} \frac{1}{i-1} d_i$ .

Formulae (10) and (11) indicate that the choice of values of quantities $n_1$ and $R_k$ is closely connected to the lexical structure and to the distribution of *hapax legomena* of the text. The shape of text is, to a certain degree, based on the choice of the properties of its initial sentence.

Table 3
The concordance of the values of $n/k$ and $Ak^b$

| Text | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n/k$ | 8.4 | 18.1 | 14.1 | 13.9 | 16.6 | 3.3 | 8.7 | 10.9 | 14.4 | 13.9 | 3.9 |
| $Ak^b$ | 8.0 | 26.7 | 13.8 | 13.9 | 17.0 | 3.1 | 9.1 | 10.0 | 14.5 | 15.3 | 2.7 |

$A$ and $b$ were estimated from the entire distributions of each text in a corpus of Turkish and Old Ottoman texts.

The values obtained from a corpus of texts (cf. the analyses presented in L. Hřebíček 1992) demonstrate an evident concordance of $n/k$ and $Ak^b$, cf. Table 3. Text 2 in this Table represents an exception: the observed values are not in concordance with each other. It must be remembered that means are characteristics depending on the normality of the respective variable. This is not the case of Text 2: the interval of the mean sentence length $E(y) \pm u_{0.01}s_y$ (where $s_y$ is the standard deviation for sentence length and $u_{0.01}$ is the 1% quantile of the normal distribution). For Text 2 the limits of the interval are [25.4; 35.5] while $n/k = 18.1$. A deeper analysis of these facts is beyond the topic of the present paper. It can be said, however, that the laws of chance are involved in the problem considered. The other texts of Table 3 have means $n/k$ which fall within the limits of their intervals.

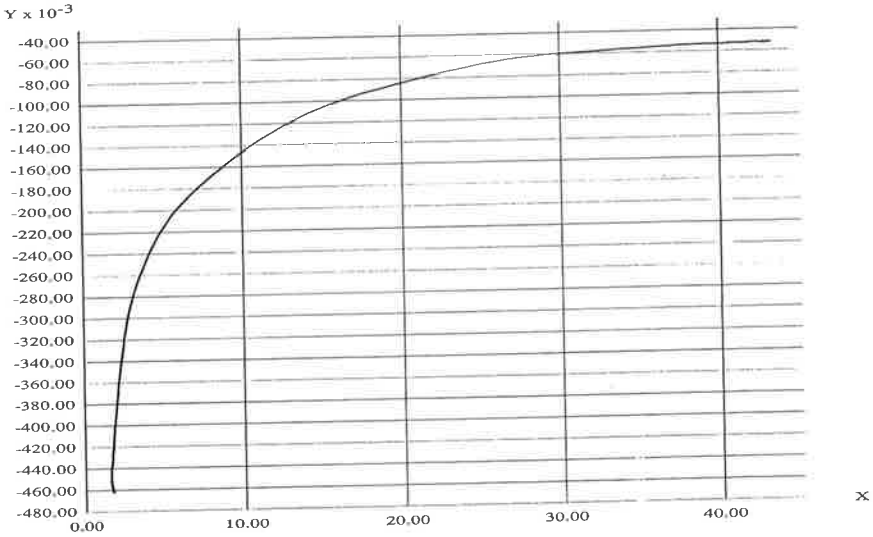



Figure 2a

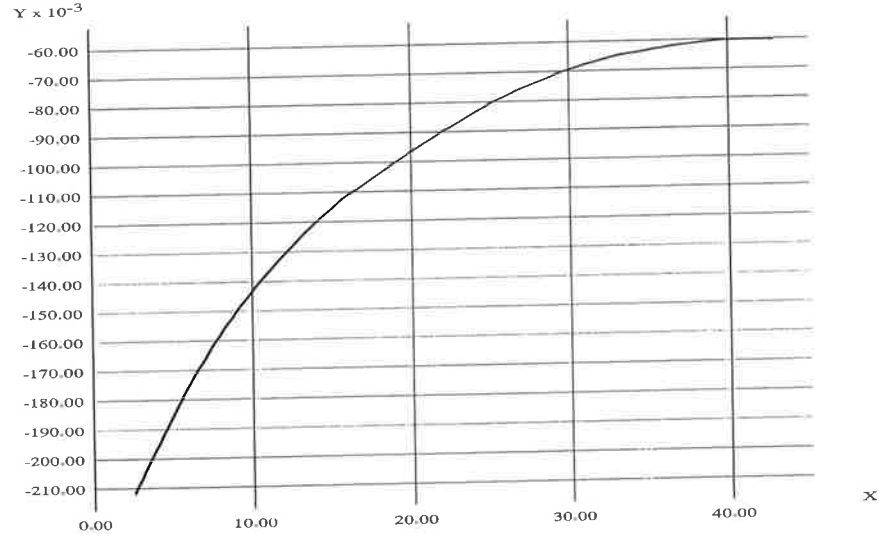



Figure 2b



Text considered as a process increasing from the first sentence to the *k*th sentence represents a series of *k* different texts. All of them have something in common, of course. The structure of the system of these texts becomes increasingly heterogeneous, more complex, and thus it can be expected that the "constant" *b* increases as *i* increases toward *k*. This is testified to by Figure 2(a) and (b) presenting the approximate course of $b_i$ observed in two texts. Their values increase toward a value which is close to zero and which corresponds to a certain strategic aim of the text. When these curves are compared with those in Figure 1(a) and (b), it becomes evident that any characteristic bearing more information presents the text strategy in a more convincing way. Text strategy appears here to be something palpable. Regardless of differences of the values of *b* for the entire text, the shape of the curve $b_i$, which depicts the increasing variable, appears to be common to all texts; this is the consequence of the MA law.

Where then are the actual constants of the text generating process? According to our presumption they are defined by formulae (9) and (10): the length of the initial sentence, the strategic aim and, of course, the laws of chance embodied in $R_k$.

* * *

We are convinced that a future unified theory of text should seek the meaning and mutual relations of many characteristics. The dynamic aspect of strategy enables one to enter the hitherto unknown spaces of text systems.

## References

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

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

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

**Dijk, T.A. van (1985).** Strategic discourse comprehension. In: Ballmer Th. (ed.), *Linguistic dynamics.* Berlin-New York, de Gruyter, pp. 29-61.

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

**Emre (1965)** = [A poem by Yunus Emre]. In: A. Gölpïnarlï (ed.), *Yunus Emre Risâlat al-Nushiyya ve Dîvân.* Istanbul, Garan, pp. 131/132.

**Hřebíček, L. (1991).** Text as a construct of aggregations. In: *Abstracts. QUALICO 91.* Trier, University of Trier, pp. 36-41.

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

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

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

**Orlov, Ju.K. (1982).** Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie "Sprache-Rede" in der statistischen Linguistik). In: Orlov, Ju.K., Boroda, M.G., Nadarejšvili, I.Š., *Sprache, Text, Kunst. Quantitative Analysen.* Bochum, Brockmeyer.

**Redhouse (1968).** = Sir J. Redhouse. In: *Redhouse Yeni Türkçe-Ingilizce Sözlük.* Istanbul, Redhouse Yayïnevi, pp. X-XI.

# Text-Picture Transinformation

*August Fenk, Klagenfurt*

## 0. Introduction

Taking a rather broad meaning of the term "text", we can consider pictures - such as the pictures in a text-book or in a comic - either as components or elements of the "text" in question, or as a special sort of "text". This second approach also gives rise to reflections on a possible "grammar" of pictorial information. It may be that this terminology can be of advantage in other contexts. The basic approach here, as the title suggests, is marked by the distinction firmly entrenched in our language between "text" and "picture", and for this reason alone can best be sketched in a terminology that retains this conceptual division. The analysis of congruences and interaction presupposes that the interacting or congruent things can be kept separate, and this not simply in the form of hypernym and hyponym, for instance.

Anyone who wishes to examine to what extent and in what respect (e.g. rhythm, "emotional colouring", ...) the melody and text of a song fit together, has already made the division between text and melody in a manner that allows both the text and the melody the possibility of autonomous existence (in fact, of course, a text such as a poem is often set to music at a later date, or an existing melody can be given a new text; and even I can hum or whistle a song although I have long forgotten the words). This analytic decompositional approach is in any case legitimate as long as we do not forget that it is only the association of the concrete text with the concrete melody that makes the concrete song. And the success of this approach does not depend on whether, from a meta-perspective, the concept of "song" is only justified by the association of text and melody. Whether it is at all meaningful to say that someone is whistling a song, strictly speaking depends on whether the melody is a sufficient condition for the concept of "song". Fortunately, for the subjective impression and the empirical analysis of the match between text and melody, what I consider to be a necessary condition, and what I consider to be a sufficient condition for the use of the term "song" is of no consequence.

However, our subject is not the song, but the "text-picture composition" or the coherence of text and picture within such compositions. We shall attempt to approach this coherence on two levels.

**Question 1:** *What are the facets and "causes" of text-picture coherence, how does this coherence arise, and what structural methods are used to create it?* This will be investigated using what are known as "logical pictures" as an example. These are an excellent material for study, since not only does the text refer explicitly to pictures, as is required of every specialised text, but also because the picture, too, refers, in often hidden and implied ways, to the text.

**Question 2:** How can text-picture coherence be defined, operationalised and measured as a scalar quantity? The method of transinformational analysis proposed in this connection is based upon the idea of subjective probability and subjective information. According to this idea, the difficulty, the comprehensibility and the informational content of a text cannot be treated as features inherent to the text, but exclusively as a relationship between the specific text and the specific recipient. The latter's entire declarative and procedural, explicit and implicit (prior) knowledge, his linguistic and content-related competence, the creativity and speed with which he generates new hypotheses - all these are involved in determining how much information the text contains for him. Thus the text is not of itself more or less difficult and informative, but these are features that relate to the specific addressee. In relation to our topic: we consider the extent to which the knowledge or the sight of a certain picture reduces the subjective informational content (and thus the cognitive load) of a text as the measure of the (subjective in the above-mentioned sense) "text-picture coherence". "Subjective" does not, however, mean that this coherence is only accessible through statements about subjective impressions. Rather, text-picture coherence, related to the specific "subject" (addressee and recipient of the text), is objectively measurable: its extent is documented in the success in guessing the text or continuing text fragments.

Before we go more closely into this aspect of quantification (see Section 2), we want to examine in detail some qualitative aspects of the "proximity" and relationship between texts and pictures.

## 1. Facets of text-picture coherence. Through the example of logical pictures.

Amongst the illustrations used for the transmission of knowledge, Alesandrini (1987) distinguishes between the "representational/ realistic" type, the "analogical" and the "abstract/logical" graphic. This last type of picture is also frequently referred to as "arbitrary", "non-representational" or "logical", "because these highly schematized visuals do not look like the things they represent but are related logically or conceptually" (Alesandrini, 1984, p. 70). The division into "graphs", "charts" and "diagrams" proposed by Winn (1987), gives names to the



most important sub-groups of "logical pictures".

Whilst "geometricisations" such as Rousseau's graphic solution (cp. Arnheim, 1980, p. 211) of the statement $(a + b)^2$ are not normally included amongst logical pictures, the "Cartesian" diagram, with its x-y coordinates at right angles and a curve that, for instance shows the weight of an infant (y-axis) related to its age (x-axis), is seen as the typical example of a logical picture. (Although the "analogisation" of numeric measurements that takes place with this diagram could also be seen as "geometricisation" in a somewhat broader sense of the word.) Further examples of logical pictures are flow charts, graphic hierarchies and Venn diagrams. The latter two types of graphic are often used to represent those relationships between two concepts (hypernym, hyponym and co-hyponyms...; disparate, overlapping or completely congruent...) that are a precondition for or components of definitions and syllogistic inferences. And in fact it is only in this type of use that these so-called logical pictures have any connection with (conceptual) logic, and it is only in this usage that we want to use the term logical picture in the narrow sense of the word. Such diagrams, in fact, can hardly be seen as "geometricisations" any more: For whether the elements of a Venn diagram look square, circular, elliptical or "amoeba-like", is, unlike the Rousseau square and the (conic section) figures in geometry, not the main issue, but is arbitrary and secondary.

Logical pictures are thus related to the characters that form the text (words) insofar as one can - in a certain respect - consider them as *arbitrary* or as *unmotivated.* They are not iconic in the sense that they directly reproduce, simulate or imitate a reference object. It is precisely for this reason that the elements of the logical picture, e.g. the axes of the Cartesian diagram, require explicit and linguistic labelling (e.g. as t, time), which will vary from case to case, before they can be interpreted as the author intends.

In another respect, however, their form is apparently not arbitrary: According to MacDonald-Ross (1979: 232), the "human habit of translating all our knowledge into spatial terms apparently predates recorded history: all our metaphors of knowledge are spatial, a good sign of conceptual antiquity." It is also argued that diagrams are "metaphorical" (e.g. Tversky et al. 1991) or that "through illustrations, abstract ideas can be concretised by means of spatial metaphors" (Winn 1988: 59, translated from German).

The role of metaphor already indicated only becomes fully transparent if the term "metaphor" is reserved for particular *linguistic* expressions. In works from the Anglo-Saxon world, however, the term "metaphor" is often a very elastic concept (cp. such expressions as "mental metaphor" and "visual metaphor"), and if graphics are also subsumed under metaphors, the view of a basically obvious relationship is blocked (Fenk 1990 a: 368; 1990 b): not only concrete logical pictures, but also entire picture *systems* (such as the Venn diagram type) are a simple *"translation"* of the spatial allusion of a metaphor into a non-linguistic two-dimensional representation. (According to this approach, the logical picture

is in principle and primarily the direct graphic translation of those spatial expressions in which our formal thought takes and had already taken concrete linguistic form at a time when, seen historically, there was no such thing as logical pictures. The individual concrete visualisation is mostly only the application of the "pictorial language" that thus developed to the specific content to be transmitted.) An example: The "hypernym/hyponym" metaphor, or the "subsumption" metaphor suggests, in order to visualise the conceptual relationship, a graphic hierarchy, while the metaphor of the "broad" concept "containing" or "including" another, suggests rather a visualisation through a Venn diagram. And since these metaphors are well-known not only to the producers of graphics, but also to the percipient, they are also more "obvious" and more "comprehensible" for the latter than a truly "arbitrary" graphic would be. In this way the graphic contributes to a reduction of the cognitive costs of text processing, without requiring much additional text (keys, explanations) to clarify the picture.

We have thus already encountered three facets of the "proximity" between text and logical picture:

a Like the words of our language, logical pictures are, in a certain respect, arbitrary.

b Logical pictures require (precisely because of their arbitrary nature) explicit linguistic or paralinguistic (e.g. mathematical) labels or the use of symbols, and their elements themselves become symbols as a result of this labelling.

c In another respect, logical pictures are not arbitrary: they are "graphic figures capturing our figures of speech" (Fenk 1990b).

A fourth way in which the pictures are "adjusted" to the text was investigated above all by Tversky et al. (1991):

d Diagrams, such as time-series and line diagrams, are mostly to be read in the same direction as the texts of the corresponding linguistic community, in Western Europe, for instance, from left to right.

It is possible to devise experiments in which tendency c (verbatim translation of the spatial metaphor!) and tendency d (adjustment to the direction of text reading!) interact and compete with each other. The first such experiments have already been described elsewhere (Fenk 1992, and Fenk, in press), so that it is sufficient here to repeat the main result: If the tree metaphor is used in a text (on the evolution of hominoidea), a bottom-to-top arrangement of the chronology in the flow chart fits the text better than the usual left-to-right arrangement that corresponds to the reading direction. The worst fit of all four directions examined (upwards, downwards, right and left) is that of right to left,



which corresponds neither with the direction of growth of a tree, nor with the reading direction within the line, nor with the reading direction from one line to the next.

But how can we *measure* how well a (draft) picture *fits* the text?

## 2. How to measure text-picture coherence

A method suggested by this present author (Fenk 1990 a, c; cp. also Fenk & Vanoucek 1992) as a means of rendering measurable the dimension we are interested in here, is based on the following consideration: *The information content of a text combined with different pictures is determined using the guessing-game technique, whose principle derives ultimately from Shannon (1951). The picture that makes the greatest contribution to reducing the information content of the text is the one that "fits" the text best (largest transinformation) and also makes the greatest contribution to rendering the text comprehensible. The extent of this "information contribution" can be determined quantitatively (in bits).* And if a text is of use in overcoming a specific problem - a puzzle, a question based on knowledge, the problem of the use of a technical device -, and it is really understood with the assistance of the pictures above all, then the contribution made by the picture to the comprehension of the text is also its contribution to the addressee's ability to solve this type of problem.

Explanatory comments:

a) Of course it is a question of the *subjective* information or the *subjective* uncertainty, i.e. the information that the text contains for the specific addressee (see introduction).

b) Of course one could also use this method to compare the "text with picture" with the "text without picture" in one of those usual and often criticised (e.g. Weidenmann 1988) treatment experiments. Our method would still show a measurable difference (extent of the picture's "advantage") whilst most other investigations known so far are always only able to report on the significant evidence of a picture advantage, which is often "measured" only very indirectly, e.g. by retention tests. (Such knowledge and retention tests are also one-sided since they do not include implicit learning, cp. Mecklenbräuker et al. 1992: 154). This "picture superiority effect" must always arise *trivially* whenever there is a combination of text and picture in which the pictures transmit any "essential" - i.e. subsequently requested - *components of the total message that are not derivable from the text*. We therefore feel that it is superfluous to add to the many investigations of this kind already carried out.

c) In comparison with Shannon's classical guessing game, time can be saved by applying - at the cost of precision - the Weltner procedure (Weltner 1970, 1973), or precision can be increased (the Fenk & Vanoucek method, 1992), with an increase in the time required. The choice of one of these three methods ultimately depends on the extent of the text fragments the guessing of which one is willing to impose upon the subjects. If one wishes to measure the information content, the subjects must in any case be expected to apply a guessing procedure. A mere estimate (cp. Westermann & Hager 1984; Hager & Westermann 1984) of the information content could not replace such measurements even if the persons estimating enjoyed the benefit of an understanding of "information content" based on information theory.

It seems that this not only useful but in principle very simple method of measuring the information contribution has not yet occurred to anyone, although in the meantime a whole flood of works (cp. the list in Glenburg & Langston 1992: 129) deal with the question of "picture-assisted text comprehension", and although this method was waiting to be discovered, prepared on the one hand by Evelyn Goldsmith's (1984) "impoverished text" method, whose author, however, had not yet encountered information theory quantification, and on the other hand by information-theory-based works that were trying to quantify text-picture transinformation:

Frank (1967) proposes determining the "semantic information" of picture descriptions (= "text") by means of guessing games. "Provided that the picture descriptions contain the entire semantic information of the corresponding picture, and nothing more" (Frank 1967: 26, translated from German), the semantic information of the text would be identical with that of the picture, and also identical with the transinformation between text and picture; what would then remain of the picture, and for the time being would be unquantifiable, would be its "aesthetic information". In order to make this measurable, Heinrich (1970) determined the total information of scanned pictures using a guessing game technique adapted for pictures by Attneave (1954); this was done under two conditions - once without and once with previous communication or naming of what was depicted in the picture (e.g. "elephant"). The "knowledge of the picture's content" leads to a reduction of the total pictorial information and the remaining residual information is then to be considered as aesthetic information ("number of degrees of freedom in the possible presentation of the object of the picture").

Our method has different aims. Nor do we assume that the precondition expressly specified by Frank in the above quotation can in fact be met. (Likewise doubt can be expressed whether the conceptual distinction between "aesthetic, semantic and syntactic information components" corresponds to empirical separability.) However, our method is, as it were, (an information-theory development of Goldsmith's "impoverished-text" method, or) the reversal of Heinrich's analysis: the subject of investigation is not the reduction of the subjective picture information resulting from knowledge of the text, but the reduction of the subjective text information resulting from knowledge of the picture. This reversal is fortunately not only applicable to the special case studied by Heinrich of the relationship between picture and picture description, but to all possible types of text-picture composition. A further advantage is that the guessing game technique is again applied to those linguistic-serial character sequences for which it is in fact tailor-made.



## References

**Alesandrini, K.L.** (1984). Pictures and adult learning. *Instructional Science 13,1, 63-77.*

**Alesandrini, K.L.** (1988). Computer graphics in learning and instruction. In: Houghton, H.A. & Willows, D.M. (eds.), *The psychology of illustration, Vol. 2: 159-188.* Springer, New York .

**Arnheim, R.** (1980). *Anschauliches Denken.* Du Mont, Köln 1980 (4th ed.).

**Attneave, F.** (1954). Some informational aspects of visual perception. *Psychological Review 61, 183 - 193.*

**Fenk, A.** (1990a). Graphische Darstellung und kognitive Repräsentation. In: Derner N., Heinze, C.D., Klaus, F.& Melezinek, A. (eds.), *Proceedings of the International Symposium "Ingenieurpädagogik '89": 365-371.* Leuchtturm, Alsbach.

**Fenk, A.** (1990b). What do logical pictures stand for? In: *Verbo-visual literacy: Mapping the field.* University of London, Summaries of the International Visual Literacy Association Symposium 1990.

**Fenk, A.** (1990c). Der Informationsbeitrag von Bildern als meßbare Größe. In: Haug, A., Melezinek, A. & Schutz, V.K. (Hrsg.), *Proceedings of the International Symposium "Ingenieurpädagogik '90": 182-184.* Leuchtturm, Alsbach.

**Fenk, A.** (1992). Pilotstudien zur Text-Bild-Interaktion. In: Melezinek, A. (ed.), *Proceedings of the International Symposium "Ingenieurpädagogik '92": 161-166.* Leuchtturm, Alsbach.

**Fenk, A.** (in press). Spatial Metaphors and Logical Pictures. In: Schnotz, W. & Kulhavy, R.W. (eds.), *Comprehension of graphics.* Pergamon.

**Fenk, A. & Vanoucek, J.** (1992). Zur Messung prognostischer Leistung. *Zeitschrift für experimentelle und angewandte Psychologie 39, 18-55.*

**Frank, H.** (1967). Über den Informationsgehalt von Bildern. *Grundlagenstudien aus Kybernetik und Geisteswissenschaften: Humankybernetik 8, 23-32.*

**Glenberg, A.M. & Langston, W.E.** (1992). Comprehension of illustrated text: Pictures help to build mental models. *Journal of memory and language 31, 129-151.*

**Goldsmith, E.** (1984). *Research into illustration.* Cambridge, Cambridge University Press.

**Hager, W. & Westermann, R.** (1984). Zur direkten Erfaßbarkeit der Verständlichkeit und des Informationsgehaltes von kurzen Texten. *Zeitschrift für experimentelle und angewandte Psychologie 31, 544-56.*

**Heinrich, P.B.** (1970). Durchführung von Rateversuchen mit Hilfe eines Rechners. *Grundlagenstudien aus Kybernetik und Geisteswissenschaften: Humankybernetik 11, 45-56.*

**Macdonald-Ross, M.** (1979). Scientific diagrams and the generation of plausible hypotheses: An essay in the history of ideas. *Instructional Science 8, 223-234*

**Mecklenbräuker, S., Wippich, W. & Bredenkamp, J.** (1992). Bildhaftigkeit und Metakognitionen. Göttingen, Hogrefe.

**Shannon, C.E.** (1951). Prediction and entropy of printed English. *The Bell System Technical Journal 30, 50-54.*

**Tversky, B., Kugelmass, S. & Winter, A.** (1991). Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology 23, 515-557.*

**Weidenmann, B.** (1988). *Psychische Prozesse beim Bildverstehen.* Bern, Huber.

**Weltner, K.** (1970). *Informationstheorie und Erziehungswissenschaft.* Quickborn, Schnelle.

**Weltner, K.** (1973). *The measurement of verbal information in psychology and education.* Berlin, Springer.

**Westermann, R. & Hager, W.** (1984). Zur subjektiven Repräsentation und direkten Erfaßbarkeit der Verständlichkeit, des Informationsgehalts und der Bildhaftigkeit von Lernmaterial. *Zeitschrift für experimentelle und angewandte Psychologie 31, 328-350.*

**Winn, B.** (1987). Charts, graphs, and diagrams in educational materials. In: Houghton, H.A. & Willows, D.M. (eds.), *The psychology of illustration, Vol. 1: 152-198.* New York, Springer.

**Winn, W.D.** (1988). Die Verwendung von Graphiken für Instruktion: Eine präskriptive Grammatik. *Unterrichtswissenschaft 16.3, 58-76.*

# The Logical and Semiotic Status of the Canonic Formula of Myth

(Some preliminaries)

*Solomon Marcus, Bucarest*

## Lévi-Strauss and the mathematical method

Lévi-Strauss' canonic formula of myth

$$f_x(a): f_y(b):: f_x(b): f_{a^{-1}}(y)$$

borrows its symbolism from elementary mathematics. Moreover, in some crucial steps of his approach, Lévi-Strauss makes use of some guiding mathematical metaphors [a few examples: Klein groups, in Lévi-Strauss (1968: 294-295); Klein bottle (1985: 214, 239); system of equations (1985: 228)], sometimes taking the form of drawings (see, e.g. 1985: 211). Sometimes, the mathematical terminology seems less metaphorical: invariant properties (1985: 226), système d'opérations logiques (1985: 227). Some terms like *redundancy* and *code* (1985: 226) are borrowed from Shannon's mathematical theory of communication: '...si les mythes d'une société autorisent toutes les combinaisons, leur ensemble devient un langage dépourvu de redondance'(1964: 339); 'Ce code cosmographique n'est pas plus vrai que les autres' (1964: 247). More important is that mathematics becomes methodologically a term of reference: 'So, we see everything that only a structural analysis of the content of a myth could give: rules of transformations permitting to move from one variant to another by operations similar to those of algebra' (Lévi-Strauss 1973: Chapter XII). Moreover, Lévi-Strauss had a deep understanding of the great methodological change occurring in the fifth decade of our century with the theory of strategic games (J. von Neumann & O. Morgenstern), with cybernetics (N. Wiener) and with the mathematical theory of communication (C. Shannon & W. Weaver); see, in this respect Lévi-Strauss (1973: Chapter XV). He understood - and this happens very seldom among social scientists - that modern mathematics is no longer dominated by the idea of quantity, as in the past, but by the idea of structure: 'However, there is no necessary connection between the concept of *measure* and the concept of *structure*. In social sciences, the structural research appeared as an indirect consequence of the development of modern mathematics, which increasingly stressed the qualitative viewpoint in contrast with the quantitative

perspective of traditional mathematics. In various fields: mathematical logic, set theory, group theory, and topology it became visible that some problems which don't accept a metric solution can however be approached in a rigorous way´ (Lévi-Strauss 1973: Chapter XV). He also showed a deep understanding of the distinction between the deterministic (mechanical) approach and the statistical one (1973: Chapter XV); in this respect, his analysis of the differences between ethnography and ethnology on one hand, and history and sociology on the other hand, is very instructive. He understood the need for a general combinatorial study of all possible types of structures resulting from some given relations [see his considerations about Anatol Rapoport's approach of cyclic phenomena (Lévi-Strauss 1972: Chapter XV)].

## The mathematical vision as a consequence of the structural vision

Against Lévi-Strauss´ explicit advice to not carry further his use of mathematical symbolism (personal communication to Maranda (Maranda & Maranda 1971: 28)), we have to observe his antinomic attitude towards mathematical (linguistic or graphic) metaphors: he introduces them just to betray them and he betrays them just to return to them, again and again.

In the same way in which Lévi-Strauss looks and accounts for the unconscious structure of society, we think that the capacity to perceive the intuitive aspects of mathematical concepts, models and theories and their relevance for social sciences belongs to the unconscious structure of Lévi-Strauss' mind. Indeed, there are enough reasons - and we have already presented some of them - to believe that Lévi-Strauss perceives the intuitive aspect of some mathematical symbols and results and has a feeling for their relevance. So, this phenomenon deserves systematic attention. It is a symptom of the fact that mythical thinking, due to its high degree of structural and systemic virtue, requires a corresponding mathematical treatment. The description of the myth given by Lévi-Strauss (1964: 246) is very near to mathematical formalization: ´La vérité du mythe n´est pas dans un contenu privilégié. Elle consiste en rapports logiques dépourvus de contenue, ou, plus exactement, dont les propriétés invariantes épuisent la valeur opératoire, puisque des rapports comparables peuvent s'établir entre les éléments d´un grand nombre de contenues différents´. So, the structuralist view is not for Lévi-Strauss a simple fashion due to his meeting with linguistics, but his genuine way to see the world. The attempts to use some mathematical tools (Hage & Harary 1983; Petitot 1988) in approaching his mythical thinking follow as a natural consequence. But more than the use of some mathematical tools, the use of a mathematical way of thinking seems to be essential in this respect.

Lévi-Strauss was aware of these facts, as can be seen from his article (1951) and its improved form as chapter III in (1973). Starting from the famous book



of Norbert Wiener (1948), who claims the impossibility of applying mathematical methods of prediction in social sciences, Lévi-Strauss argues in a very convincing way that linguistics, especially phonology, fulfill all the conditions making possible (for Wiener) a mathematical study: linguistic behavior is situated at the level of unconscious thinking: the influence of the observer on the object to be observed is negligible; writing is old enough to lead to series long enough to make possible a mathematical analysis (the available series in Indo-European, Semitic or Chinese linguistics are of the order of about 4000 or 5000 years).

But Lévi-Strauss goes farther and raises the problem of whether some other social phenomena could not allow a treatment similar to the linguistic ones. Starting from Kroeber´s treatment of women's fashion, Lévi-Strauss builds a formal model of the rules of marriage and of kinship systems, based on the modalities assuring the circulation of women within a social group. Once this hypothesis of replacing a biological system of relations by a sociological one is adopted, it remains only - as Lévi-Strauss observes - to develop the mathematical study of all types of exchange that can be conceived among $n$ partners, in order to find out the rules of marriage valid in the society under consideration.

So, formalisation is for Lévi-Strauss a natural consequence of his structural vision and it is motivated by its explanatory capacity. We need a formal model in order to improve our understanding! But this understanding goes, for Lévi-Strauss, through the linguistic metaphor. Marriage rules and kinship systems are viewed as a language, i.e. as a set of operations aimed at assuring a certain type of communication among individuals and groups. *Women* are the *words* of this language; they circulate among various social groups in the same way in which words circulate between individuals using the everyday language. This linguistic mediation has several reasons (we tried to analyse them in Marcus 1969, 1974), one of them being just that pointed out by Lévi-Strauss: among all social sciences, linguistics is methodologically the most advanced. This was also the reason for which Roman Jakobson called linguistics ´the mathematics of social sciences´. The need for a linguistic and a mathematical perspective follows, for Lévi-Strauss, from his general idea that social processes are basically processes of communication. His first step is always a detailed description of the social relations under consideration, but then he tries to build the formal dynamic models which point out the naturally unconscious structure of societies and explain all observed phenomena by means of some types of transformations, whose invariants express the essential aspects of the process (see, especially, Lévi-Strauss 1973, chapter XV). In this respect, we could insert Lévi-Strauss into the tradition of Felix Klein´s Erlangen Program, based on the idea that to understand a field of knowledge means to understand the basic groups of transformations underlying it and to find out the corresponding invariant properties. Let us recall, in this respect, that among the four conditions imposed to a model in order to define a structure, the second one is the following: ´...any model

belongs to a group of transformations, each transformation corresponding on a model of the same family in such a way that the set of these transformations is a group of models' (Lévi-Strauss 1973, chapter XV).

Obviously, Lévi-Strauss' explicit mathematics is poor, but the mathematical potential of his ideas is tremendous.

## The general aspect of the canonic formula

Having in mind the genuine structural thinking of its author and his spontaneous tendency towards mathematical metaphors, let us try to make a radiography of the canonic formula of myth, starting from its literal appearance.

The mathematical aspect concerns both terminology and notation. Terminology includes words like: formula, terms, functions, inverse, linear, non-linear. Notation includes: use of the functional symbolism with both arguments and indices [$f_x(a)$, $f_x(b)$, $f_y(b)$, $f_{a^{-1}}(y)$], use of the sign of division :, and use of the notation of the inverse $a^{-1}$ of a. There is also the notation :: or ≅, rather logical than mathematical, but whose plausible meaning is very ambiguous. Indeed, it can equally mean an equivalence relation, a similarity (resemblance) relation or an analogy or maybe some other things too. At first glance, these different types of binary relation may seem the same or at least their differences may seem negligible. But logic and semiotics go beyond a first glance, in order to obtain a more accurate description of the relations under consideration. We will first introduce the logical apparatus leading to the difference between equivalence and similarity (resemblance), while later we shall discuss two different ways to describe analogy. In a further step, metaphors will be also considered. Provisionally, we shall adopt for ≅ the hypothesis that it is a similarity relation, leading mathematically to what is called in the theory of binary relations a tolerance relation, whose explanation follows.

Given a set A, the cartesian product AxA is the set of all ordered pairs <x,y>, where x and y belong to A. Any part R of AxA is called a *binary relation* in A. When <x,y> belong to R, we write xRy. If xRx for any x in A, R is said to be *reflexive*. If from xRy it follows yRx (for any x, y in A), R is said to be *symmetric*. If from xRy and yRz it follows xRz (for any x, y, z in A), R is said to be *transitive*. Any reflexive and symmetric relation in A is called a *tolerance relation* in A. A tolerance relation which is transitive in A is said to be an *equivalence relation* in A.

Tolerance relations were introduced by C. Zeeman (1962) in connection with what he calls the topology of the brain. The particular situation considered by Zeeman is that of indiscernability: how near should two points be in order to be perceived as a single point? Mathematically, it could be described as follows: x and y are e-near if the distance between x and y is inferior to e (here e is a strictly positive number). We write in this case $xR_ey$ and we observe that $R_e$ is



a tolerance relation, but not an equivalence relation. Obviously, if $e_1 < e_2$ and x and y are $e_2$-near, then they are also $e_1$-near, so it is interesting to take the largest possible value for e. Another example of a tolerance relation which is not an equivalence relation is the relation of synonymy in natural languages. One of the possible ways to test the quality of a dictionary of synonyms is just to check to what extent it fulfills the property of synonymy to be symmetric, but not (always) transitive. The reason why synonymy is not always transitive is given by the possibility that the set C(x,y) of contexts where x and y can be interchanged does not intersect the set C(y,z) of contexts where y and z can be interchanged.

If generally speaking the relation ≅ is one of similarity, thus a tolerance relation, the approximation of ≅ by an equality, i.e. an equivalence relation, leads to the global interpretation of the canonic formula as a proportion: the equality between two ratios. Obviously, this is not a proportion in the proper numerical sense. No element in the formula is a number, so no division is possible. There is only a formal analogy with a proportion. Indeed, given four numbers m, n, p, r such that m:n = p:r, we could read it as "m is with respect to n like p with respect to r". This aspect of the proportion keeps its meaning even when m, n, p, r are no longer numbers, but qualitative entities. In this way we get a *metaphorical proportion*, where the sign = is no longer read as equality, but as similarity. This means that = is no longer an equivalence, but a tolerance. It deserves special attention.

## The metaphorical proportion (again about ≅)

From the beginning Lévi-Strauss showed a great interest in it, as can be seen from his article (1945), where he observes that phonology is with respect to social sciences what nuclear physics is for natural and exact sciences. But the metaphorical proportion already appears to Aristotle, who analyzes examples like : oldness is with respect to the life what evening is with respect to the day. The same type of analogy is considered by Kant (Prolégomènes à toute métaphysique future, pp.146-147) quoted from Perelman & Olbrechts-Tyteca (1988). Let us observe that a metaphorical proportion such as A/B = C/D, where A, B, C, D are qualitative entities, is made possible by a previous, implicit metaphor, resulting from the analogy between the quotient A/B (as well as C/D) and the ordinary quotient between two numbers. This hidden metaphor appears sometimes in scientific research. Let us remember George D. Birkhoff's (1933) aesthetic measure of an object, given by the ratio O:C, where O is the order while C is the complexity of the object under consideration. Birkhoff succeeds in showing that to some simple visual or sonorous objects we can associate numerical measures of their order and of their complexity; but in less simple situations we are not (yet?) able to express O and C by numbers. For instance,

we don't know how to express numerically the order and the complexity of a Beethoven symphony. Another older example, from the past century, is the Weber-Fechner law claiming that when a sequence of external stimuli proceeds in geometric progression, the corresponding sensations (internal responses) are in arithmetic progression. Here too the empirical evidence accounts only for some particular cases, while the general assertion is both a cognitive metaphor and an explanatory hypothesis. Similarly, we assume that the canonic formula of myth, already tested on some particular myths, claims to account for a general situation.

Very important is the metaphorical proportion in the field of temporality (Marcus 1985). For instance, in a very comprehensive context we may assume that the relation between biological and chronological time is similar to the relation between psychological and chronological time, which is similar to the relation between social and chronological time (each time the numerator being the logarithm of the denominator); but these proportions are rather hypothetical, their empirical evidence being still weak. The theoretical importance of this type of relation should not be underestimated. They allow us to make bold connections between geological time and linguistic time (see the field of historical linguistics) and between psychological time and relativistic time (Marcus 1993), opening the way to a transdisciplinary approach.

Interesting metaphorical proportions have already been used in anthropology. So, the foreign monetary system is invested in some societies with a religious meaning, leading to proportions of the type 12/1 shilling = 12 apostles / Christ or 10 cents / 1 shilling = 10 commandments / God (see P. Maranda 1981: 34). But a more significant situation is the use of a metaphorical proportion as a reciprocity symbol underlying all forms of exchange (Maranda 1981: 31). Let us refer again to Pierre Maranda (1981: 26-27): 'A discrete encoder $a_i$ (member of the set A and thus operationally contiguous with the other members of the same set) and a discrete decoder $b_i$ (member of the set B, etc.) are linked through the intermediary of a discrete message or concatenated units of information $c_i$ (member of the set C, etc.) which passes from $a_i$ to $b_i$. But $a_i$ and $b_i$ do not only the first one give and the second receive. They both evaluate the message. By virtue of the axiom of reciprocity, $a_i$ knows that what he gives is binding over $b_i$, a fact of which $b_i$ is aware. $b_i$'s operations are performed by referring $c_i$ to a scale through statements of equivalence or non-equivalence: $c_i$ is weighted by being mapped into the rest of C. The evaluation always takes the form of the proportion, viz. $c_i/b_i = d_i/a_i$, i.e., 'this, which I give you, is to you what that, which you give me in return, is to me' (Maranda 1981: 26-27). Such a use of metaphorical proportion is even bolder than the previous ones, in view of the strong heterogeneity of the entities occurring in the numerator and in the denominator ($b_i$ and $c_i$, and respectively $a_i$ and $d_i$, belong to different types of entities, one of which is a person ($b_i$ resp. $a_i$), the other ($c_i$ resp. $d_i$) being usually a non-person)



A recent example comes from political science. In order to point out the main ideas of his book, Samuel Huntington (1968) proposes what he calls three equations involving various social and political processes: social mobilization / economic development = social frustration, social frustration / mobility opportunities = political participation, political participation / political institutionalization = political instability. So, we have three equalities of the form a/b = c, c/d = e, e/f = g, but a, b, c, d, e, f, and g are not numbers; they express some concepts and processes in sociology and political theory. It was Herbert Simon, a Nobel laureate in economics, who argued the metaphorical status of Huntington's 'equations'; we then tried to deepen this analysis (Marcus 1990). Here we have a special case of metaphorical proportion. We could think, for instance, of what happens if in the second relation we replace the expression of social frustration as it follows from the first relation.

## Controversies concerning the status of analogy

Of some help in understanding the metaphorical proportion could be the famous treatise by Perelman & Olbrechts-Tyteca (1988). J. St. Mill (Système de logique, vol. II., p. 90), quoted from Perelman & Olbrechts-Tyteca, considers analogy weaker than similarity, which is weaker than identity. For him, the value of analogy is to permit the formulation of an hypothesis to be tested by induction. As a matter of fact, as Perelman & Olbrechts-Tyteca observe, analogy is a special case of similarity: it is a similarity among structures; it contrasts with the usual similarity, which is among objects. So, J.St. Mill's hierarchy becomes: identity, similarity between objects, similarity between structures. What is, in this respect, the place of the metaphorical proportion and, particularly, of the canonic formula of myth? We interpret them as analogies, i.e. as similarities between structures. This means that in the formula 'A is with respect to B what C is with respect to D' we have to focus on similarities neither between A and C, nor between B and D, but on the similarities between two relations: the relation between A and B, on the one hand, and the relation between C and D, on the other hand.

In order to illustrate the above idea, let us consider some examples. We begin with numbers. The similarity between 18 and 21 is very weak: they accept 3 as a common factor. The similarity between 42 and 49 is also weak: they accept 7 as a common factor. But despite the fact that 18 is different from 21 and 42 is different from 49, we have 18/42 = 21/49. Another example is the famous solar metaphor for the atom: planets are with respect to the sun what electrons are with respect to the nucleus of the atom. Here, the analogy is neither between planets and electrons, nor between the sun and the nucleus. Planets and electrons are from any point of view very different, their heterogeneity is almost total. The sun and the nucleus of the atom are also very hetero-

geneous. The analogy is reflected by the words *with respect*, it is purely relational. It is not the substance, the internal structure of the electrons which is involved in the analogy with the internal structure of the planets; it is not the analogy between the structure of the sun and that of nucleus of the atom which is in question; the analogy concerns the *behavior* of the electrons with respect to the nucleus, on the one hand, and the *behavior* of the planets with respect to the sun, on the other hand.

What makes the difference between these two situations? When we compare the nucleus of the atom with the sun, we have in view an object-attribute predicate, while comparing the hydrogen atom and the solar system we have to cope with a relational predicate. We will discuss this in another paragraph. In a more expressive way, this difference was described by P. Grenet (quoted by Perelman & Olbrechts-Tyteca 1988: 501): "Ce qui fait l'originalité de l'analogie et ce qui la distingue d'une identité partielle, c'est-à-dire de la notion un peu banale de ressemblence, c'est qu'au lieu d'être un *rapport de ressemblence*, elle est une *ressemblence de rapport*". But Grenet adds: "Et ce n'est pas là un simple jeu de mots, le type le plus pur de l'analogie se trouve dans une proportion mathématique." Perelman & Olbrechts-Tyteca disagree with this last assertion, claiming that "on n'y voit pas ce qui précisément caractérise, selon nous, l'analogie et qui a trait à la différence entre les rapports que l'on confronte". They distinguish A is to B what C is to D' between the *theme* (A, B) and the *phore* (C, D), where the phore is better known and helps to understand the theme. For instance, adopting the solar metaphor for the atom, the theme is given by the electrons and the nucleus of the hydrogen atom, while the phore is formed by the planets and the sun, because the reason to adopt this metaphor was just to reduce what is less known, the atomic system, to what is better known, the solar system, in other words, to better understand the former by means of the latter.

But Perelman & Olbrechts-Tyteca (1988: 501) add a new condition for analogy, which seems for them to be a consequence of the asymmetry already observed between theme and phore: they require that theme and phore belong to different domains (usually, one of them refers to the sensible world, while the other refers to the spiritual world). This is a very critical point in their approach, which touches directly the problem we have to face when dealing with the canonic formula of myth. Indeed, this formula is of the type 'A is to B what C is to D', where it is hard to say that A and B, on the one hand, and C and D, on the other hand, belong to different domains. But what surely we can assert is the difference in complexity between the left part and the right part of the formula.



**Analogy as a relational-predicate mapping**

Another representation of analogy was proposed by Gentner (1982). Following this author, an analogy between a base system B and a target system T (to be investigated) is a relational-predicate mapping M from the objects of B to the objects of T (if F is a relational-predicate, if $b_1$ and $b_2$ are objects of B, if $t_1$ and $t_2$ are objects of T, and if $M(b_1) = t_1$, $M(b_2) = t_2$, then from the truth of $F(b_1,b_2)$ follows the truth of $F(t_1,t_2)$), which is not an object-attribute mapping (given the object attribute A, the truth of A(b) does not imply the truth of A(t), where t = M(b)). Obviously, the base system corresponds to the component *phore*, while the target system corresponds to the component *theme* in the description of the analogy given by Perelman & Olbrechts-Tyteca. An example is Huntington's 'system of equations'. We can consider a numerical base system B = {p, q, r, s, t, u, v} satisfying the relations p/q = r, r/s = t, t/u = v and the conceptual target system T = {a, b, c, d, e, f, g} whose elements are the sociological-political entities considered by Huntington. Let us also consider in B two relational predicates R and S. By R(x, y) we mean: if x is increasing, then y is increasing; by S(x, y) we show that when x is increasing, y is decreasing. So, we have R(p, r), R(r, t), R(t, v), S(q, r), S(s, t), S(u, v). Now let M be a relational-predicate mapping from B to T, with M(p) = a, M(q) = b, M(r) = c, M(s) = d), M(t) = e, M(u) = f, M(v) = g. Thus, we have R(M(p), M(r)) = R(a, c), R(M(r), M(t)) = R(c, e), R(M(t), M(v)) = R(e, g) and corresponding relations for the relational predicate S. We can conclude that writing a/b = c, c/d = e, e/f = g expresses in an analogical way the fact that when a(c or e) is increasing, c(e or g) is increasing and when b(d or f) is increasing, c(e or g respectively) is decreasing. 'Increasing' and 'decreasing' involve only some order relations, they don't require numbers.

Now the problem is: can we make explicit all the above parameters when dealing with the canonic formula of myth? Obviously, our target system is the right part of the formula, but are we able to reduce ≅ to a rigorous relational-predicate mapping between $f_x(a):f_y(b)$ and $f_x(b):f_{a^{-1}}(y)$?

Gentner does not say it explicitly, but it is clear that a difference in nature or complexity between the base system and the target system is obligatory, otherwise there would be no reason to investigate the latter by means of the former. With respect to the distinction between a relational-predicate mapping, and an object-attribute mapping, it is important to determine to what extent the existence of one type of such mappings is independent from the existence of the other type.

**The singularity of the canonic formula with respect to the status of analogy**

It is generally accepted that the second part of the formula is more difficult than the first, in view of the (so far) mysterious transformations from a to $a^{-1}$ and from y as function to y as term. So, despite the order of the two quotients in the proportion, giving the impression that we compare A:B to the better known relation C:D, the reality is just the opposite: A:B raises no problem, while C:D (mainly D) is to a large extent enigmatic. So, should we decide that A:B is the phore and C:D is the theme in Lévi-Strauss' proportion? In order to answer this question, we have to observe that the distinction between theme and phore is based on two facts: (a) the knowledge of the phore usually anticipates (in the chronological sense) the knowledge of the theme (the knowledge of the solar system anticipates chronologically the knowledge of the atom); (b) the intellectual effort needed to understand the theme is larger than that needed to understand the phore, so there is an asymmetry between theme and phore, the former being more enigmatic than the latter (when the solar metaphor was adopted as an explanatory hypothesis concerning the structure of the atom, this structure was less known than that of the solar system; and perhaps even today this gap still exists). Sometimes, the point (a) has to be replaced by another one: (a') the phore belongs to the sensible world, while the theme is only intelligible. This is the case, for instance, in the Aristotelian analogy discussed by Perelman & Olbrechts-Tyteca (1988: 501): in the same way in which the eyes of the bat are blinded by the light of the day, the intelligence of our soul is blinded by the most natural obvious things. But in many cases the knowledge of the sensible aspects of the world precedes the knowledge of the aspects which are only intelligible, just because the former are usually more easily understood than the latter, so a and a' to a large extent overlap. As a matter of fact, both the solar system and the atomic system belong to the material world and both are only partially observable, because they are beyond the usual order of magnitude of everyday phenomena (one of them belongs to the infinitely large, the other belongs to the infinitely small). There is however a difference in complexity, maybe not of an intrinsic nature, but due to the historical development of scientific knowledge.

Now let us turn back to the canonic formula of myth. Its singularity is given by the fact that, having the form 'A:B is like C:D', C and D are not independent with respect to A and B, as in the other examples considered; C is dependent on both A and B, because C borrows from A the function and from B the term, while D is using a function that cannot be defined before knowing the term of A and is using a term which cannot be determined before knowing the function occurring in B, so D is also dependent on both A and B. There is nothing of this type in the solar-atom proportion or in the other proportions considered. The definition of the kernel of an atom or of its other components



does not refer to the components of the solar system; the oldness and the life of a human being can be understood with no reference to the various parts of the day.

**The strange irregularity of ≅**

We understand in this way why it is obligatory to begin with A = $f_x(a)$ and B = $f_y(b)$; but the correct reading of the canonic formula is not 'A is to B what C is to D', but 'As well as A is to B is C to D', in order to point out, to make clear, that the theme is C:D, while A:B is only the phore. Let us observe that the usual, normal order is theme-phore, because we want to stress what the aim of the analogy is. Linguistically however we can reverse the order, for instance we can say 'what evening is with respect to the day is the oldness with respect to our life'; a similar thing happened with the Aristotelian proportion discussed above, which in Perelman & Olbrechts-Tyteca's version is: 'De même que les yeux des chauve-souris sont éblouis par la lumière du jour, ainsi l'intelligence de notre âme est éblouie par les choses les plus naturellement évidentes'; here the order phore-theme is not obligatory, we can replace it by the order theme-phore: 'L'intelligence de notre âme est éblouie par les choses les plus naturellement évidentes de même que les yeux des chauve-souris sont éblouis par la lumière du jour'. In contrast with these situations, the syntactic order in the canonic formula is obligatory A:B ≅ C:D and this may give the impression that A:B is the theme, but the real theme is C:D, while A:B is only the phore. As a matter of fact, just because C and D are defined by means of both A and B, it is not the movement from A:B to C:D (as in the usual analogies), but the passing from A:B to A:B ≅ C:D that is characteristic for the canonic formula. The difference between A and B, on the one hand, and C and D, on the other hand, is, as we have already observed, not one of nature, but one of complexity. The condition imposed by Perelman & Olbrechts-Tyteca on A and B, on the one hand, C and D on the other hand, to belong to different domains in order to be able to read A:B = C:D as analogy, is too restrictive. They are wrong in this respect when they refuse to give numerical proportions the status of analogy. Obviously, when A = 2, B = 3, C = 16, D = 24, there is no difference of nature between these four components; they all are numbers. But there is a difference of complexity between A and B on the one hand, C and D on the other hand, because A and B are prime numbers, while C and D are not.; so, the proportion A:B = C:D is asymmetric; A:B is an irreducible fraction, while C:D is not; so this proportion is an (a special case of) analogy, with A:B as phore and C:D as theme. As a matter of fact, we start with the theme and we try to reduce it to the phore. The non-trivial case of this situation occurs when C and D are very large numbers, whose decomposition into prime factors is a difficult problem (let us remember that the classical algorithm giving this

decomposition works in exponential time).

But accepting the essential asymmetry of the canonic formula we accept the singular situation of the relation ≅ in this formula; we read it as similarity and even as an equivalence, but rigorously speaking it is neither one nor the other. As a matter of fact, ≅ has no regularity property. It is not reflexive, because reflexivity would be in contradiction to the requirements for (A, B) and (C, D) to differ either in nature or in complexity; it is not symmetric, because we cannot start with C:D (we could do it if we would separate the definition of A, B, C, and D from the writing of the proportion). So, ≅ can be interpreted, strictly speaking, neither as a tolerance relation, nor as an equivalence relation. It is a kind of irreflexive binary relation, a unidirectional similarity (i.e., non-symmetric), because it has to indicate a metamorphosis, a transformation whose nature remains to be clarified.

## The canonic formula requires both a paradigmatic and a syntagmatic reading

All researchers of the canonic formula of myth have agreed that its enigmatic component is just the last one $f_{a^{-1}}(y)$. But now we realize that an equally enigmatic component is just that which never raised problems: the relation denoted by ≅. It was denoted by :: in Lévi-Strauss (1958), then this notation was changed in ≅ (Lévi-Strauss 1973), but in both cases, as the notation suggests, it was of thought as an equivalence relation, as is attested by the following quotation showing the way in which Lévi-Strauss interprets his formula: "Here, with two terms, a and b, being given as well as two functions, x and y, of these terms, it is assumed that a relation of equivalence exists between two situations defined respectively by an inversion of *terms* and *relations,* under two conditions: (1) that one term be replaced by its opposite (in the above formula a and $a^{-1}$); (2) that an inversion be made between the *function value* and the *term value* of two elements (y and a) (Lévi-Strauss 1963).

As we already pointed out, everyday language does not make a rigorous distinction between equivalence, similarity and analogy; this is the reason why the relation ≅ was alternatively interpreted as equivalence, similarity or analogy, giving the impression that these words are synonyms. But logic distinguishes them, as we have shown above and, as we have argued, none of them succeeds in capturing exactly the situation of ≅ in the canonic formula. To be more precise, equivalence (as a binary relation which is reflexive, symmetric and transitive) and similarity (as a tolerance relation, i.e. a binary relation which is reflexive and symmetric) are excluded as possible interpretations of ≅, while analogy is rejected if we keep rigorously to its meaning in 'Traité de l'argumentation' (Perelman & Olbrechts-Tyteca 1988), but could be accepted with an adequate modification of its meaning, as was shown above. But even with this



improvement, we feel that something essential concerning the interaction between the left and the right part in the canonic formula is not yet captured. One aspect of what is missing was discussed above: the components C and D are defined by means of the components A and B. But it is not only this, in other words, this is only a part of a more general feature of the canonic formula. A feeling of this feature has already been expressed in the literature. Maranda & Maranda (1971: 25) observe that analogy is essentially 'linear'; it cannot formalize the twists found in myths and which call for a 'non-linear' formalization. By 'linear' Maranda & Maranda mean just the symmetry property of the analogy, the fact that both of its parts are in the same plane. But more important is what follows. Lévi-Strauss already pointed out the role of the mediation process in myth (see, for instance, his book (1962) and chapter XI in Lévi-Strauss (1973)). This mediation is an essential part of the canonic formula, which should be understood, as Maranda & Maranda (1973: 26) point out, as the figuration of a mediating process where some dynamic roles are expressed more accurately than in a simple analogy model. In this formula (b) is the mediator; (a) is the first term, which expresses, in connection with the socio-historical context, a dynamic element (specifying function $f_x$) under the impact of which the item unfolds. The other function, $f_y$, which is opposed to the first one, specifies (b) in its first occurrence. Thus (b) is alternately specified by both functions, and thus can mediate opposites.

To put the above remark in a more explicit form and into a broader context, it has to be stressed that while analogy is of a paradigmatic nature, the relation ≅ in the canonic formula acquires its correct meaning in both a paradigmatic and syntagmatic context. The myth formula is, in each particular use, a class of stories and it captures the essence of this class. Let us remember Lévi-Strauss' assertion (1973, chaper XI): 'The substance of the myth is located neither in the style, nor in the syntax or in the narrative modality, but in the *story* it tells us'. The mediation process has to be understood in time, it is essentially a narrative structure and the myth formula proposes a segmentation of this structure. By mediation, the initial oppositions are attenuated. The result of the mediation is just the conversion of the term a in the function $a^{-1}$ (the inverse of a) and of the function y in the term y. The exact meaning of these assertions will be discussed later. But we can already assert that what is usually named the equivalence between A:B and C:D in the myth formula is a part of an iterative, generative process, which remains to be discussed.

## The components of the formula

In order to understand the nature of the transformations taking place when passing from the left to the right part of the formula, we need first to clarify the nature of the entities called *terms* and *functions.* Usually in mathematics the

terms are the components of a sum (finite or infinite). For instance, in expressions like $u + v$, $u_1 + u_2 + ... + u_n + ...$ we say that $u, v, u_1, u_2, ..., u_n, ...$ are the terms of the corresponding sums (finite in the first case, infinite in the second case, when the sum is called a *series*). It may happen however that one uses the word *term* for a component of an entity other than a sum; for instance, we speak about the terms of a finite or infinite sequence, such as $u_1, u_2, ..., u_n$ or $u_1, u_2, ..., u_n, ...$. A finite sequence of n terms is defined as a mapping of the set $\{1, 2, ..., n\}$ into a set A; an infinite sequence is defined as a mapping of the set $\{1, 2, ..., n, ...\}$ into a set A; so, what we call the term $u_i$ of a sequence is the value taken by the above mapping when the value of the argument is the natural number i. The terms can be constants or variables; in the first case the custom is to denote the terms by letters from the beginning (or the middle) of the alphabet, while in the second case we have to distinguish between terms representing independent variables (in this case the custom is to denote them by letters from the end of the alphabet: $x, y, z, x_n, y_n, z_n$) and terms represented by functions (i.e., dependent variables). Let us recall that the mathematical concept of a function is defined as a system of three objects $<A, B, f>$, where A and B are sets, while f is a rule which associates to each element x in A one element f(x) in B.

Does the above mathematical definition of a function fit with the way in which Lévi-Strauss is using the word *function*? The trap into which we are drawn is that, while the answer to this question is negative, the graphic appearance of the functional symbol f in the canonic formula would suggest that the answer is affirmative. Indeed, the notation $f_x(a)$ means, at first glance, the value the function $f_x$ takes when the argument is equal to a. However a doubt appears by observing that the symbol f is logically parasitic, although rhetorically, as the initial letter of the word *function*, it calls our attention to the idea of a function. Otherwise, no information is lost if the canonic formula is written by completely ignoring the symbol f:

$$x(a):y(b) \cong x(b):a^{-1}(y).$$

This last form particularly calls our attention to another possible way of interpreting *term* and *function* in this sense in logic. 'A class is a set of terms which can be substituted for each other and still be argument to the same propositional function without altering its truth value' (Maranda 1981: 32). So, a term is understood as something susceptible to being an argument of a propositional function. This fits with the interpretation of x and y as propositional functions and it also fits with expressions such as *function value* and *term value* used by Lévi-Strauss (see one of the above quotations). Speaking of *term value* we implicitly recognize that the term is a variable which can take various values, to which various values of the function will correspond. From all the examples analyzed by Lévi-Strauss it follows that for him *function* is equivalent to *role*,



while *term* has to be interpreted as *character* or *object*. But this fits completely with the above logical interpretation, because any role generates a propositional function and any character (or object) is a potential value of the argument of such a function.

But there is nothing in the logic of propositions enabling us to give a meaning to the 'inverse $a^{-1}$ of the argument a, with $a^{-1}$ interpreted as a propositional function' and to explain the capacity of a propositional function - in our case y - to be considered the argument of $a^{-1}$. These mysterious metamorphoses made Carroll (1977: 671) consider that most commentators of the canonic formula 'have never been completely certain of the meaning of this passage'.

## References

**Birkhoff, G.D.** (1933). *Aesthetic measure.* Cambridge, Cambridge University Press.

**Carroll, M.P.** (1977). Leach, genesis and structural analysis: a critical evaluation. *American Ethnologist 4, 663-677.*

**Gentner, D.** (1982). Are scientific analogies metaphors? In: Mial, D.S. (ed.), *Metaphors; problems and perspectives.* Sussex, The Harvester Press and New Jersey, Humanities Press: 106-132.

**Hage, P. & Harary, F.** (1983). *Structural models in anthropology.* (= Cambridge Studies in Social Anthropology No 46). Cambridge, Cambridge University Press.

**Huntington, S.** (1968). *Political order in changing societies.* New Haven, Yale University Press.

**Lévi-Strauss, C.** (1945). L'analyse structurale en linguistique et en anthropologie. *Word 1,2, 1-21.*

**Lévi-Strauss, C.** (1951). Language and the analysis of social laws. *American Anthropologist 52, 155-163.*

**Lévi-Strauss, C.** (1958) *Anthropologie structurale.* Paris, Plon.

**Lévi-Strauss, C.** (1963). The structural study of myth. In: *Structural Anthropology 1.* Garden City, New York, Doubleday: *206-231.*

**Lévi-Strauss, C.** (1964). *Mythologiques I: Le cru et le cuit.* Paris, Plon.

**Lévi-Strauss, C.** (1968). *Mythologiques: L'origine des manières de table.* Paris, Plon.

**Lévi-Strauss, C.** (1973). *Anthropologie structurale deux.* Paris, Plon.

**Lévi-Strauss, C.** (1985). *La potière jalouse.* Paris, Plon.

**Maranda, E.K. & Maranda, P.** (1971). *Structural models in folklore and transformational essays.* The Hague, Mouton.

**Maranda, P.** (1981). Anthropological analytics. In: Rossi, I. (ed.), *The logic of culture.* New York, Bergin: 23-41.

**Marcus, S.** (1969). Lingvistica, stiinta pilot. *Studii si Cercetari Lingvistice 21, 235-245.*

**Marcus, S.** (1974). Lingusitics as a pilot science. In: Sebeok, Th.A. (ed.), *Current Trends in Linguistics, vol. 12, 2871-2887.* The Hague, Mouton.

**Marcus, S.** (1985). *Timpul.* Bucharest, Albatros.

**Marcus, S.** (1990). Metaphor in science. A case study. *European Journal of Semiotic Studies 2,2, 231-238.*

**Marcus, S.** (1993). *Vers une approche transdisciplinaire du temps.* (to be published)

**Perelman, C. & Olbrechts-Tyteca, L.** (1988). *Traité de l'argumentation. La Nouvelle Rhétorique.* Bruxelles, Éditions de l'Université de Bruxelles.

**Petitot, J.** (1988). Approche morphodynamique de la formule canonique du mythe. *L'Homme 28 (2-3), 106-107: 24-50.*

**Wiener, N.** (1948). *Cybernetics, or control and communication in the animal and the machine.* M.I.T. Press and Wiley, New York.

**Zeeman, E.C.** (1962). The topology of the brain and visual perception. In: Fort, M.K. (ed.), *Topology of 3-manifolds and related topics.* Englewood Cliffs, N.J., Prentice-Hall: 240-256.

# The Distribution of Imaginistic Information in Oral Narratives:
## a model and its application to thematic continuity[1]

*Wolfgang Wildgen, Bremen*

**Abstract**

In prior research we have developed formal tools for the description of imaginistic content in oral narratives (cf. Wildgen 1990 and forthcoming). The purpose of this paper is to show, how this "information" is sequentially organized. We distinguish between semantic and syntactic information and analyse the "unification" of these basic types of information in a sentential representation, The model is applied to a short oral narrative of the Bremen corpus and we are able to show patterns of thematic continuity and thematic dynamism accessible with the aid of the tool developed in the first sections.

### 1. Language and the flow of information

In this paper we will apply the notion of information developed in Dretske's book *"Knowledge and the flow of information"* and deepen the notion of information against the background of actual dynamic theories (cf. Wildgen & Mottron 1987)

Dretske (1986) considers 'meaning' and 'information' to be two different concepts.

> "Typically, of course, we communicate with one another, exchange information, by exploiting the conventional meaning of signs. We convey information by using signs that have a meaning corresponding to the information we wish to convey. But this practice should not lead one to confuse the meaning of a symbol with the information, or amount of information, carried by the symbol.
>
> According to this usage, then, signals *have* a meaning but they *carry* information. What information a signal carries is what it is capable of "telling" us, telling us *truly*, about another state of affairs. Roughly speaking, information is that commodity capable of yielding knowledge, and what information a signal carries is what we can learn from it " (Dretske 1986: 44).

[1] This paper is a partial version of chapter 8 in the monograph (in preparation): A Systemic Image and Process centred Model of the Meaning of Sentences and Narrative Texts.

The integrative feature of Dretske's notion of flow of information is that the different sub-channels leading to knowledge from the external causes (the world), through sensation and perception up to communication are treated as *one* continuous flow of information. The specific individual character a message has at the different stages is relevant for meaning but not for (semantic) information. The material realization of the channels is also irrelevant; what is relevant are the dependency relations between the sender (s) and the receiver (r).

The amount of information received is related to the amount of information sent via the difference called either "noise" or more specifically 'environment' (ibidem: 19, 49). These ideas, which in a more qualitative manner take up concepts of classical information theory, may be applied to the problem of the multiple channels which relate the event and the narrative.

A specified flow of information would be as that illustrated in table 1.

Table 1
A set of levels in the transformation of information on the world.

| distal | proximal | cognitive | linguistic | |
|---|---|---|---|---|
| object/--> event | object/--> event | coding--> of the object/ | coding--> of the meaning event | utterance |
| world (a) | sensory field (b) | mind (c) | language (d) | (e) |

The imaginistic representations of narrative units (i.e. those telling the kernel of the story, what happened) is one stage in the informational flow; we have described its syntactic and semantic regularities (in Wildgen 1990 and forthcoming). The next stage is the coding of the same information in terms of a conventional system of phrasal and sentential patterns. Although at this micro-level the existence of gestalt-like pictures formalized in our imaginistic syntax is far less plausible (the production is too quick and combinatorially too complex), the *same* information which is given as an imaginistic unit must be preserved in a different code, i.e. we must be able to find the traces of the imaginistic gestalts and must explain how highly combinatorial and discrete devices can code the holistic entities.

The solution to this problem necessitates two stages:

- breaking down the holistic information into quanta (features) and their syntax,
- finding the traces of the imaginistic gestalt in the new coding. If the formal model of subsentential syntax mirrors the real appearance of these structures as quanta, classes, groups and rule-governed structures, this should not hide the fact that we consider it to be a phase of one and the same 'flow of information'.



The format in which the flow of information at the subsentential level (sentence, clause, phrase, word structure) is formulated should correspond to the long traditions of grammar writing since antiquity, to descriptive devices of modern structuralism and to the core of the generative descriptions formulated since Chomsky (1957). It seems that an attribute-value-grammar (AV-grammar) corresponds roughly to this desideratum. It uses a neutral, classical, formal language, a kind of Aristotelian hierarchy of 'genus proximum' and 'differentia specifica' and knows only one principle of composition called 'unification'. We shall not go beyond this level of technicality and shall try to show how the information contained in narrative units is distributed over word classes and syntactic devices. Our main objective will be to guarantee that the discretization and putting into pieces of the narrative unit does not lose the spatial and dynamic contents. The static character of the descriptive tool should not be a filter of dynamic information, as we believe most formal grammars are in fact. Thus although our description is conservative in its main techniques, it is innovative in the concern for dynamic and spatial information carried by this specific channel.

The general framework of an informational view of the cognitive system and language advocated by Dretske (1986) has been used for the construction of a grammar since the mid-eighties. The basic ideas of such an application can be summarized as follows:

a) All linguistic information organized in a grammar gets the same data-format. It is a list of attributes and their values. As values may be again attributed, hierarchical graphs of attributes and their values are allowed. The basic restriction is that these graphs be acyclical.[1]
b) The basic operation which controls the concatenation of strings and the construction of phrasal and sentential wholes is called "unification". It consists in a matching of the feature lists (attributes with their values) which are contributed by the elements of the construction in question.

[1] Cf. Pollard & Sag (1987: 28): "intuitively, a feature structure is just an information-bearing object that describes or represents another thing by specifying *values* for various *attributes* of the described thing,... Feature structures are standardly notated by attribute-value matrices (AVM's)"

The motivation for the first principle may be technical (computational simplicity). Thus Shieber (1986: 10) argues:

> "In fact, viewed from a computational perspective, it is not surprising that so many paradigms of linguistic description can be encoded directly with generalized features/value structures of this sort. Similar structures have been put forward by various computer scientists as general mechanisms for knowledge representations and data-types."

The second principle may also be seen as a simple computational device. Its linguistic signification could therefore seem to be rather small and even to contradict the intuitive ideas of IC-structure, valence-patterns, anaphoric processes and other rather global structures. In the 'Construction Grammar' of Fillmore and Kay a deeper linguistic claim is put forward which puts flesh on the technical notion of 'unification'. "The grammar contains a set of constructions. A string of words is a sentence of the language if and only if it can be assembled and given an interpretation by unifying a subset of this set of constructions, and a sentence is as many ways ambiguous as there are distinct assemblies of constructions which it can be shown to realize" (Fillmore & Kay, 1987:29).

Thus a specific grammar may be seen as a kind of lexicon of constructions, and syntactic structures are the result of an interaction between a lexicon of words (morphological constructions) and a lexicon of syntactic constructions (both are specific for one language but contain some invariant core which may be approached by comparative methods). This type of grammar is, however, descriptive and does not start with a priori principles. In a sense it returns half-way to American descriptivism (taking profit of the sophisticated techniques developed in generative grammar).[1]

So far we have discussed the general integration of our model into the current development in the domain of formal grammars. Our problem, however, goes beyond these grammars. We have to ask the question: How is 'imaginality', if it exists in lexical items (morphemes and morphological constructions), transported by unification and how can global imaginistic representations at the level of narratives (at the textual level) come out of this information and these constructional devices?

With the background of Constructional Grammar in mind we can advance two more specific subquestions:

Q1: How is imaginistic information encoded in the lexicon (mainly in the lexicon of verbs and morphological constructions on the level of the verb)?

[1] Fillmore and Kay (1987: 29) say that: "Unification based Construction Grammar is a species of generative grammar".



Q2: What are the types of constructions which allow the emergence of imaginistic gestalts based on local information in the constituents (cf. Q1)?

In this perspective we assume that the process of unification is a mental device, which amalgamates information and filters out grammatical combinations (eliminating ungrammatical ones).

In answering the first question we can report results of current research based on a corpus of 50 narratives, the second question can only be given a preliminary answer.

## 2. The encoding of imaginistic information in the lexicon of verbs and some basic constructions of the verb

In order to encode information in the framework of an attribute-value system, we must categorize the information which was encoded in our imaginistic unit (cf. chapter 5 and Wildgen 1990). The information may be subdivided into:

- lexical (semantic) information,
- syntactic information (word categories, phrasal structures, sentential structures).

In the case of verbs, which will be analyzed in this chapter, the lexical category of the word is 'V = verb'. Although the verbal prefix is separately marked, the analysis is given for the whole lexical entry. We distinguish seven attributes and their values.[2] The first order values of DP presuppose a classification of semantic roles, which is fully exposed in Wildgen (forthcoming) and will be summarized here.

The different semantic roles are characterized by the system of dynamic configurations. The possible dynamic configurations are nested and hierarchically structured . We distinguish:

- primary agents (they are the support of the process and do not disappear in the process): A, P

- secondary agents (they appear and disappear in the process): I, B.

[2] The design of a computerized data-base was developed by Dr. Joachim Liedtke, the classification was done by the research group under the direction of the author in the frame of the research project "Erzähldynamik". We thank the German research association (DFG) for its financial support.

The "casemes" defined by our configurational criterion are called:

- A (Agent) - P (Patient) (primary roles)

- I (Intermediary) - B (Binding force) (secondary roles).

The label I summarizes a plurality of forces which are linearly intermediarte between A and P. Dependent on the domain of interpretation it can be a path (interlocal locomotion), a metastable phase on a quality scale (quality space), an instrument (action space), or an object (change of possession).

The role B (binding force) has a rather variable realization. Configurationally it is an intermediary force parallel to the primary sequence A-I-P. It, therefore, calls for a second dimension in state space (cf. its derivation in C.T.-semantics from the umbilics, Wildgen 1982a: 85-92). It can be parallel to A (a helper of the agent), to P (a beneficiary of the event) and to I (a secondary instrument, a medium of exchange).[3]

Table 2 gives the attributes and their values.

In the data-base the values for the attributes are specified for every verb-token in the corpus of narratives. In table 3 and 4 we show two specimens of our data-base; first 8 verbs (ordered alphabetically by their stems) are analysed, then 8 verbs are treated in the alphabetical order of their prefixes.[4] If more than one feature is present, the abbreviations are written one after the other.

Example: 'ab-holen': DP = iq . The type of motion of the agent is interlocal (i) and quantitative/qualitative (q).

As this table is simply the beginning of a large list (with 2500 items) it was not selected in order to show very simple or very interesting examples. We shall, therefore, comment on some problematic classifications in table 3.

If 'veralbern' has the roles A and P and the second order values *sensual* for A and *sensual* for P, this means that A and P are emitter and receiver in a communicative interaction. The pair f(A, P) = (s,s) which is the basic configuration of communication is defined by the reciprocity of perceivable signs and their reception. In this example no specific message is implied, therefore I is not specified.

[3] A mathematically explicit analysis is given in Wildgen (1985a: 208-212). The compactified elliptical umbilic has a two-dimensional behaviour-space and shows consecutively (along a linear path through the bifurcation set) the configurations: (A, B, I), (B, I, P). The attractor B disappears and I (object) is caught by P. The underlying schema is that of giving: (A ,I ,P).

[4] TX = text (number in the corpus of oral narratives analyzed in Bremen), TS = clause (German: "Teilsatz"), i.e. its number in the sequence of clauses into which the text is decomposed in our corpus.



Table 2
A list of attributes and values for the analysis of verbs

| Attribute | values (1st order) | | values (2nd order) | |
|---|---|---|---|---|
| Dynamic pattern | sem. roles: | | category of substrata | |
| DP | agent: | A | interlocal: | i |
| | intermediary f.: | I | local: | l |
| | binding f.: | B | sensual: | s |
| | patient: | P | mental: | m |
| | | | quanti./qualit.: | q |
| | | | abstract: | a |
| | | | state: | st |
| | | | position: | p |
| processual phase: PP | ingressive: | ig | | |
| | egressive: | eg | - abrupt: | ra |
| | self-refer.: | sr | - successive: | rs |
| | resultative: | re | | |
| | limit-line: | ll | | |
| | limit-surface: | ls | | |
| direction: DI | temporal: | te | | |
| | spatial: | sp | | |
| | input: | in | | |
| | output: | ou | | |
| kind of motion K | special type: | | se | |
| | in all other cases: | | no | |
| force F | big: | bi | | |
| | small: | sm | | |
| verbal mode VM | negation: | ng | | |
| | question: | qu | | |
| | subjunctive: | sj | | |
| | let do sth.: | ld | | |
| | want to do sth. | wd | | |
| | can do sth.: | cd | | |
| | must do sth.: | md | | |
| | should do sth.: | sd | | |
| evaluation: E | positive: | po | | |
| | negative: | ne | | |

Table 3
Analysis of the first verbs ordered by their stem

| prefix | stem | translation | A | I | B | P | Further attributes PP | DI | K | F | VM | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **ahnen** | to suspect | m | a | Ø | Ø | sr | | | | | |

Na, und da hab'ich schon Schlimmes geahnt.
Well and then I already *suspected* something terrible

| prefix | stem | translation | A | I | B | P | Further attributes PP | DI | K | F | VM | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ver** | **albern** | to make fun of | s | Ø | Ø | s | eg | | se | | wd | ne |

nu, die wollen sie veralbern!
and they want to *make fun of* her!

| prefix | stem | translation | A | I | B | P | Further attributes PP | DI | K | F | VM | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **ärgern** | to be annoyed | m | Ø | Ø | Ø | sr | | se | bi | | ne |

und denn hab'ich mich so geärgert,
and then I was so *annoyed*,

| prefix | stem | translation | A | I | B | P | Further attributes PP | DI | K | F | VM | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **ärgern** | to annoy | s | Ø | Ø | Ø | | | | bi | wd | ne |

daß einer denn mich ärgern will,
that someone wants to *annoy* me,

| prefix | stem | translation | A | I | B | P | Further attributes PP | DI | K | F | VM | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **rum** | **ballern** | to ring out | s | Ø | Ø | Ø | eg | sp | | | bi | ne |

daß meine Reifen so rumballerten,
that my tyres were *ringing out* so much

| prefix | stem | translation | A | I | B | P | Further attributes PP | DI | K | F | VM | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **aus** | **bauen** | to remove | l | Ø | q | i | ll | sp | | | | |

und hat ihn ausgebaut
and *has removed* it



Table 3. Continuation

| prefix | stem | translation | A | I | B | P | Further attributes PP | DI | K | F | VM | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ein** | **bauen** | to put in | l | Ø | q | i | ll | sp | | | | |

Naja, und wir haben die denn auch eingebaut,
Well, and then we also *put* them *in*,

| prefix | stem | translation | A | I | B | P | Further attributes PP | DI | K | F | VM | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **befreien** | to rescue | l | Ø | Ø | q | re | | | | sj | po |

er hätte sie auf elegante Art befreit
he *rescued* her in an elegant manner

In the first example we have with 'ärgern' a mental process which is emotional. The processual phase is sr (self-referential) as the verb is reflexive. In the next token with 'ärgern' a patient (P) is introduced. The emotional process is located in P whereas A makes a perceivable (probably a communicative) action, A has the processual type s (sensual).

In 'rumballern' the tyres produce a perceivable noise (processual phase, egressive is in this case emissive). The semantic role A takes the value s (sensual).

The two tokens of 'ausbauen' are interesting as their number of roles exceeds the number of syntactically realized noun-phrases. In both cases the binding force is not specified by a noun phrase; it is coded by the verb. In the first example the subject has been eliminated by subject ellipsis, a regular process in German coordinative constructions. The values of the roles (A, B, P) are: (local, qualitative, interlocal). In each case the engine (P) is moved interlocally, whereas the agent (A) makes movements with his limbs (local), the car (B) which is the background (a secondary participant) of the process is changed qualitatively; it loses ("ausbauen") or recovers ("einbauen") an important piece.

In 'befreien' the agent (A) moves locally (with his arms), the patient is freed (changes qualitatively). This specimen shows that a detailed analysis of the example and its meaning in context is necessary in order to classify exactly the predicative centres of a narrative. For the construction of imaginistic representations the attributes - dynamic pattern (DP), processual phase (PP), the first and second order values of DP and the first order values of PP - are of central importance. The other features complete the qualitative picture given by DP and PP.

Table 4
Analysis of some verbs by their prefix

| prefix | stem | translation | A | I | B | P | Further attributes PP | DI | K | F | VM | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ab** | **bekommen** | take off | I | Ø | Ø | q | re | | | | | |

also quasi dann -äh- den -äh- Mantel abzubekomm',
then to *take off* the coat

| prefix | stem | translation | A | I | B | P | Further attributes PP | DI | K | F | VM | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ab** | **düsen** | to go off | i | Ø | Ø | Ø | eg | | | se | | bi |

Ich wieder abgedüst
I *went off*

| prefix | stem | translation | A | I | B | P | Further attributes PP | DI | K | F | VM | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ab** | **fahren** | to drive by | i | i | Ø | Ø | | | | | | |

und -äh- mit dem hab'n wir einige Annoncen abgefahr'n,
and with this one we have *driven by* several advertisements,

| prefix | stem | translation | A | I | B | P | Further attributes PP | DI | K | F | VM | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ab** | **holen** | to pick up | iq | Ø | Ø | i | re | | | | sd | |

er soll den Trümmerhaufen mal abholen da vorne, ja,
he should *pick up* that heap of rubble over there, yes,

| prefix | stem | translation | A | I | B | P | Further attributes PP | DI | K | F | VM | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ab** | **lösen** | to take off | I | I | q | I | re | sp | | | | |

und mit/ und löst dann mit der anderen Seite quasi die, die Decke oder den Mantel ab,
and *takes off* with the other side the cover or coat,

| prefix | stem | translation | A | I | B | P | Further attributes PP | DI | K | F | VM | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ab** | **machen** | to take off | I | Ø | Ø | q | ll | sp | | | md | |

quasi ohne daß ich groß -äh- da den Mantel abmachen mußte,
without having to *take off* the coat.

| prefix | stem | translation | A | I | B | P | Further attributes PP | DI | K | F | VM | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ab** | **schleppen** | to take in tow | i | i | Ø | i | | | | | | |

ADAC schleppt ab, also bis zur Tankstelle,
ADAC *takes us in tow* , you know down to the filling station,



Table 4. Continuation

| prefix | stem | translation | A | I | B | P | Further attributes PP | DI | K | F | VM | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ab** | **springen** | to jump off | i | Ø | Ø | Ø | ll | | | bi | | |

da is'mir die Kette abgesprungen,
when my chain *jumped off*,

Tables 3 and 4 give examples from the first pages of the classification which contains all verb-tokens in the corpus of 50 oral narratives. The examples we gave show that the image- and process-centred information contained in the components developed so far can be integrated in the format of an attribute-value table. All the techniques and technical devices available for grammars of the unification type can therefore be applied. As such an application is not the concern of this article, we shall only exploit these techniques insofar the dynamics of texts is concerned.

## 3. The hierarchical organization of the information and its unification

If we consider all the classificatory specifications in our lexicon of verbs and verb-centred constructions as the central information in a narrative unit we can represent every narrative by a sequence of classificatory vectors (i.e. sequences of classifications as shown in table 3 and 4). The labels are those of table 2. Every vector contains information which contributes to the characteristics of a unit. We can distribute the attributes of table 2 into one semantic and one syntactic set. The syntactic set is elaborated and contains new attributes. The basic attribute VP (valence pattern) has as values the set of complements of the verb (in the sense of valence theory): The subject (1), the direct object (accusative = 2), the genitive object (3), the indirect object (dative = 4), the prepositional object (5), etc. (cf. Engel 1988: 185-198) and the combinations of these elements: (1,2), (1,3), (1,2,3), etc. The sequential order in a specific sentence is given by attribute O (order). The attributes belong to different hierarchical levels; the underlying relation of order is symbolized as ">". On the basis of the hierarchical structure we can build structural trees. The attributes F, E and VM, O are global attributes and directly dependent on SEM and SYN respectively.

Table 5
Semantic attributes and values

| **attributes A** | | **Attributes B** | | | | |
|---|---|---|---|---|---|---|
| **DP >** | **Cat. (C)** | **(PP) >** | **(DI)** | **(K)** | **(F)** | **(E)** |
| A | i | ig | te | se | bi | po |
| I | l | eg | sp | no | sm | ne |
| B | s | sr | in | | | |
| P | m | re | ou | | | |
| | q | ll | | | | |
| | a | ls | | | | |
| | st | | | | | |
| | p | | | | | |

Example of a simple hierarchical tree (horizontally arranged):

**"run"** analysed as :

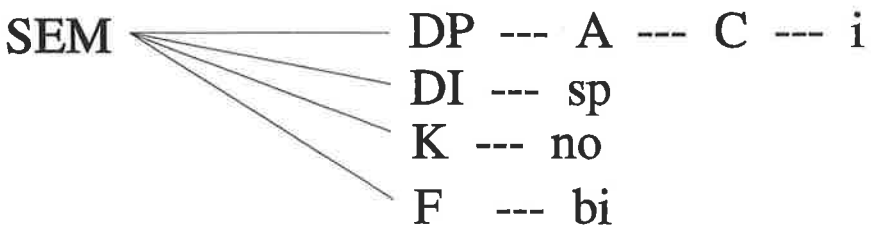



Figure 1. The attribute-value tree for "run"

The attributes VP, O, TY and PR are new and represent specifically the surface structure of the sentence. Thus the verb 'wegschlagen' (to beat off) in the narrative unit "schlag die Decke weg" (knock the coverlet (off the bed)) has two vectors: a semantic and a syntactic one:

-semantic vector: SEM(DP(A(l), P(i)), PP(ll), D(sp), K(0), F(bi), E(0))
-syntactic vector: SYN(VP(1, 2(TY(np))), O(0,2), VM(0))[5]

[5] DP(A(l), P(i)) means that the attribute DP has the first order value A(agent) and P(patient) and the second order value l for A and i for P; VP(1,2) means that the verb is bivalent with fillers of type 1 = nominative noun phrase and 2 = accusative noun phrase.



Table 6
Syntactic attributes and features

| **Valence pattern(VP)** | **Type(TY)(second order attribute)** | |
|---|---|---|
| 1, 2, 3, 4, 5, | np (noun phrase) | pr (pronoun) |
| (1,2), (1,3),... | vp (verb phrase) | ad (adjective) |
| (1,2,3) , ... | ap (adjective phrase) | n (noun) |
| | se (sentence) etc. | |

| **third order attributes** | **global attributes** | | |
|---|---|---|---|
| **preposition(PR)** | **verbal mode(VM)** | | **order(O)[6]** |
| on | ng | qu | (1) |
| at | ld | sj | (1,2) |
| in | cd | wd | (2,1) |
| | sd | md | (5,1) |

The semantic vector says that the verb (in a specific sentence token) has two semantic roles (Agent and Patient) and that the first has the category of process called *local* (l) as the beating implies a movement of the limbs, the second has the category *interlocal* (i) as the coverlet is moved in space, the processual phase (PP) is characterized by the transition through a limit line (ll) which separates a closed and an opened coverlet, the type of directivity is spatial (sp) and force (F) is big (bi). The attributes Kind (K) and Evaluation (E) are not specified (0).

The syntactic vector says that the valence of the verbs is of type [1 (= subject), 2 (= object)]; the order is (0,2), i.e. we have a case of subject ellipsis, the type of constituent for the object (2) is noun-phrase (np); the attributes PR and VM are not specified. As the constituent 1 (subject) is not realized, its type is not specified.

If we introduce SEM (semantic) and SYN (syntactic) as basic attributes we get a tree, which can be represented by a bracketing. Its maximal depth is given by a vector like:

(SEM (DP (A(i)))) and (SYN(VT(1(TY(np(PR(in)))))))

[6] We start from a standard order given by the valence pattern. **O** is only specified if this order is inverted or if a position forseen in the standard is empty. In the latter case we simply insert the VP value : 0, cf. table 7.

The notation gives first the attribute of the highest type, e.g. DP(A) and then in parentheses attributes of a lower type or so-called features.

Every unit of the narrative may be represented by the semantic and the syntactic vector of its verb. The sequence of units (i.e. the text) is thus represented by a sequence of pairs of vectors (the semantic and the syntactic vector of a unit). Thus every text may be transformed into a sequence of vector-pairs and we can develop an analysis of the text, based on the information contained in this sequence of vector-pairs. If the phonological shape is important for a text (e.g.. a poetic text), the vector pair may be elaborated by a phonetic/phonological vector to a vector triple. The sequence of the phonetic/phonological vectors would allow to describe rhymes and other phonological properties in sequences of units. For specific analyses we can also concentrate on the syntactic or the semantic vector alone or on partial vectors. As an example we shall analyse the first narrative of our corpus, which is the shortest one (as the corpus is ordered by the length of the narratives).

(1) Ich bin einmal über'n Rembertiring gefahren

(2) da is' mir die Kette abgesprungen

(3) und denn -äh- is' die Luft auch noch 'rausgegangen

(4) und denn hab' ich mich so geärgert,

(5) daß ich das Fahrrad denn weggeworfen hab'

Simplified translation (without particles and hesitations)

(1) Once I drove across the Rembertiring (street in Bremen)

(2) Suddenly my chain fell off

(3) and then even the air escaped

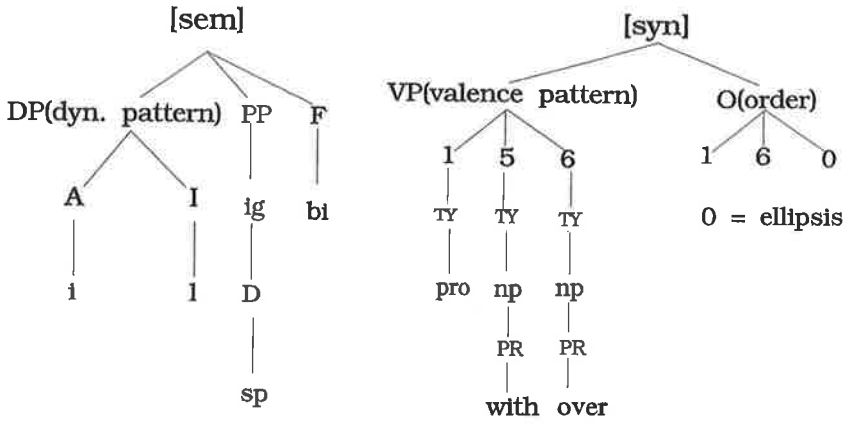
(4) and then I became so angry (at myself)

(5) that I threw away my bicycle.

If we use the more involved graphical notation, we can represent the semantic and syntactic information contained in the verb (as used in the text above) by separate attribute-feature trees and their unification. The unification mechanism is only sketched here. As a technical background to our notation cf. Pollard & Sag (1987). In Fig. 2 we sketch the attribute-value-structures of the contextually



completed sentence: "I drive on a bicycle across the Rembertiring". The global attribute O indicates that in the realized sentence the prepositional phrase (5) has been eliminated.




**drive (I, with the bicycle, over the Rembertiring)**

**completed unit**: [(ich)(mit dem Fahrrad)(über den Rembertiring)]
**realized unit:** [(ich)(über den Rembertiring)]

Figure 2. AV-analysis of the completed unit and its realization

The semantic and the syntactic information associated with the verb 'fahren' can now be unified, i.e. we calculate a unification which decides:

1. If the dynamic pattern (DP) fits the valence pattern (VP). If this is not the case, we can consider several options:

   1.a. A syntactic valence is not role-relevant (in general we assumed that the values 6 to 11 of VT are not role-relevant).

   1.b. A syntactic function is role-relevant but not interpreted by a role.

   1.c. A semantic role is not syntactically realized. It may be incorporated in the verb or in other constituents or be implicit in the context.

2. In those cases where no restrictions on coinstantiation and agreement are present, the process of unification cumulates the information which does not require a matching.

We shall describe these operations for the short narrative cited above. In order to focus on the contrast between the DP-constituents of the [sem]- and VP-constituents of the [syn]-component we shall only use a partial representation and unification procedure.

In sentence (1) the semantic role I is left to the context, we can complete "mit dem Fahrrad" (with the bicycle). The valence governed NP (Directional) is not role-relevant.

In sentence (2) the roles in DP and the constituents in VP match, i.e. DP and VP are saturated and no valence is left without semantic interpretation. Whereas the first example shows the phenomenon of partial unification, where DP(A(i)) is unified with VP(1), and DP(I(i)) is left unsaturated, the second example saturates the DP-component completely. It transmits processual information to the nominal/pronominal components. The order of principal constituents is, however, inverted: (A,B) --> (B,A).

In (3) the Patient (the tyre of the bicycle) is not mentioned. It can be inferred from knowledge about partinomic relations in the field 'bicycle' (cf. Fig. 6).

In (4) the DP(A(m)) is unified with VP(1), but for VP(2) in [SYN] there is no role-specification in the dynamic pattern (DP) of the verb. It is an effect of the self-referentiality of mental verbs that the agent can appear twice, once as the origin and once as the destination of the process.

In the last sentence the semantic roles (A, P) correspond to the role specified in the syntactic component. A is marked by the nominative(1), P by the accusative(2). The further features - local motion (l) by the agent and interlocal motion (i) by the patient - are taken over into the unified expression.

In a much longer narrative (67 narrative units) Liedtke (1990: 240) found a rather complicated distribution of participants. Most of the participants in his list belong to one of the following classes:

a1) The protagonist, who is at the same time the narrator (I / Uwe (Christian name) / my leg / we), N = 30.

a2) The helpers (friends of the protagonist (friends, parents of the friends / people / we), N = 5.

b) The antagonist, the spider on his bed (house spider / beast / my special friend / the big one / the mistress), N = 10.

c) All other participants have a much lower frequency. We can, however, single out a network of spatial frames, ordered by spatial embedding:



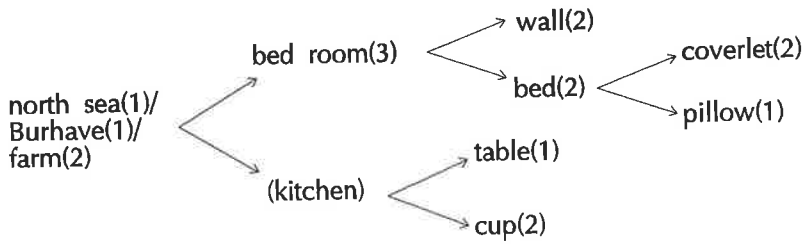



Figure 3. The network of spatial frames (number of tokens)

In total the network contains 10 items (one element 'kitchen' is implicit) and 17 occurrences. If we unify a1) and a2) in *one* category, we get a very classical picture:

- Two major agents; one protagonist, the other antagonist,

- the spatially organized frame, the scenery in which the struggle between the main agents occurs. The overall scheme is that of capture:

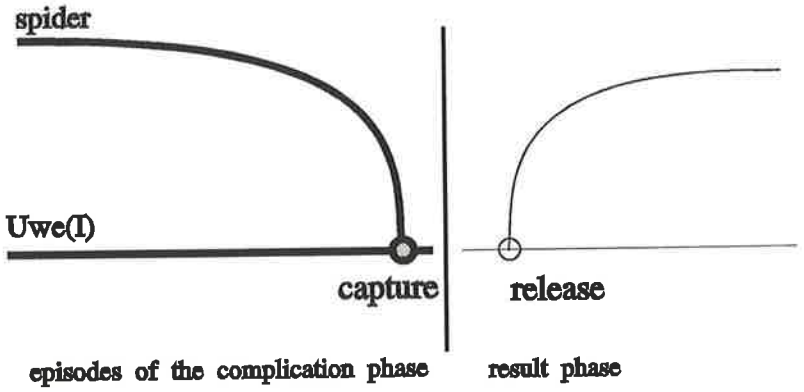



Figure 4. The overall schema of the narrative

In the last three units (65-67) the captured spider gets its freedom again.

In our context the distribution of explicit, lexical information over the three major thematic components (cf. above) is of primary interest. For this purpose we propose a rough scale of lexical (informational) weight.

0: implicit reference (in context, by ellipsis),
1: repeated pronominal reference,
2: repeated nominal reference,
3: initial / new mention by nominal / pronominal reference of name,
4: complex description by a nounphrase with at least one attribute (adjective, relative clause).

The following list shows the sequences for the categories $a_1$, $a_2$, b, and c mentioned above:

$a_1$) Protagonist: ich (3), ich (1), ich (1), (1), (1), (0) (0), (0), (0), (0), (0), (0), (0), (1), (1), (1), (1), (1), (1), 0, 1, 1, dich (1), ich (1) (1), Uwe (3), ich (1), (0), (1), wir (1). Sum of weights: 24, relative weight: 0,8 (N = 30).

$a_2$) Helpers of the protagonist: bei Freunden (3), die (1), den Eltern, meinem Freund (4), alle Leute (3), sie (1). Sum of weights: 12, relative weight: 2,4 (N = 5).

b) Antagonist: fette, dicke Hausspinne (4), das Vieh (3), sie (1), sie, groß und breit (4), die (1), sie (1), sie (1), sie, die Große (4), die Dame, die (4), die (1): Sum of weights: 24, relative weight: 2,4 (N = 10).

d) Spatial frames

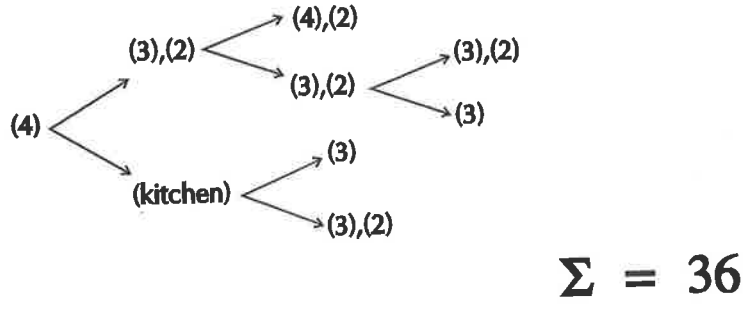



Figure 5. The distribution of weights in the network of spatial frames
Sum of weights: 36, relative weight: 2,77 (N = 13).

The absolute weight of the four categories exhibits the following order:



The scale of relative weights shows that the protagonist, who is simultaneously the narrator has by far the lowest mean weight. The highest weight is given to the spatial frames:

spatial frames (2.77) > antagonist = helpers of the protagonist (2.4) > protagonist (0.8).

Summarizing, we can say that in terms of frequency the protagonist dominates the narrative (35 occurrences including the helpers); the spatial frames (15) are even more often mentioned than the antagonist (10). In terms of total semantic weight, the spatial frames are dominant and the protagonists and antagonists have identical values.

The relative semantic weight shows very low values for the protagonist who is the narrator (0,8) and a much higher weight for all other participants. The higher values of spatial frames are mainly due to the hierarchical organisation of the frames in this story (which leads to a rich lexical elaboration) and to its descriptive character.

As Liedtke (1990: 242f) has shown, the occurrence of new participants is not equally distributed over the narrative. If we consider different episodes of the story, most new participants are found in the introductory phase of every episode.

In general we can see how the processual information contained in the verb (in its contextual reading) is distributed over the verbal construction, which is the basic sentence scheme. Further information on time, space and manner, in the form of conjunctions and sentence adverbials may be added. The sentential mechanism of the German language is, however, not our primary concern here. Our objective was to show that imaginistic information can be handled with the mechanisms of attributive-value notation and the syntactic filter called 'unification'.

## 4. Thematic coherence and thematic dynamics in narrative texts

The sketch of an attributive-value description which organizes the imaginistic information is not very far-reaching. Nevertheless such a simple mechanism allows interesting analyses of thematic coherence and thematic dynamics in a narrative text. We shall only deal with the sequence of the [SEM] attribute-value structures. A parallel analysis can be made on the basis of the [SYN]-vector, if we add the feature V (central verb) to our notation.

We can transform the text of narrative units into a sequence of semantic vectors with their lexical fillers (including the contextually or lexically realized ones).

1) SEM[DP(A(i), I(i)), V]
ich (Fahrrad) fahren

2)SEM[DP(A(i), P(q)), V, PP(l), D(sp)]
Kette mir abspringen

3)SEM[DP(A(i), P(q)), V, PP(l), D(sp)]
Luft (Fahrrad) rausgehen

4) SEM[DP(A(m)), V, PP(sr), F(bi), E(ne)]
ich mich ärgern

5) SEM[DP(A(l), P(i)), V, D(sp)]
ich das Fahrrad wegwerfen

We can consider the 'bicycle' and its parts as a partinomic network of the following shape:

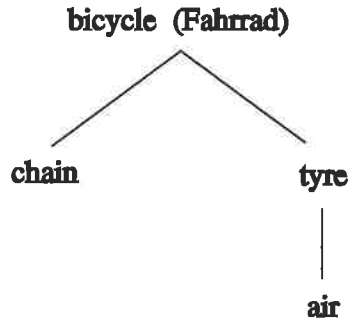



Figure 6. The partinomic network of 'bicycle'

The narrative takes its major roles from two sources:

a) the hero (in this case the narrator himself); we abbreviate as HE

b) the bicycle and its parts (we abbreviate as BI).

The distribution over the sequence and the roles is given in table 7.



Table 7
The distribution of roles and values in the narrative

| | Agent | Instrument | Patient | Hero | Bicycle |
|---|---|---|---|---|---|
| 1 | He(i) | Bi(i) | --- | (i) | (i) |
| 2 | Bi(i) | --- | He(q) | (q) | (i) |
| 3 | Bi(i) | --- | Bi(q) | --- | (i), (q) |
| 4 | He(m) | --- | He(m) | (m) | --- |
| 5 | He(l) | --- | Bi(i) | (l) | (i) |

The bicycle which was first an (implied) instrument, becomes an agent in (3) (implicitely already in 2). In (4), however, the hero comes back on the stage, but with a change in processual level (PP), from spatial (i) to mental (m); finally the hero solves the conflict, becomes (spatial) agent again, the bicycle is a (spatial) patient of the final scenario. Thus the situation in (1) is reestablished, in the sense that **He** dominates the stage. The transitions between initial and final situations are smoothened by the following structures.

- The prominence of **Bi** in (2) is prepared by the (implied) instrumental role of **Bi** in (1),
- the processual level of **He** is lowered (from i to q) as **He** becomes Patient and thus is downgraded,
- in (3) **He** is not present (we can infer that **He** is the (terrified) observer),
- in (4) the role of **He** as the observer is put into the foreground, and his emotional reaction is described.
- Finally in (5) **He** takes the initiative again and solves the problem by force.[7]

Syntactically the sequence is more complicated; additional entities are introduced (mainly for the orientation in global space and time), some roles are contextually interpreted (bicycle in (1) and tyre in (3)), and motivational links

[7] If you ask why **He** throws the bicycle away, my everyday experience (in Bremen) tells me that he had just stolen it before but I cannot garantee for this explanation.

are introduced between (4) and (5) such that (5) is syntactically subordinated to (4).

## 5. Imaginistic dynamics and imaginistic coherence

In the last sections our procedure was still conservative. Although we considered dynamic and action-oriented information, the format of our description remained that of attributes and their values. The coherence, namely of the text, and its dynamics were represented as patterns of repetition. We shall briefly point to further elaborations which are contained in a monograph in preparation called: "A Systematic Image and Process Centred Model of the Meaning of Sentence and Narrative Texts".

We can start with a vocabulary of imaginistic units classified by their valency **v** (v = 1,2,3). In Wildgen (1990) twenty units were derived from a basic list of vectors (in semantic space-time). They all use a normed two-dimensional space-time domain (t x r); t = (0, 1), r (-1, 0, +1); unit vectors which fill the two-dimensional matrix of a narrative text are constructed on the basis of (t x r) unit-squares. In Fig. 7 the basic list of vectors in space (t x r) is given. The Figures 8a, b, c show the sublists of mono-valent, bi-valent and tri-valent imaginistic units. The moves of the protagonist on r (semantic space) are positive, and the moves of the antagonist on r are negative.

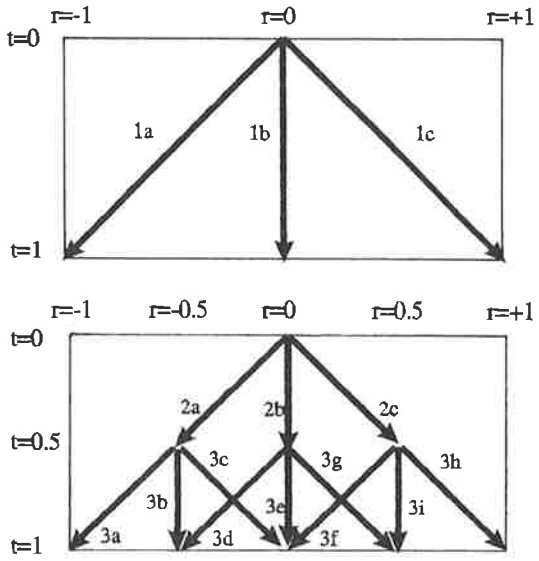



Figure 7. The basic list of vectors



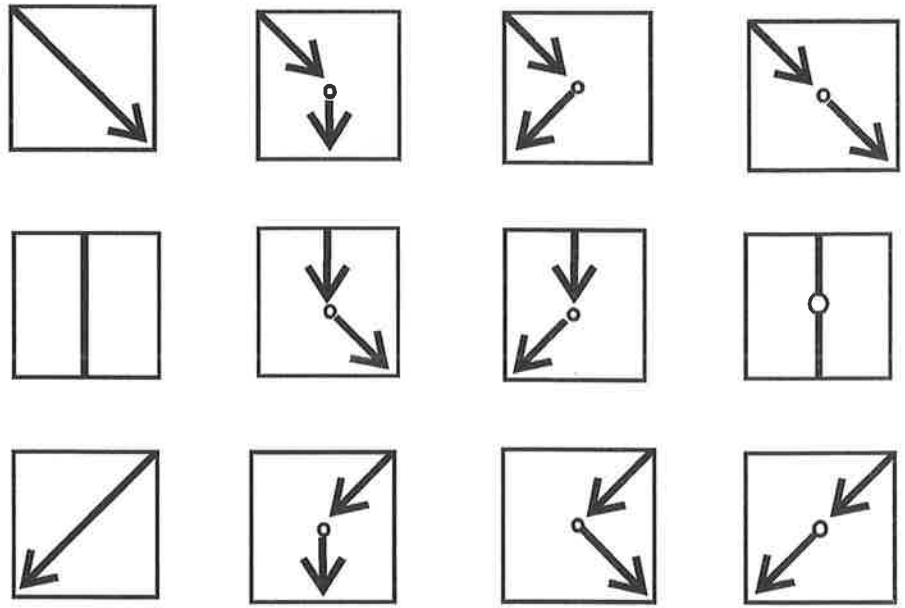


Figure 8a. The list of mono-valent imaginistic units

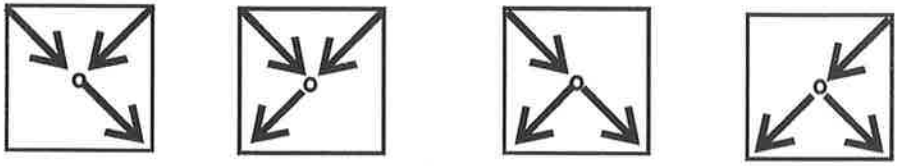


Figure 8b. The list of bi-valent imaginistic units

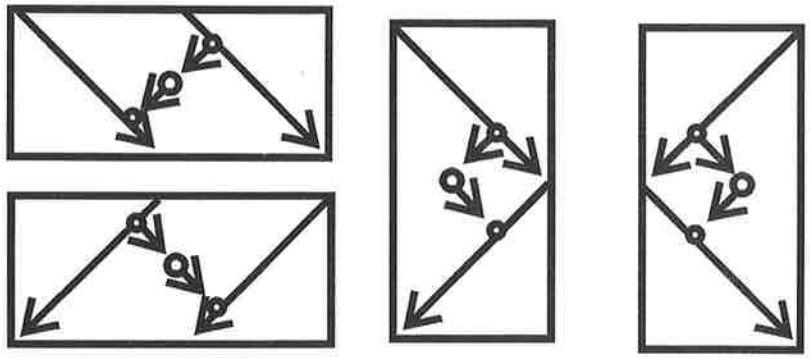


Figure 8c. The list of tri-valent imaginistic units

If we use these units as a fund of pictorial shorthand for the description of what happened in the short narrative analysed in the previous section, we get the imaginistic matrix exposed in Fig. 9. The values of the second order attribute (kind of substrata) are only given for the major force (with a vector r = |1|). The intermediary forces are only represented in the case of tri-valent units.

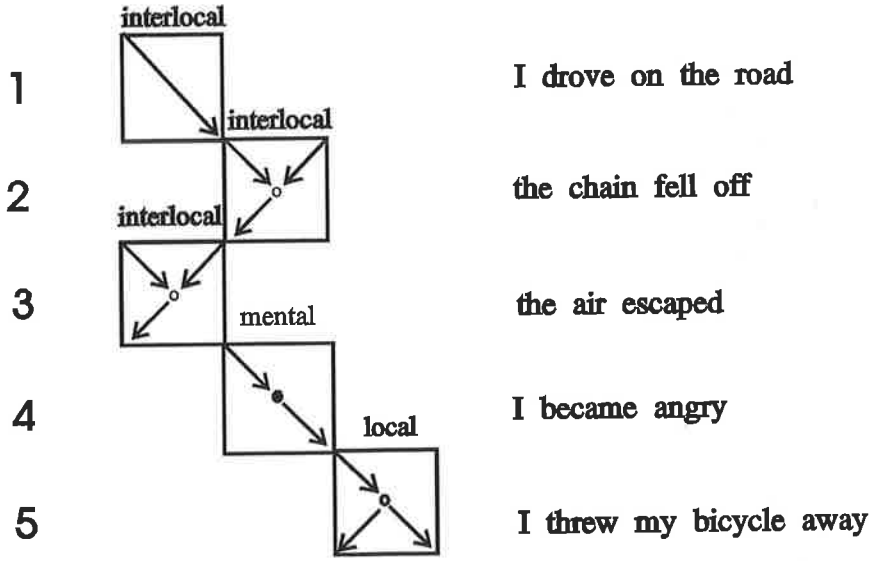



Figure 9. A simplified transcription of the narrative in the imaginistic code

We can see a line of protagonistic forces which in (2) and (3) is interrupted. The representation does not show that the antagonistic forces come out of the instrument implicit in (1): the bicycle. The complicating scheme is repeated in (2) and (3) and is therefore the centre of gravity of the story. The entity (4), which contains a mental process, is an evaluation (a further marking of the climax). The unit (5) where the protagonist dominates the antagonist (contrary to units (2) and (3)) is the result (cf. Labov 1972).

This short analysis points to the restrictions of an attribute-value model used in the previous chapters and shows the gain of naturalness which may be arrived at, if we change systematically the framework of our description from:

| | | |
|---|---|---|
| static | to | vectorial (dynamics) |
| linear | to | multi-dimensional |
| linguistic order | to | spatial order (in a cognitive framework). |



This sketch of one short textual analysis shows that we cannot only write a grammar (unification and attribute based) with the imaginistic information extracted from the text, we can also describe semantic (and syntactic) coherence pattern in a narrative using an abstract imaginistic representation in the same way in which we coded imaginistic information in an attribute-value notation, we can recode attribute-value information (partially) into an imaginistic representation of the type developed in Wildgen (1990 and forthcoming). In this sense imaginistic and attribute value dexriptions are parallel but they have different global properties such that at specific levels of description one of them or a specific mixture (hybrid) is more appropriate.

## References

**Chomsky, Noam** (1957). *Syntactic Structures.* The Hague, Mouton.

**Dretske, Fred I.** (1986). *Knowledge and the Flow of Information.* Cambridge (Mass.), Bradford M.I.T. Press.

**Engel, Ulrich** (1988). *Deutsche Grammatik.* Hweidelberg, Groos.

**Fillmore, Charles & Paul Kay** (1987). *Construction Grammar Lecture.* Stanford, LSA Summer Institute.

**Gazdar, Gerald, Ewan Klein, Geoffrey Pullum & Ivan Sag** (1985). *Generalized Phrase Structure Grammar.* Cambridge (Mass.), Harvard U.P.

**Haken, Hermann** (1988). *Information and Self-Organization. A Macroscopic Approach to Complex Systems.* Berlin, Springer.

**Labov, William** (1972). The Transformation of Experience in Narrative Syntax, chapter 9 of Labov, W., *Language in the Inner City. Studies in the Black English Vernacular.* Philadelphia, University of Pennsylvania Press: 354-396.

**Liedtke, Joachim** (1990). *Narrationsdynamik. Analyse und Schematisierung der dynamischen Momente im Erzählprodukt.* Tübingen, Narr.

**Pollard, Carl & Ivan A. Sag** (1987). *An Information-Based Syntax and Semantics, Vol. 1: Fundamentals.* CSLI Lecture Notes 13, Stanford.

**Shieber, S.** (1986). *An Introduction to Unification-Based Approaches to Grammar.* CSLI Lecture Notes 4, Stanford, CSLI.

**Wildgen, Wolfgang** (1982). *Catastrophe Theoretic Semantics. An Elaboration and Application of René Thom's Theory.* Amsterdam, Benjamins.

**Wildgen, Wolfgang** (1990). Sketch of an Imaginistic Grammar for Oral Narratives. In: Wagner, K.H & Wildgen, W. (eds.), *Studien zur Grammatiktheorie. Bremen, Universitätsverlag*

**Wildgen, Wolfgang** (forthcoming 1992/1993). *A Systematic Image and Process Centred Model of the Meaning of Sentences and Narrative Texts* (monograph submitted to Benjamins, Amsterdam). Ms., Bremen.

**Wildgen, Wolfgang & Laurent Mottron** (1987). *Dynamische Sprachtheorie: Sprachbeschreibung und Spracherklärung nach den Prinzipien der Selbstorganisation und der Morphogenese.* Bochum, Brockmeyer.

# The Influence of Context on Narrative Structure

*Erik Møller, Copenhagen*

## Introduction

In the last 20 years, much European linguistics has shown that the way language is used is dependent on the specific situation in which it is produced. Speech is never produced in empty space, but in a concrete situation with a specific speaker and listener with specific aims etc. The analyses showing this have normally been qualitative, not quantitative; and linguists have tended not to discuss the fact that speech is not a homogeneous entity, but established by an unknown number of different genres. In this article, I want to investigate - in a quantitative way - how context influences the structure of one specific oral genre, the narrative. By context I mean the speech event, the number of participants, and the fight for the floor; all this has a decisive influence on the way the narrator builds up his narrative.

## 1. The narrative and its communicative function

My definition of the narrative takes as a point of reference the definitions made by Labov (1972), Labov & Waletzky (1967) and Quasthoff (1980). A narrative is a social reconstruction of a past event, in which the narrator himself has taken part. The narrative has a point (cf. Polanyi 1989:46ff) and is structured chronologically in such a way that at least two of the sentences make up a temporal juncture (Labov & Waletzky 1967:25).

In an interaction, each narrative has a communicative function. On the basis of speech act theory, I have defined three communicative functions: information, self-aggrandisement, and entertainment (inspired by the work of Quasthoff 1980, Vogel 1986). All three functions may be present, but normally one is more dominant than the others.

*The narrative containing information* is a narrative which is told with the purpose of informing the listener, i.e. it gives arguments in favour or against something, persuades, or exemplifies something to the audience. It is the most factual of the three types, and it is the narrative which is most constrained by a contract of realism. *The narrative containing self-aggrandisement* focusses on the narrator himself. The narrative is used to put the narrator in a favourable



light by attributing a series of positive characteristics to himself. The narrative is about the narrator even if it looks as if it is not. In certain cultures, e.g. the Danish, it is the least accepted type of the three since it is close to the stigmatised genres 'boasting' and 'lie'. *The narrative containing entertainment* is used by the narrator to create a socially pleasant atmosphere by focussing on peculiar accidents or representing something in a grotesque way - all with a humorous tone. This type of narrative is most remote from the claim of objectivity and has, though to a lesser degree than the narrative containing information, a built-in claim to realism in its structure (see Møller 1993 for a discussion of communicative function).

## 2. Data: The Copenhagen Study in Urban Sociolinguistics

### 2.1. *The theoretical basis*

My data is drawn from the corpus my colleagues and I collected for The Copenhagen Study in Urban Sociolinguistics (Gregersen & Pedersen 1991), a quantitative sociolinguistic study of the Copenhagen speech community, inspired by the work of William Labov and Lesley and James Milroy (Labov e.g. 1972, 1984, L. Milroy 1980(87)). However, the Copenhagen Study deviates from these studies with respect to the importance assigned to various aspects. More emphasis was placed on the syntactic and pragmatic levels than in the above-mentioned works which have described sociolinguistic variation mainly in terms of phonetic variation. Furthermore, the above studies have been concerned mostly with *inter*individual differences with respect to traditional speaker variables like class, age, and gender, whereas the Copenhagen study has focussed equally much on *intra*individual differences, as manifested in different speech styles. Our informants were recorded in two situations (speech events): an interview and a group session. These two contexts were analysed in terms of their stylistic variation, i.e. in terms of level of formality. The speech material of the two speech events was divided into formal and informal sections on the basis of the conversational structure, topic, and phonetic performance (Albris 1991:58ff). Labov 1984:32 has claimed that narratives are always part of informal speech. In our material, this assumption turned out to be largely correct, since more than 70% of the narratives were told in sections with informal speech. The question is whether the stylistic analysis makes an independent analysis of the influence of the speech event superfluous. Quite a lot of analyses of phonetic variables show that this is indeed the case. If one compares any phonetic variable in two different speech events, holding style constant, one finds that it is dependent on style only and not on the speech event, though certain speech events naturally further certain styles. However, Romaine 1984 proposes that this only holds for specific levels in language. For phonetics, phonology, and syntax, Romaine

claims, language as an abstract agent exercises power over expression. For semantics and pragmatics, however, 'speakers exercise more power over, or are more active agents in those aspects or parts of the language system where intrinsically meaningful choices exist' (Romaine 1984:34). If this is true, the distinction formal - informal speech is too abstract to be used for all levels of analysis; the dependence of the speech event seems to be stronger for the area of semantics and pragmatics than for e.g. phonetics because those levels are more in the hands of the speakers. To conclude, the question is whether the speech event has any influence on the narrative structure even when we hold style level constant.

### 2.2. *The recordings*

As stated above, the informants were recorded in two different speech events, the conversational interview and the group session (a family gathering). The conversational interview was a relatively less controlled interview than the sociolinguistic interview Labov conducted (Labov 1984:32). We tried to incorporate much of the criticism researchers like Nessa Wolfson (1976, 1982), Ronald Macaulay (1984, 1991:16ff), and Lesley Milroy (1987:41ff) have stated. We did not elicit narratives as Labov did by asking the famous danger-of-death-question; when people wanted to tell narratives we let them tell them, but if they didn't we didn't elicit any. By conducting the fieldwork in this way we were able to find out which of our informants *chose* to tell narratives in the two speech events, and which did not - and which genres they chose to use instead - obviously of interest for the pragmatic variation approach (for a detailed discussion of the fieldwork, see Albris et al. 1988:53ff, Møller 1993). The interview situation was created by the fieldworkers as a kind of experimental situation; the other speech event, a family gathering, was a combination of a socially natural situation and an arrangement made by the fieldworker. Two or three family members met for several hours, drinking coffee and chatting. The fieldworker took part as a natural participant, not trying to be invisible (Albris 1991:56).

### 2.3. *The informants*

The data was collected in one neighbourhood, Nyboder, in central Copenhagen (for a description of the area, see Albris et al. 1988:8ff, Gregersen et al. 1991: 10ff). All the informants were born and bred in the area, but it was not a requirement that they lived in the area at the time of recording in order for them to participate as informants in the study. Many of the informants were acquainted with each other, maybe not at the time of the recording but in their childhood. The informants were chosen according to their age, sex, and class.



For this specific study I have chosen 9 interviews and 5 group sessions with the same 9 informants, aged between 25-39 years. The informants were all recorded in an interview and in a group session with their brother or sister, sometimes with their spouse and children (but they did not participate as informants in the study because they had not grown up in the area):

Table 1
Number of narratives per 100 minutes in interview and group session[1]

| Informants | Interview | | | Group Session | | |
|---|---|---|---|---|---|---|
| | narr. | min. | % | narr. | min. | % |
| Jeanne A | 1 | 150 | 0.67 | 0 | 120 | 0.00 |
| Brith J | 6 | 120 | 5.00 | 4 | 120 | 3.33 |
| Alice M | 33 | 110 | 30.00 | 27 | 170 | 15.88 |
| Thorkild R | 5 | 110 | 4.55 | 20 | 170 | 11.76 |
| Dan O | 10 | 92 | 10.87 | 8 | 120 | 6.67 |
| Jane L | 4 | 95 | 4.21 | 2 | 90 | 2.22 |
| Eva P | 0 | 102 | 0.00 | 5 | 90 | 5.56 |
| Frank C | 0 | 77 | 0.00 | 2 | 150 | 1.33 |
| Jes T | 6 | 102 | 5.88 | 14 | 150 | 9.33 |
| Total | 65 | | | 82 | | |

*Legend:* % = number of narratives per 100 minutes

In table 1 we see that the number of narratives told is similar in both speech events, 65 narratives in the interview versus 82 narratives in the group session. But as we have more informants telling narratives in each group session, the narratives turn up with greater frequency here than in the interview. Normally a group session is looked upon as more informal than an interview, and my results correspond with this assumption. Narratives are often told in informal private situations, and a family gathering is obviously private and more informal than an interview between two strangers.

Out of the total number of narratives, 36 of the narratives in the interview and 50 in the group session were told to inform. The number of narratives told in order to entertain was 13 in the interview and 27 in the group session. With

[1] In the group session, Dan O participated together with his sister Ebba R. At the time of the recording she was 20 years old, and because of her age placed in the age group of young people between 14-24 years old.

respect to the narratives told to give an impression of self-aggrandisement, the numbers were 7 and 3 respectively. Finally, there were 9 narratives in the interview and 2 in the group session with uncertain communicative function. For the following quantitative analysis I concentrate on the narratives with the communicative functions to inform and to entertain.

## 3. Methods

In order to describe the surface structure of the narrative, I follow the definition in Labov (1972) and Labov & Waletzky (1967), although slightly revised (Møller 1991). According to Labov a fully-formed narrative will manifest the following five elements:

1) abstract
2) orientation
3) complicating action
4) evaluation
5) coda

A narrative begins with a short summary of the narrative, the abstract, to introduce the genre in the conversational structure, and its specific theme. Then a series of orientation sentences follow to make it possible for the audience to understand the plot, e.g. time, place, and interactants are introduced. Then comes the actual retelling of the plot, the complicating action. This consists of sentences told in chronological order 'and so ... and so ...'. After the complicating action we are normally given an evaluation, although this may also be placed at other junctures in the narrative structure. The function of the evaluative sentences is to emphasise why the narrative is told, to highlight and justify the tellability of the plot. Finally, there are some comments, called coda, that bring the audience back from the complicating action in the narrative universe to the time of the telling, the present.

The five structural elements are defined at the level of the sentence. In this study, I focus on the distribution of orientation, complicating action, and evaluation, because of their frequency. I have chosen to use a chi-square test to decide whether differences were significantly different or not (level for significance is standard, i.e. 0.05). I want to find out whether the two types of narrative (the ones to inform and the ones to entertain) have different structures according to type and to speech event. The differences are evident from significantly different frequencies of orientation, complicating action, and evaluation, i.e. I report quantitative differences resulting in quantitatively different substructures for qualitatively different substyles.



## 4. Results

In section 4.1, I shall analyse the length of the narrative in the two different speech events. In section 4.2, I concentrate on the narrative structure. In both sections my data is the narratives containing the communicative functions to inform and to entertain.

### 4.1. *The length of the narrative*

In the interview we did not have any fight for the floor. The fieldworker - in accordance with his instructions - did not interrupt the informant or try to produce narratives himself. Thus, the informant had all the time he needed to tell the narrative. In the group session, however, we did have a fight for the floor. The narrator had to work hard at keeping the floor. If the audience does not want to listen to the narrative and interrupts the narrator, the 'face' of the narrator will be insulted (cf. Polanyi 1989:45). The narrator has to tell his narrative more quickly, to avoid having to start all over, or he will be in danger of losing his right to narrate.

In tables 2 and 3, I compare the number of sentences in narratives with the communicative functions to inform and entertain in the interview situation and the group session.

Table 2
Sentences in narratives with the communicative function to inform per 100 narratives[2]

| Interview | Group Session | Chi-square | P |
|---|---|---|---|
| 750 / 36 / 2083.3 | 681 / 50 / 1362.0 | 65.449 | $< 10^{-8}$ |

Both tables show that both types of narrative are significantly longer in the interview situation than in the group session.

The results indicate that in a situation with fight for the floor, the narrator will tell shorter narratives than in a situation without any fight. The length of a narrative can, thus, be seen as a function of the interactional structure[3].

[2] Tables 2 and 3 should be read as follows: the numbers under the headings Interview and Group Session are the number of sentences, the number of narratives and the number of sentences per 100 narratives. Then we have the result of the chi-square test and the level of significance. Tables 3 and 4 are structured in the same way.

[3] The results from tables 2 and 3 contradict Nessa Wolfson's claim that narratives - and she is referring to the narratives collected by the fieldwork methods used by

Table 3
Sentences in narratives with the communicative function to entertain per 100 narratives

| Interview | Group Session | Chi-square | P |
|---|---|---|---|
| 375 / 13 / 2884.6 | 373 / 27 / 1381.5 | 106.02 | $< 10^{-8}$ |

In tables 4 and 5, I want to establish whether there is a difference in length between narratives with different communicative functions holding the speech event constant.

Table 4
Number of sentences per 100 narratives in the interview

| to inform | to entertain | Chi-square | P |
|---|---|---|---|
| 750 / 36 / 2083.3 | 375 / 13 / 2884.6 | 26.69 | $25(10^{-8})$ |

Table 5
Number of sentences per 100 narratives in the group session

| to inform | to entertain | Chi-square | P |
|---|---|---|---|
| 681 / 50 / 1362.0 | 373 / 27 / 1381.5 | 0.049 | 0.82 |

In the interview, narratives with the communicative function to entertain are significantly longer than narratives which inform (table 4). In the group session, however, we do not find any significant difference (table 5).

The results from tables 2 to 5 show that narratives containing both communicative functions have a significantly higher number of sentences in the interview than the same narrative types in the group session. In the interview a significantly higher number of sentences was found in the narratives which entertain

Labov - told in an interview are shorter than narratives told in other speech events (Wolfson 1976:192). The results, however, are in accordance with Quasthoff's claim that narratives in the interview are told under hothouse conditions (Quasthoff 1980:26). The reason we find such different conclusions may be that Labov elicited his narratives, whereas the narratives Quasthoff and I have as our data are naturally produced, i.e. they are told because the informant himself wanted to tell something in this specific genre.



than in the narratives which inform. Such a difference was not found in the group session. We can interpret this as an indication of the fact that the communicative function determines the length of the narrative in the interview to a higher degree than the interactional structure, whereas in the group session with the resulting fight for the floor, the interactional structure determines the length to a higher degree than the communicative function.

### 4.2.1. *Analysis of the narrative structure in the interview*

Below, an analysis is given of the total narrative structure for the narratives which inform and entertain in the interview:

Table 6
Elements of structure in the interview[4]

| Struct.element | Narratives to | | Total |
|---|---|---|---|
| | inform | entertain | |
| Abstract + coda | 73 (64.7) | 25 (32.3) | 97 |
| Evaluation | 135 (134.0) | 66 (67.0) | 201 |
| Com. action | 210 (244.0) | 156 (122.0) | 366 |
| Orientation | 333 (307.3) | 128 (153.7) | 461 |
| Total | 750 | 375 | 1125 |

D.F. = 3, Chi-square = 23.16, P = 0.000037

In table 6, we find a significantly different structure between narratives which inform and those told to entertain in the interview. A chi-square test for each of the elements orientation, complicating action, and evaluation for the two narrative types reveals that for the elements orientation (chi-square 6.431, sign. = 0.025) and complicating action (chi-square 14.213, sign. = 0.001), there is a significantly different frequency, whereas there is no significant difference between the frequency of evaluation in the two types. Consequently, we find two different substructures: one for narratives to inform, with a relatively high num-

[4] The table should be read as follows: because of the low frequency of the structural elements abstract and coda, I have combined them into one group. The numbers in the brackets are the expected numbers based on marginal frequencies. The abbreviation D.F. means degree of freedom.

ber of orientation sentences and relatively few complicating action sentences, and one for narratives to entertain, containing a structure with relatively few orientation sentences and a relatively high number of complicating action sentences.

### 4.2.2. *Analysis of the narrative structure in the group session*

The question is whether the analysis below will show differences for the narratives in the group session as we found in table 6 for the narratives in the interview.

Table 7
Elements of structure in the group session

| Struct.element | Narratives to | | Total |
|---|---|---|---|
| | inform | entertain | |
| Abstract + coda | 35 (32.3) | 15 (17.7) | 50 |
| Evaluation | 81 (89.9) | 58 (49.1) | 139 |
| Com. action | 233 (227.0) | 118 (124.0) | 351 |
| Orientation | 332 (331.8) | 181 (181.2) | 513 |
| Total | 681 | 372 | 1053 |

D.F. = 3, Chi-square = 3.58, P = 0.31

In the group session, we cannot detect the sort of difference in structure found for the interview narratives. The absence of significance in table 7 indicates that, in the group session, narratives have one and the same narrative structure irrespective of narrative type.

### 4.2.3. *Analysis of the narrative containing the communicative function to inform in the two speech events*

Below, an analysis is given of the overall structure of narratives which inform in the two speech events.

Table 8 indicates that narratives containing the communicative function to inform have a significantly different structure according to the speech event in which they are produced. More precisely, the difference is between complicating



action sentences (chi-square 4.452, sign. = 0.05) and evaluation sentences (chi-square 8.815, sign. = 0.01), whereas we do not find any difference as far as orientation is concerned. In summary, we have two different structures for the narrative to inform: in the interview one with relatively few complicating action sentences and relatively many evaluation sentences; in the group session one with relatively many complicating action sentences and relatively few evaluation sentences.

Table 8
Elements of structure in the narratives to inform

| Struct.element | Narratives to inform in | | Total |
|---|---|---|---|
| | the interview | the group session | |
| Abstract + coda | 72 (56.1) | 35 (50.9) | 107 |
| Evaluation | 135 (113.2) | 81 (102.8) | 216 |
| Com. action | 210 (232.2) | 233 (210.8) | 443 |
| Orientation | 333 (348.5) | 332 (316.5) | 665 |
| Total | 750 | 681 | 1421 |

D.F. = 3, Chi-square = 24.20, P = 0.00002

This difference between the frequencies for complicating action and evaluation in the two speech events can be related to Labov's normative claim that evaluation sentences is a less sophisticated way of evaluating the narrative compared to embedded evaluation, i.e. evaluation which forms part of the other structural elements (e.g. the embedded evaluation may be part of an orientation sentence, Labov e.g. 1972:370ff). For the narratives analysed here, this would mean that when the narrator is in a speech event in which he has to fight for the floor, he has to use a more sophisticated narrative structure, and this is done by focussing more on the complicating action and less on the evaluation. In such a situation evaluation is transformed into embedded evaluation and is told as part of the complicating action.

### 4.2.4. *Analysis of the narrative to entertain in the two speech events*

I now want to investigate whether the overall structure of the narratives which entertain is different in the two speech events.

Table 9 once again shows a significantly different structure of the narrative according to the speech event in which the narrative is produced. Specifically, we find a structural difference in the orientation sentences (chi-square 9.521,

sign. = 0.001) and complicating sentences (chi-square 4.969, sign. = 0.05), on the other hand there is no difference in the evaluation sentences. Thus, we have two different narrative structures for the narratives which entertain: in the interview we have one structure with relatively identical frequency of orientation and complicating action sentences, in the group session we have a structure with a relatively high number of orientation sentences and relatively few complicating action sentences. The result is in itself not surprising, as the results in table 7 showed that we did not have two different substructures for narratives containing information and entertainment respectively in the group sessions.

Table 9
Elements of structure in the narratives to entertain

| Struct.element | Narratives to | | Total |
|---|---|---|---|
| | inform | entertain | |
| Abstract + coda | 25 (20.1) | 15 (19.9) | 40 |
| Evaluation | 66 (62.2) | 58 (61.8) | 124 |
| Com. action | 156 (137.6) | 118 (136.4) | 274 |
| Orientation | 128 (155.1) | 181 (153.9) | 309 |
| Total | 375 | 372 | 747 |

D.F. = 3, Chi-square = 17.32, P = 0.0006

#### 4.2.5. *Summary of the analysis of narrative structure*

My analyses have shown that we find different structures for the narrative according to the speech event in which it is produced, even though the style analysis showed that we had the same informal speech style. The difference in narrative structure in the narratives containing information and entertainment in the interview cannot be found in the narratives with similar functions in the group session; here the two types of narrative have similar structures. Both types of narrative have different structures according to the speech event. The result may be interpreted as an indication of the fact that the narrative structure has greater importance when the communicative function is established in the interview than in the group session. Let us illustrate the result of the analysis in the following figure to sum up where we found structural differences:



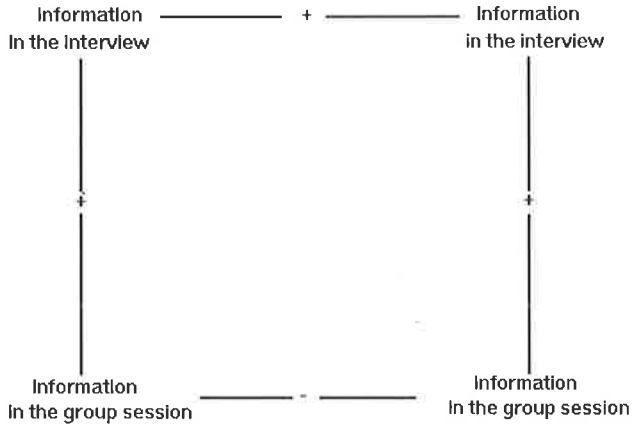



Figure 1. Structural differences between the narratives to inform and entertain in the interview and the group session
*Legend:* + = significant difference, - = non-significant difference

## 5. Discussion

In the American tradition of narrative analysis, we find that in the quantitative paradigm the narrative is analysed without regard to the context and the communicative function of the narrative under analysis. This paradigm is dominant both in sociolinguistics and in psycholinguistics (e.g. Labov 1972, Mandler 1984, Wolfson 1982). On the other hand, in the qualitative paradigm both context and communicative function are discussed, but we do not see any reflection on the fact that narratives fall into different subtypes containing specific substructures (e.g. Polanyi 1985). In my study (Møller 1991, 1993), I have tried to integrate these two paradigms.

Both subtypes of narrative, the one primarily told to inform and the one primarily told to entertain, had significantly fewer sentences in the group session than in the interview. We thus find a connection between competing for the floor and the length of the narrative. A speech event which inherently includes competition for the floor produces shorter narratives than a speech event without any competition. However, we have to remember that the speech event with a fight for the floor was a family gathering, and in Danish families we normally find democratic relations among all adult family members; no member has more formal power than the others. Thus, we still do not know anything about the length of the narrative in speech events with a fight for the floor and asymmetrical power relations.

In the interview, we found a different narrative structure between narratives which inform compared with those which entertain. A priori, one would expect the differences found in the interview to be accentuated in the group session, since with the fight for the floor the narrator would have to profile his narratives according to the communicative function he wanted to get across. But this did not happen. Both types of narrative had identical narrative structures (i.e. not significantly different from each other). And furthermore, in the group session, both types of narrative had a significantly different structure compared with the interview situation. The structural variation between narratives containing the communicative functions to inform and entertain disappears so to speak in the group session. Since there are no problems with assigning communicative functions to narratives in the group sessions, it follows that the narratives in the group session have to profile their communicative functions at other levels than the structural one at sentence level. A possible place to look for an explanation is in the way the discourse is performed.

In the group session there is a tendency for narratives to be produced in sequences; this means that the narratives - as regards the way they are placed in the conversation and as regards their propositional structure - have an air of belonging to a greater unit than the narrative itself. This seemingly parallel ordering of narratives can be seen as an indication of the fact that participants in a discourse try to build up coherence at discourse level. Coherence is among other things established through turn taking, or rather a kind of meta turn taking, i.e. a tacit negotiation about both the conversational structure, the genre, its possible communicative function, and theme: 'In the following part of the discourse we shall decide to tell narratives, and listen to each other's narratives, about theme x to have a good laugh'.

When someone begins a new sequence, some longer careful abstract is often found in the first narrative. The narrator tries to make clear that he starts a new sequence, maybe with a new genre introduced, and he is careful to mark the communicative function, as in this abstract: 'I am just going to tell you something funny about my trip to Germany without being able to speak a single damned German word'. The narrator explicitly marks the genre (a narrative: the word tell), the communicative function to entertain (something funny), and the theme 'problems with foreign languages'. When the first narrative is over, another speaker takes over by saying: 'Once I was in France', and everybody knows that a humorous narrative about not knowing the French language will follow. That is, when a narrative is told with a specific communicative function, the interactants expect the next speaker to go on telling a narrative with a similar communicative function, and so on until a speaker explicitly states that he is going to change the genre or communicative function.

The analyses of the narratives to inform and entertain in the two speech events indicate that the communicative function of a narrative can be communicated at several levels, often in a complicated interplay. In this interplay, the



narrative structure - to a greater or lesser extent - can communicate a specific communicative function. The specific structure is thus a resource which may be triggered by certain speech events. My analyses suggest that the speech event must be without fight for the floor, and situations must be such that the single narrative is not dependent on other narratives in longer sequences.

## 7. Conclusion

My analysis has shown that it does not suffice to look at speech divided into formal and informal speech when one wants to conduct analyses at the level of genre. Done in this way, the narrative would always be part of the analysis of informal speech. Furthermore, Labov and others have suggested that we have a narrative structure which is the narrative structure. This is not the case. In the analyses one has to ask the pragmatic question: 'Which function is fulfilled in this situation in this speech event?', and to find out how the relationship between narrative structure, communicative function and speech event is construed. The narratives - or in Quasthoff's words hothouse narratives - told in the interview have, as my analyses have shown, a narrative structure different from narratives told in the less artificial climate of the group session. The question is thus how representative hothouse narratives are compared to narratives told in speech events in social reality. It is my belief that my analyses have shown differences but have not told everything about these differences. The reason that it is not possible to come closer to the solution of this problem is that my data material - collected in the traditional sociolinguistic way - has certain limitations. To get further, we have to record narratives from various contexts to find out how different contexts influence narrative structure. From a methodological point of view this means that we have to venture into more anthropologically influenced fieldwork, methods such as those used by Heath 1983. We have to follow our informants in the real world and record narratives told in the kitchen, in the church, in the pub, told in order to inform, to entertain - in short to create life and the meaning of life.

## References:

**Albris, J.** (1991). Style Analysis. In: Gregersen, F. & Pedersen, I.L. (eds), *The Copenhagen Study in Urban Sociolinguistics*. Copenhagen, Reitzel: 46-107.

**Albris, J., Gregersen, F., Holmberg, H., Møller, E., Pedersen, I.L., Thomsen, O.N** (1988). *The Copenhagen Study in Urban Sociolinguistics. Interim Report*, Institute of Nordic Philology, Copenhagen University.

**Gregersen, F., Pedersen, I.L.** (eds) (1991). *The Copenhagen Study in Urban Sociolinguistics*. Copenhagen, Reitzel.

**Gregersen, F., Albris, J., Pedersen, I.L.** (1991). *Data and Design of the Copenhagen Study.* In: Gregersen, F., Pedersen, I.L. (1991): 5-44.

**Heath, Sh. B.** (1983). *Ways with Words. Language, Life, and Work in Communities and Classrooms.* Cambridge, Cambridge University Press.

**Labov, W.** (1972). *Language in the Inner City - Studies in Black English Vernacular.* Philadelphia, Pennsylvania University Press.

**Labov, W.** (1984). Field Methods of the Project on Linguistic Change and Variation. In: Baugh, J., Sherzer, J. (eds.), *Language in Use.* Englewood Cliffs, New Jersey, Prentice Hall: 28-53.

**Labov, W., Waletzky, J.** (1967). Narrative Analysis: Oral Versions of Personal experiences. In: Helm, J. (ed.), *Essays on the Verbal and Visual Arts.* Seattle, University of Washington Press: 12-45.

**Macaulay, R.K.S.** (1984). Chattering, Nattering, and Blethering: Informal Interviews as speech events. In: Enninger, W., Haynes, L. (eds), *Studies in Language Ecology.* Wiesbaden, Steiner: 51-64.

**Macaulay, R.K.S.** (1991). *Locating Dialect in Discourse. The Language of Honest Men and Bonnie Lassies in Ayr.* New York, Oxford University Press.

**Mandler, J. M.** (1984). *Stories, Scripts, and Scenes: Aspects of Schema Theory.* Hillsdale, New Jersey, Erlbaum.

**Milroy, L.** (1980). *Language and Social Networks.* Oxford, Blackwell, 2nd ed.

**Møller, E.** (1991). Narratives in the Sociolinguistic Interview. In: Gregersen, F., Pedersen, I.L. (eds) (1991): 241-336.

**Møller, E.** (1993) *Mundtlig fortælling - fortællingens struktur og funktion i uformel tale.* (Oral narrative - the structure and function of narrative in informal speech). Copenhagen, Reitzel.

**Polanyi, L.** (1985). Conversational Storytelling. In: van Dijk, T. (ed.), *Handbook of Discourse Analysis 3.* London, Academic Press: 183-201.

**Polanyi, L.** (1989). *Telling the American Story. A Structural and Cultural Analysis of Conversational Storytelling.* Cambridge, Mass., MIT Press.

**Quasthoff, U. M.** (1980). *Erzählen in Gesprächen. Linguistische Untersuchungen zu Strukturen und Funktionen am Beispiel einer Kommunikationsform des Alltags.* Tübingen, Narr.

**Romaine, S.** (1984). The Status of Sociological Models and Categories in Explaining Linguistic Variation. *Linguistische Berichte 90, 25-38.*

**Vogel, Th.** (1986). Verbal Planning and Narrative Structure: The Belfast Narrative. In: Harris, J., Little, D., Singleton, D. (eds), *Perspectives on the English Language.* Proceedings of the First Symposion on Hiberno-English, University of Dublin: 245-255

**Wolfson, N.** (1976). Speech Events and Natural Speech: Some Implications for Sociolinguistic Methodology. *Language in Society 5, 189-209.*

**Wolfson, N.** (1982). *The Conversational Historical Present Tense in American English Narratives.* Dordrecht, Foris Publications.

# The Statistical Structure of a Text and its Readability

*Juhan Tuldava, Tartu*

0. The term "statistical structure of a text" is used to signify certain formal quantitative characteristics of the text as a whole, such as the length of the text (measured in terms of number of words) and the mean length of the sentence or the word in the text. The relationship between the readability of the text and those statistical characteristics may be explicit or implicit. In doing this, one must take into consideration the polysemantic nature of the term "text readability". One can speak about objective and subjective text readabilities.

By subjective text readability we mean how difficult or easy the text is for a certain reader or listener, how well the reader (listener) can comprehend the text, and, in need, remember and reproduce it. There are various ways of measuring the subjective readability level of a text, such as measuring the speed of reading, asking questions about the subject matter of the text, cloze test, asking for an informant's or expert's opinion. The state of the reader or listener (age, education, occupation) is of importance in assessing the subjective readability of a text.

The objective readability of a text, or its "complexity or sophistication", is expressed by its structural or content characteristics which have experimentally proved to correlate with subjective assessments of the text difficulty level. Objective readability characteristics can singly or in combinations be used as a reliable tool in subjective readability prognostication. The objective readability of a text should also be subject to common sense assessment, so for example a text may be difficult to read and comprehend, if it contains mostly long sentences and words. In the same way, it can be said that a large number of unfamiliar new words in the text will inevitably make the comprehension of the text difficult.

We should like to discuss below some possible relations between statistical structural characteristics and the objective readability of texts proceeding from text structure and contrastive data analysis. The objective readability of a text can in its turn serve as support in assessing subjective readability.

1. The length of the text and the size of its vocabulary are text parameters which can be directly measured. The length of the text is usually measured in words or word forms and it is denoted by N (there are other ways of measuring the text length, e.g. in characters, in sentences). The size of the text vocabulary

is the total of different words in the text which is usually measured by the number of word forms (V) or by the number of lexemes (L) in the text. When we speak about "lexemes" then the different word forms (inflectional forms) are all treated under one form, its basic form as a rule. For example the word forms *brother, brother's, brothers, brothers'* will all be reduced to the lexeme (or lemma) BROTHER.

The size of the vocabulary, that is the number of different words (word forms or lexemes) in a certain text, formally characterizes how "rich" the vocabulary of the text is: in case of texts of the same size a larger vocabulary will testify to a larger variety in the choice of words. This should in its turn be reflected in the readability level of the text, i.e. the larger the text vocabulary, the more difficult the text should be. Naturally the vocabulary size alone cannot be decisive in assessing the readability level of a text. The readability level of a text can be rightly assessed by further qualitative analysis and contrastive analysis including other formal indices.

The ratio between the vocabulary size and the text length is often expressed through the so-called TTR index (TTR is abbreviated English "type-token-ratio"). In other words it is V/N or L/N and it expresses the relative richness of the vocabulary of the text under discussion. The reversed ratio (N/V or N/L) expresses the average level of repetition of a word form or lexeme in the text; consequently the higher the degree of repetition, the easier the text considered should be.

We shall use these primary and secondary (derived) indices to assess and compare four texts: A, B, C and D. The first three texts are passages of comparatively equal size from Estonian prose expressing an author's speech[1].

As can be seen from the data presented in Table 1, Texts A and B should be of approximately the same readability level whereas text C should be somewhat easier than the first two texts.

2. The relative indices described above can be used only in comparing texts of more or less the same length, as the ratio between the vocabulary size (V or L) and the text length (N) is a variable: with a longer text the vocabulary size does not grow linearly but somewhat slower, i.e. with growing N the V/N (or L/N) ratio will decrease monotonously and the ratio N/V (or (N/L) will correspondingly increase. Table 2 presents the data on Text B.

How can one then compare the vocabulary richness in texts of different sizes and make conclusions about the relative readability degree of the texts? The following techniques can be suggested.

[1]) Text A is a random sample from "Kolme katku vahel I", a novel by Jaan Kross (published 1970). Text B is a random sample from "Kartulikuljused", a novel by Aimée Beekman (1967). Text C is a random sample from "Tondiöömaja", a novel by Heino Kiik (1970). (See: Kaasik et al. 1977).



**Table 1**
Comparison of four texts

| | N | V | V/N | N/V |
|---|---|---|---|---|
| Text A | 4991 | 2861 | 0.573 | 1.744 |
| Text B | 4998 | 2869 | 0.574 | 1.742 |
| Text C | 4996 | 2738 | 0.548 | 1.825 |
| Text D | 10000 | 5630 | 0.563 | 1.776 |

**Table 2**
Data on Text B

| N | V | V/N | N/V |
|---|---|---|---|
| 1000 | 731 | 0.731 | 1.368 |
| 2000 | 1333 | 0.667 | 1.499 |
| 3000 | 1865 | 0.622 | 1.608 |
| 4000 | 2404 | 0.601 | 1.664 |
| 4998 | 2869 | 0.574 | 1.742 |

In the case of texts of relatively small sizes (N < 10000) and of smaller differences in length (less than twice as large), an approximate assessment can be obtained by using H.H. Somers' (1968) formula:

$$(1) \qquad S = \frac{\ln \ln V}{\ln \ln N} ,$$

where S is a relative vocabulary richness measure and ln is the natural logarithm.

Let us use formula (1) to compare Text B (N = 4998, V = 2869) and Text D (fictitious), which is twice the length of Text B (N = 10000, V = 5630).

Text B : ln ln 2869/ln ln 4998 = 2.0746/2.1420 = 0.969;
Text D : ln ln 5630/ln ln 10000 = 2.1559/2.2203 = 0.971.

So it can be concluded that the vocabulary richness level and the readability level are approximately the same in the two texts compared (however Text D has a slightly higher value of S index).

When the texts to be compared are of markedly different length (one is more than twice as long as the other) and also in general cases another approximation formula can be used. The formula is based on a more precise statistical analysis of the vocabulary size and the text length ratio (cf. Tuldava 1980):

$$T = \frac{\ln \ln N}{\ln \ln \frac{N}{V} + A}, \tag{2}$$

where T is the relative vocabulary richness, A is a constant depending on the language the text is in and to some extent also on the genre. In Estonian prose texts A will approximately be 6, in Russian texts 5 and in English texts 4.[2)]

Let us compare Text B and Text D again, this time using formula (2):

Text B : ln ln 4998/(ln ln 1.742 + 6) = 0.3958;
Text D : ln ln 10000/(ln ln 1.776 + 6) = 0.4077.

In this case the more marked vocabulary richness of text D is more evident and consequently text D can be supposed to be more difficult that Text B.

3. To make assessments essentially more precise some extra calculations should be performed. First, the dynamics of vocabulary growth, i.e. the relation between vocabulary size (V or L) and text length (N), should be calculated at least for one of the texts compared. This can be expressed quite precisely through the following function (Tuldava 1987:100):

$$V = Ne^{-a(\ln N)^b}, \tag{3}$$

where e is the basis of the natural logarithm, and a and b are constants.[3)] Then

[2)] Formula (2) is based on the linearized form of formula (3) (see next page) where $\ln|\ln(V/N)| = \ln\ln(N/V)$ and $\ln|a| = A$. The parameter b refers to the speed rate of vocabulary growth and appears as index to T = 1/b in formula (2). The constant |A| is a characteristic which is different for various languages (partly related to the degree of analyticism of a language). The approximate values of A have been found out on the basis of several experiments on English, Russian, and Estonian texts.

[3)] Formula (3) is a modification of the power function expressing the relation between TTR index (V/N) and text length (N) in the form $V/N = \alpha N^{\beta}$ (as a variant of Menzerath's law; see Altmann 1983). By taking the logarithm both of V/N and N we get $\ln(V/N) = a(\ln N)^b$ and then through antilogarithm back to $V = Ne^{a(\ln N)^b}$ (here a < 0), i.e. to formula (3). The formula has the advantage of meeting the limit conditions: V = 1 and N = 1. Good correspondence to empirical data has been proved on experimental material from several languages (see: Tuldava 1980 and 1987) (For another method see: Maas 1972).



to be able to compare vocabulary richness of texts the text lengths should be standardized, i.e. the text length for which the values for the constants a and b are known should be extra- or interpolated through the formula (3) to the text length of the other text and the V value should be established.

Let us take an example of the application of formula (3). We know the dynamics of vocabulary growth in Text B in its five points (see Table 2). To calculate a and b values the equation must first be linearized. This can be done by double logarithm-taking:

$$\textit{Formula (3)} \Rightarrow \frac{V}{N} = e^{-a(\ln N)^b}$$

$$\Rightarrow \ln\left|\ln \frac{V}{N}\right| = A + b \ln\ln N$$

where $A = \ln|a|$. By calculating the coefficients of linear correlation and regression between $Y = \ln|\ln V/N|$ and $X = \ln\ln N$ for Text B with the help of the method of least squares we come to the correlation coefficient r = 0.9999, A = -6.36 and b = 2.7. To reduce formula (3) to its original expression we must find the value of the constant a, using the expression $e^A = e^{-6.36} = 0.0017$. Consequently the vocabulary growth in Text B is determined by the equation

$$V = Ne^{-0.0017(\ln N)^{2.7}}.$$

Comparing the relative vocabulary richness in Text B and Text D, determined by the more precise method (by extrapolating the data of Text B to suit the text length of Text D, i.e. N = 10000) we shall have for text B:

$$V = 10000e^{-0.0017(\ln 10000)^{2.7}} = 5054,$$

which is lower than in Text D (where V = 5630) for the same text length (N = 10000).

4. When the different words in a text (word forms or lexemes) are arranged in their order of frequency of occurrence, a frequency dictionary (FD) of word forms or lexemes will be produced. FD can be used in many ways to calculate quantitative characteristics, which may be related to the level of text readability. First it is worth mentioning that the relation between rank (r) and frequency (F) can be expressed approximately through a function known as Zipf's law (Zipf 1949):

$$F_r = Cr^{-1}, \tag{4}$$

where $F_r$ expresses the frequency of occurrence of a word at the $r$th rank, and C and t are constants. This function describes the whole FD, but usually three zones of frequency, beginning, medium and end zones of the FD, can be distinguished, each with a slightly different t value. To solve our problem we must consider the final frequency zone with the constant $t_3$. In particular the final part of the FD includes rare words, the proportion of which in a text indicates the variety of the vocabulary and consequently the readability level of the text. The relationship between the constant $t_3$ and the occurrence of rare words is inversely proportional, i.e. a smaller $t_3$ value indicates a larger role (proportion) of the rare words in the text. Given the texts A, B, C, the $t_3$ values will correspondingly be 0.48, 0.50 and 0.55. That is, in other words, by the number and proportion of rare words in the text and by the supposed level of difficulty, the texts can be ordered in falling degree of difficulty A - B - C.

5. The t-characteristic of the word rank distribution is closely related to the index of rarity, well known in statistical linguistics, which is the ratio of words which occur only once in the given text (the so-called "hapax legomena") to the length of the text or the size of the text vocabulary. The rarity index is expressed by $V_1/N$ or $V_1/V$ (for word forms) and $L_1/N$ or $L_1/L$ (for lexemes) with $V_1$ and $L_1$ expressing the number of word forms or lexemes that occur only once in the given text.

The results for the three texts analyzed are presented in Table 4.

Judging by the index of rarity the three texts should also be ordered in falling degree of difficulty: A - B - C.

As was the case with the quantitative indices discussed above, the concrete value of the index of rarity also depends on the length of the text. So the texts to be compared must be of equal length or a number of recalculations should be made. The approach may be the same as was described in the case of the TTR index, i.e. the indices of the relative amount of rare words can be calculated using formula (1): ln ln $V_1$/ln ln N or ln ln $V_1$/ln ln V.

Formula (2) can also be used for this purpose by using the A value 5 for Estonian texts and 3.5 for English texts and replacing V by $V_1$. Formula (3) can likewise be used if V is replaced by $V_1$ (or L by $L_1$ - if lexemes have been counted) and we know the dynamics of change of $V_1$ (or $L_1$) dependent on text length (N).

The relation between the rare words in the text and the text length manifests itself in the $V_1$ (or $L_1$) value diminishing proportionally with the growing N. This can be demonstrated clearly by comparing a single text and the sum of 20 texts (see Table 5).

As can be seen, the ratio $V_1/N$ changes more rapidly with the growing text length than $V_1/V$. Experiments have shown that as the ratio $V_1/N$ is more sensitive in the comparison of texts and authors' styles, it can be recommended to be used in comparing texts and measuring their degree of difficulty by reading.



**Table 4**

| | N | V | $V_1$ | $V_1/N$ | $V_1/V$ |
|---|---|---|---|---|---|
| Text A | 4991 | 2861 | 2336 | 0.466 | 0.817 |
| Text B | 4998 | 2869 | 2270 | 0.454 | 0.791 |
| Text C | 4996 | 2738 | 2156 | 0.432 | 0.787 |

**Table 5**

| | N | V | $V_1$ | $V_1/N$ | $V_1/V$ |
|---|---|---|---|---|---|
| Text B | 4998 | 2869 | 2270 | 0.454 | 0.791 |
| Sum total (20 texts) | 99898 | 30733 | 21760 | 0.218 | 0.708 |

6. Besides considering the number of words in the text which occur only once, it is also possible to analyze the proportion of the number of words in the text which appear there twice, three times, four times and so on. Then the so-called "lexical spectrum" is formed, which may be related to the degree of readability of the text. The lexical spectrum expresses the distribution of words in the text and is one of the best indices of the deep structure of the text. The full lexical spectrum of a text can be calculated (prognosticated) if the values of N, V and $V_1$ (or N, L and $L_1$) are known. The Waring-Herdan distribution model can be used for the purpose (Herdan 1964; Muller 1976; cf. also Ratkowsky, Hantrais 1975):

$$(5) \qquad V_{i+1} = \frac{V_i\,(a + i - 1)}{(b + 1)},$$

where i is the frequency of the word (i = 1,2,...) and the parameters a and b are estimated according to the formulas:

$$a = (Q - M - 1)^{-1};$$
$$b = aQ;$$
$$M = V/N;$$
$$Q = (1 - V_1/V)^{-1}.$$

The cumulative lexical spectrum (i.e. the frequencies of words occurring one,

two, three, ... times successively summed) makes it possible to measure the "coverage" of the texts by its vocabulary, beginning from the end of the frequency list, i.e. from the zone of rare words. For instance, $V_{1-3}/N$ will indicate the ratio of the summed frequency of words occurring once, twice and three times to the length of the whole text. The indices of this kind correlate as a rule with the index of rarity (see above) and they point to the richness of the vocabulary and probably to the degree of difficulty the text may offer for its reader.

7. The text coverage can also be analyzed starting from the first ranks in the frequency dictionary. As a rule the role of the 10, 50 and 100 most frequent words in the text is measured by $F^*_{10}/N$, $F^*_{50}/N$ and $F^*_{100}/N$ (often expressed in per cent). According to the FD of the three texts compared the following distributions can be noted (see Table 6).

The text coverage by the most frequent words is called the "concentration of the vocabulary" in the given text. If word forms are counted it also indicates the degree of "abstractness" of the text as the most frequent words tend to be of very general meaning (most of them are auxiliary words). This is most clearly seen in analytical languages, e.g. the most frequent word forms in English texts are (Kučera, Francis 1967): *the, of, and, to, a, in, that, is, was, he,* altogether covering up to 25% of the text. In our example the texts are ordered A - C - B, which does not correlate with the indices of text readability. This can be accounted for by the individuality of style in Text A: the frequency of the most frequent word form *ja* (meaning 'and') being 5.7% whereas in Text B it is 2.7% and in Text C 3.5% (the medium value in 20 various texts being 3.2.%).

**Table 6.**
The text coverage in per cent

| | The number of most frequent word forms (r = rank) | | |
|---|---|---|---|
| | r = 1 - 10 | r = 1 - 50 | r = 1 - 100 |
| Text A | 13.8 | 25.0 | 31.3 |
| Text B | 11.0 | 21.4 | 28.3 |
| Text C | 12.9 | 24.4 | 31.2 (%) |

To measure the difficulty level of a text it is reasonable to study the vocabulary and text concentration not including the coverage of all the words in the text but including only the frequencies of certain kinds of words. So the difficulty degree of a text is closely related to the index of the proportion of unknown words in the text and in the vocabulary, the number of long words (e.g. words over seven characters) among the 100 most frequent words, and their text coverage, the proportion of abstract nouns in the text, the number of verbs among the 50 or 100 most frequent notional words and many other text characteristics which may be related to the readability level of a text.



While calculating the vocabulary concentration (of various kinds of words) one must take into consideration that the values of the indices will be dependent on the length of the text, and only texts of equal length can be directly compared. Our primary experience has shown that by comparing texts of different length, the universal relation between the size of the vocabulary and the length of the text expressed in formula (3) is valid here too. The symbol V in the formula is to be replaced by $F^*$, i.e. the cumulative frequency of words under examination. While comparing texts of different lengths the constant values for at least one of the texts are to be found and then they are to be extrapolated to the length of the other text(s) the way it was demonstrated with the V and N relation. For approximate evaluation the formula (1) can also be used (replacing V by $F^*$).

8. In conclusion we would like to present a universal formula for measuring text readability (Tuldava 1975)[4]:

$$R = \bar{i} \, lg \, \bar{j} \qquad (6)$$

where R is text readability (difficulty), $\bar{i}$ stands for the mean word length expressed in syllables, $\bar{j}$ is the mean length of sentences in words, lg is the decimal logarithm (as in the original formula, but instead of lg the natural logarithm ln can be used, which will result in different numerical values, but it will not affect the relative values). The word length can be measured in characters instead of doing it in syllables; the mean sentence length may be measured in characters instead of words and so on, which will all result in slightly different values from the values achieved using the original formula (6), but in principle the differences cannot be significant.

The three texts under discussion yield the following results when we use formula (6). The data are presented in Table 7.

The texts can be ordered in diminishing order of difficulty: A - B - C, which was also demonstrated by the statistical characteristics of text structure mentioned above (TTR, S, T, V(N), $t_3$, $V_1/N$ and $V_1/V$). It may be added that the results correspond very well with the intuitive evaluation of the named texts' readability degree made by a number of native informants-experts. The results

[4]) The translation of the article with the theoretical substantiation of the formula will be published in Glottometrika 14.

correspond also with the well-known Flesch's Reading Ease formula (Flesch 1948) and with some other formulas (for a review of existing formulas see Klare 1974-1975).

**Table 7**

| | $\bar{i}$ | $\bar{j}$ | $\lg \bar{j}$ | $R = \bar{i} \lg \bar{j}$ |
|---|---|---|---|---|
| Text A | 2.34 | 19.1 | 1.28 | 2.99 |
| Text B | 2.36 | 10.4 | 1.02 | 2.41 |
| Text C | 2.22 | 9.1 | 0.96 | 2.13 |

There exists a great number of other statistical characteristics of the text structure: J. Mistrík's index of stereotype level (Mistrík 1967), M. Těšitelová's indices of vocabulary richness (Těšitelová 1972), B. Golovin's statistical style indices (Golovin 1971) and others. The relationship of those text characteristics with text readability needs further research.

9. Finally, it should be noted that as no author has ever performed statistical testing concerning the above mentioned indices (TTR, $V_1/N$, R etc.) one always resigned to the ranking of texts according to these measures. However, as G. Altmann in a letter to me has pointed out, in all cases at least an asymptotic test is possible. As an example we shall present his considerations about the testing of the index R (formula 6).

In order to test the difference of R in two texts, we set up the criterion

$$u = \frac{R_1 - R_2 - [E(R_1) - E(R_2)]}{\sqrt{V(R_1) + V(R_2)}}$$

that can be used in the case of large samples. Here u is the normal variate, $R_x$ (x = 1,2) is defined in formula (6), $E(R_x)$ is the expected value of $R_x$ and $V(R_x)$ is its variance. The texts are independent, thus $Cov(R_1,R_2) = 0$. Under the null hypothesis, $E(R_1) = E(R_2)$; thus we need not compute them - unless a and b in the two texts are different.

V(R) can be derived as follows: The mean word length ($\bar{i}$) depends on mean sentence length (j). According to Tuldava (1975)

(a) $\quad i = a + b \ln \bar{j}$

or according to the law of Menzerath (Altmann 1983):





(b) $\quad i = a\bar{j}^b$.

Thus R can be written for our purposes as

(a) $\quad R = (a + b \ln \bar{j}) \ln \bar{j}$

or

(b) $\quad R = (a\bar{j}^b) \ln \bar{j}$.

In order to derive V(R) we use the expansion in Taylor series and obtain

$$V(R) = (dr/d\bar{j})^2 V(\bar{j})$$

which in case (a) yields

$$V(R) = \left(\frac{a + 2b \ln \bar{j}}{\bar{j}}\right)^2 V(\bar{j})$$

and in case (b)

$$V(R) = [a\bar{j}^{b-1}(1 + b \ln \bar{j})]^2 V(\bar{j})$$

where $V(\bar{j})$ is the variance of the mean sentence length, i.e. $V(\bar{j}) = V(j)/n$, n being the number of sentences in the sample. Thus we have the possibility to set up two formulas

(c) $$u = \frac{R_1 - R_2}{\sqrt{\left(\frac{a + 2b \ln \bar{j}_1}{\bar{j}_1}\right)^2 V(\bar{j}_1) + \left(\frac{a + 2b \ln \bar{j}_2}{\bar{j}_2}\right)^2 V(\bar{j}_2)}}$$

and

(d) $$u = \frac{R_1 - R_2}{\sqrt{[a\bar{j}_1^{b-1}(1 + b \ln \bar{j}_1)]^2 V(\bar{j}_1) + [a\bar{j}_2^{b-1}(1 + b \ln \bar{j}_2)]^2 V(\bar{j}_2)}}$$

The coefficients a and b need not be equal in the two texts (but in that case $E(R_1) - E(R_2)$ must be put in!), but if Menzerath's law is true they must be equal for the same language or at least for the same class of texts. Of course,

$\bar{j}_x$ in the squared expressions should be replaced by its expected value, but in our case it can be estimated from our samples. Formulas (c) and (d) can be further modified according to whether we consider $\bar{j}_1 = \bar{j}_2$ or not, and according to whether we consider $V(j_1) = V(j_2)$ or not. For small n (c) and (d) are distributed according to Student's t and the number of degrees of freedom can be found e.g. by means of the Welch-procedure.

## References


**Altmann, G.** (1983). H. Arens' "Verborgene Ordnung" und das Menzerathsche Gesetz. In: Faust, M., Harweg, R., Lehfeldt, W., Wienold, G. (eds.), *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik.* Tübingen, Narr: 31-39.

**Flesch, R.F.** (1948). A new readability yardstick. *Journal of Applied Psychology 32, 221-233.*

**Golovin, B.N.** (1971). *Jazyk i statistika* . Moskva, Prosveščenie.

**Herdan, G.** (1964). *Quantitative linguistics.* London, Butterworths.

**Kaasik, Ü., Tuldava, J., Villup, A., Ääremaa, K.** (1977). Eesti tänapäeva ilukirjandus proosa autorikõne lekseemide sagedussõnastik (A frequency dictionary of lexemes of modern Estonian prose fiction). *Acta et Commentationes Universitatis Tartuensis 13, 5-140.*

**Klare, G.K.** (1974-1975). Assessing readability. *Reading Research Quarterly 10, 62-102.*

**Kučera, H., Francis, W.N. (1967).** *Computational analysis of present-day American English.* Providence, R.I., Brown University Press.

**Maas, H.-D.** (1972). Über den Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift für Literaturwissenschaft und Linguistik 2, 73-96.*

**Mistrík, J.** (1967). Matematiko-statističeskie metody v stilistike. *Voprosy jazykoznanija No.3, 42-52.*

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

**Ratkowsky, D.A., Hantrais, L.** (1975). Tables for comparing the richness and structure of vocabulary in texts of different lengths. *Computers and the Humanities 9, 69-75.*

**Somers, H.H.,** (1966). Statistical methods in literary analysis. In: Leed, J. (ed) *The Computer and the Literary Style.* Kent, Ohio 1966, 128-140.

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

**Tuldava, J.** (1975). Ob izmerenii trudnosti teksta. *Acta et Commentationes Universitatis Tartuensis 345, 102-120.*

**Tuldava, J.** (1980). K voprosu ob analitičeskom vyraženii svjazi meždu ob'emom slovarja i ob'emom teksta. *Acta et Commentationes Universitatis Tartuensis 549, 113-144.*

**Tuldava, J.** (1987). *Problemy i metody kvantitativno-sistemnogo issledovanija leksiki.* Tallinn, Valgus.

# Comparison of Texts on Familiar or Unfamiliar Subject Matter

*Jaan Mikk and Jaanus Elts, Tartu*

It is generally accepted that words which keep repeating in various texts are more familiar than words which do not occur so frequently. They are understood more completely and with less effort. Long frequent words have become reduced to short ones, and so familiar words tend to be short. We can suppose that any widely discussed subject matter is mostly expressed by familiar and short words. Texts on subjects unfamiliar to the average reader include more rare words than texts on well-known subjects. Consequently, texts on familiar or unfamiliar subject matter should also have certain differences in their statistical characteristics.

The purpose of this paper is to establish the differences in the statistical characteristics of two kinds of texts: those on some well-known subject and those on some unfamiliar subject. The degree of familiarity of a set of texts is established, the texts are analyzed and the results compared.

We analyzed 48 Russian texts of about 1900 letter spaces each in popular-scientific books on biology. One or two texts were picked at random from each book to have more variety in text characteristics.

## Analysis of texts

All the texts were computer-analyzed. The analysis included the following procedures:

1. Word length was measured in letter spaces. Sentence length was measured in words and in letter spaces.

2. The principal form of every word in the text was determined with the help of the programs for morphological analysis worked out in Kiev by N.A. Darčuk and her colleagues (Avtomatizacija..., 1984). The part of speech the words in the text belonged to and the frequency of occurrence in the text were found.

3. The frequency values of the words in our texts were found in a Russian frequency dictionary given to us by D.A. Buchštab from Moscow University.



4. As the degree of substantival abstractness and the number of terms proved to have a significant influence upon text comprehension in our study of comprehension of physics texts (Kukemelk & Mikk 1991), the degree of abstractness of every noun in the text and the role of terms were measured.

Ways to measure the degree of abstractness in text have been proposed by R.F.Flesch (1950) and P.J.Gillie (1957). Flesch counted personal (concrete) words in the text and Gillie counted the abstract suffixes of nouns in the text. We have also suggested a scale of abstractness of nouns. As our way of measuring abstractness correlated best with text difficulty (Mikk 1974: 134-135), we measured substantival abstractness by grading the nouns in the text as follows:

1 - concrete nouns designating things directly perceivable by senses (e.g. man, stone);
2 - nouns designating phenomena and processes perceivable by senses (e.g. rain, light);
3 - abstract nouns designating objects and notions imperceptible directly by senses (e.g. evolution, cell).

The nouns in the text were grouped into three classes to measure "terminologicality" of text (Elts 1992a):

1 - nouns that are not terms;
2 - terms encountered in everyday speech;
3 - terms that are not used in everyday speech (scientific terms).

We measured about 190 text characteristics. We will not give the full list of the characteristics. A list of the most informative characteristics and their mean values are given in Table 1 (some more characteristics are presented in Elts & Mikk 1991).

## Experimental determination of text familiarity levels

The experiment was carried out in two secondary schools in Tartu (Estonia) in 1989. The subjects were 148 pupils in the 7th, 8th and 10th grades in Russian-speaking schools. Russian was the mother tongue of the testees and all the texts and questions were in Russian. As all the pupils were expected to participate in all the tests, incomplete test answers were not included in the experiment analysis and finally the answers of 124 testees were analysed.

To establish the level of subject familiarity, a test of free response questions on the subject matter was administered. The test was of four variants, eight to ten questions in each. To compare texts on their familiarity the questions in the

test must represent all possible questions on the text. A representative sample of questions contains some dozens of them (Mikk 1988). Question sets were drawn up on all the texts analyzed. The variants were randomly distributed among the testees. The testee had not read the text before. The testee was given one point for any correct answer. The maximum number of points a testee could score was eight to ten. Each text was characterized by the mean value of the correct answers in percent which was later used as the text subject familiarity index.

**Results of the investigation**

We calculated linear correlations between the index of familiarity and the text characteristics. Many characteristics were of the same type, e.g. No 11 - the proportion of sentences of 13 or more words, No 12 - the proportion of sentences of 14 or more words, and so on. All the characteristics were determined to find the best one of them. The data suggested that the optimal sentence length for the 8-9 grade pupils was 10-11 words (Elts & Mikk 1991). As characters of the same type have good correlation ($r_{11,12} = 0.97$), we decided to discuss only one of them in this paper. The characteristic with the highest correlation indices were included in Table 1. All the correlations in Table 1 were statistically significant at the 0.01 level (with 46 d.f.). The coefficients with absolute values over 0.46 were significant at the 0.001 level.

The mean percentage of the correct answers prior to the reading of the text was 5.48 and its standard deviation was 5.53.

There were many text characteristics of significant correlation with text familiarity. To study how different the texts were we compared the mean values of the characteristics of 12 texts on some unfamiliar subject with those of 12 texts on some familiar subject. A text was considered familiar if the mean values of the correct text answers prior to the reading of the text were within the range of 7.5 to 24.5 percent (mean value = 13.6). In texts on unfamiliar subjects the value could not exceed 1.2 percent (mean value = 0.8). The results of the comparison are given in Table 2.

All the differences except one between the mean values characterizing the texts on unfamiliar/familiar subjects in Table 2 are statistically significant (d.f. = 22).

We worked out some regression equations to predict the degree of subject familiarity in a text by text characteristics. As the correlations between the text characters are high (Table 3), it was not necessary to include all of them in the regression equation. For example, the correlation coefficient between Characters No 31 and No 65 was 0.79. If we include the first one in the formula the second one may be left out without essential loss in the accuracy of the formula.



Table 1
Characteristics of texts and their correlations with text familiarity

| No. | Characteristics | Mean | STD | Corr. |
|---|---|---|---|---|
| 11 | Proportion of sentences of 13 or more words | 0.53 | 0.22 | -0.44 |
| 31 | Proportion of sentences of 90 or more letter spaces | 0.61 | 0.24 | -0.53 |
| 53 | Proportion of segments of 9 or more words between two consecutive verbs | 0.40 | 0.15 | -0.37 |
| 65 | Nominal phrases of 2 or more nouns in the sentence | 1.21 | 0.58 | -0.47 |
| 78 | Proportion of words of 9 or more letters | 0.25 | 0.07 | -0.53 |
| 90 | Mean length of sentences in letter spaces | 118 | 36.7 | -0.44 |
| 91 | Mean number of letters in a word | 6.27 | 0.60 | -0.54 |
| 93 | Modification ratio | 0.44 | 0.09 | 0.38 |
| 97 | Mean frequency value of the words according to the frequency dictionary | 969 | 207 | 0.37 |
| 101 | Proportion of words with frequency values below 30 in the frequency dictionary | 0.63 | 0.05 | -0.42 |
| 103 | Mean frequency value of the nouns according to the frequency dictionary | 26 | 17 | 0.58 |
| 104 | Mean rate of repeating nouns in the text | 1.35 | 0.13 | -0.42 |
| 107 | Proportion of nouns with frequency values below 30 in the frequency dictionary | 0.85 | 0.07 | -0.53 |
| 109 | Proportion of nouns in the text | 0.35 | 0.05 | -0.49 |
| 233 | Mean terminological index of nouns | 1.53 | 0.24 | -0.45 |
| 235 | Percentage of nouns of the second grade of abstractness | 28.2 | 11.2 | 0.37 |
| 236 | Percentage of abstract nouns | 23.6 | 16.0 | -0.49 |
| 242 | Percentage of scientific terms | 13.4 | 11.2 | -0.54 |

The computer program Statgraphics was used to calculate the equations. The first argument in the formula is the characteristic of the highest correlation with subject familiarity. The second argument is the characteristic of the highest partial correlation (Sachs 1982: 456) and so on.

We made two regression analyses. The first was concerned with the characteristics with significant correlation of 0.001 with the subject familiarity level ($r \geq 0.46$). The regression equation was as follows:

$$Y_{210} = 7.12 - 9.35X_{31} + 0.153X_{103} \quad (1)$$

The indices in the formula show the number of the corresponding characteristics in Table 1. The standard error of the estimation is 4.06 and the coefficient of multiple correlation is 0.69. The standard error of the constant is 2.10 and the standard error of the coefficients is 2.55 and 0.035 respectively.

Table 2
Comparison of the mean values of the characteristics in the text on familiar/unfamiliar subjects texts

| No. | Characteristics | Mean value in | | t- test of difference |
|---|---|---|---|---|
| | | unfamiliar texts | familiar texts | |
| 11 | Proportion of sentences of 13 or more words | 0.72 | 0.43 | 3.74 |
| 31 | Proportion of sentences of 90 or more letter spaces | 0.78 | 0.41 | 4.88 |
| 53 | Proportion of segments of 9 or more words between two consecutive verbs | 0.47 | 0.30 | 3.66 |
| 65 | Nominal phrases of 2 or more nouns in the sentence | 1.63 | 0.82 | 3.38 |
| 78 | Proportion of words of 9 or more letters | 0.31 | 0.21 | 4.07 |
| 90 | Mean length of sentences in letter spaces | 149 | 93 | 4.25 |
| 91 | Mean number of letters in a word | 6.74 | 5.88 | 4.11 |
| 93 | Modification ratio | 0.45 | 0.51 | 1.80 |
| 97 | Mean frequency value of the words according to the frequency dictionary | 861 | 1052 | 3.08 |
| 101 | Proportion of words with frequency values below 30 in the frequency dictionary | 0.66 | 0.60 | 3.39 |
| 103 | Mean frequency value of the nouns according to the frequency dictionary | 14.0 | 40.5 | 4.37 |
| 104 | Mean rate of repeating nouns in the text | 1.42 | 1.26 | 3.23 |
| 107 | Proportion of nouns with frequency values below 30 in the frequency dictionary | 0.89 | 0.79 | 4.64 |
| 109 | Proportion of nouns in the text | 0.37 | 0.32 | 2.45 |
| 233 | Mean terminological index of nouns | 1.73 | 1.40 | 3.76 |
| 235 | Percentage of nouns of the second grade of abstractness | 23.3 | 34.5 | 2.62 |
| 236 | Percentage of abstract nouns | 34.5 | 12.2 | 4.74 |
| 242 | Percentage of scientific terms | 23.5 | 5.5 | 5.28 |

The other regression analysis concerned the characteristic with 0.01 significance level (correlation ≥ 0.37) with subject familiarity. The results of the analysis are as follows:

$$Y_{210} = 12.8 - 3.71X_{91} + 21.7X_{93} + 0.098X_{103} + 0.13X_{235}. \quad (2)$$

The standard error of estimates is 3.57, the coefficient of multiple correlation is 0.77, the standard errors of the coefficients are 0.97, 5.83, 0.035 and 0.051 respectively.



Table 3
Correlations between text characteristics[1]

| No | 11 | 31 | 53 | 65 | 78 | 90 | 91 | 93 | 97 | 101 | 103 | 104 | 107 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 31 | .95 | | | | | | | | | | | | |
| 53 | .44 | .50 | | | | | | | | | | | |
| 65 | .81 | .79 | .57 | | | | | | | | | | |
| 78 | .61 | .73 | .59 | .68 | | | | | | | | | |
| 90 | .90 | .91 | .51 | .89 | .69 | | | | | | | | |
| 91 | .59 | .71 | .61 | .68 | .96 | .69 | | | | | | | |
| 93 | -.38 | -.34 | -.04 | -.43 | -.14 | -.26 | -.05 | | | | | | |
| 97 | -.60 | -.56 | -.41 | -.53 | -.58 | -.52 | -.61 | -.20 | | | | | |
| 101 | .63 | .66 | .51 | .74 | .68 | .65 | .69 | -.28 | -.75 | | | | |
| 103 | -.23 | -.26 | -.35 | -.33 | -.40 | -.25 | -.40 | .07 | .18 | -.40 | | | |
| 104 | .30 | .31 | .16 | .38 | .23 | .29 | .25 | -.22 | -.13 | .02 | -.35 | | |
| 107 | .32 | .35 | .37 | .42 | .32 | .33 | .32 | -.24 | -.25 | .56 | -.78 | .24 | |
| 109 | .56 | .57 | .62 | .81 | .68 | .60 | .67 | -.47 | -.66 | .72 | -.37 | .25 | .36 |
| 233 | .18 | .28 | .47 | .39 | .44 | .25 | .47 | -.04 | -.29 | .34 | -.46 | .53 | .44 |
| 235 | .02 | -.02 | -.16 | .03 | .03 | .00 | -.02 | -.03 | -.11 | .00 | .36 | -.19 | -.37 |
| 236 | .44 | .53 | .42 | .47 | .62 | .50 | .59 | -.10 | -.41 | .42 | -.36 | .39 | .38 |
| 242 | .30 | .41 | .62 | .51 | .60 | .40 | .60 | -.13 | -.39 | .47 | -.44 | .42 | .49 |

Table 3, cont.

| No. | 109 | 233 | 235 | 236 |
|---|---|---|---|---|
| 233 | .39 | | | |
| 235 | .06 | -.41 | | |
| 236 | .42 | .54 | -.22 | |
| 242 | .52 | .92 | -.33 | .60 |

## Discussion

The statistical text characteristics correlated well with the percentage of correct answers in the text (Table 1). What might account for it? Why did the students who had not read the texts give more correct answers to the questions on the texts of shorter sentences and fewer terms?

The explanation may lie in a third factor in the process: the familiarity level of the subject matter in the community. Familiar topics tend to be treated in shorter words and sentences, with the use of fewer terms and abstract nouns. Naturally, familiar topics will yield more correct answers than unfamiliar topics. So the characteristics of nonread text correlate with percent of correct answers on its content. This is in accordance with the conclusion of R.C.Anderson and A.Davison: "The presence of long sentences and complex words in a text in some way reflects or is correlated with complexities of subject matter, but need

[1] The numbers of the characteristics are the same as in Table 1. The correlation coefficients are printed without zero. The coefficients with an absolute value over 0.23 are significant at the 0.05 level, those over 0.37 - at the 0.01 level and those over 0.46 at the 0.001 level.

not directly cause a text to be difficult" (Anderson & Davison 1988: 48).

Of course we are speaking here about statistical relationships. Writing "familiar topics" we mean that in general people know more about the content of texts on these topics. We consider here statistical averages of two dimensions: people and text.

We computed about two hundred textual characteristics to test a hypothesis about their relationship with subject familiarity and to find the best predictors of subject familiarity. The hypothesis proved true for about 50 characteristics.

We found the following groups of text characteristics in which texts on familiar/unfamiliar topics differ.

1. **Length of sentences.** Sentences proved to be shorter in texts on familiar topics. In our experiment texts on unfamiliar topics contained sentences of 149 letter spaces and texts on familiar topics 93 letter spaces on the average (Table 2). Measuring the sentence length in letter spaces proved to be more informative than measuring it in words (cf. the correlations of Characteristics No 11 and 31 in Table 1). A similar result was obtained in testing texts in Estonian (Mikk 1974: 148). The characteristic of the proportion of long sentences in the text is more informative than the average sentence length (the proportion of the sentences of 90 or more character spaces has the highest correlation in this group of characteristics).

There are some other characteristics related to the length of sentences. Texts on unfamiliar topics contain more noun phrases in the sentence and there are more verbs distanced from each other (Characteristics No 53 and 65 in Tables 1 and 2.)

2. **Length of words.** Texts on familiar topics contain shorter words. The texts on unfamiliar topics in our experiment had an average word length of 6.74 letters and the texts on familiar topics 5.88 letters.

3. **Frequency of words in speech.** The text on familiar topics contained more words of high frequency on the average. Among the text characteristics of word frequency levels the most informative was the mean frequency of nouns in the text - its correlation with the familiarity index was the highest in its absolute value in Table 1 (Characteristic No 103, correlation of 0.58). The mean frequency value of nouns in the frequency dictionary was 14 for the texts on unfamiliar topics and 40 for the texts on familiar topics (Table 2). The mean frequency of nouns proved to be much smaller than the mean frequency of all words in our texts (see Table 1, the mean value for all the words was 969).

It is surprising that the mean rate of repeating nouns had a negative correlation with the subject familiarity ($r = -0.42$) (Table 1). We might suppose that the more nouns are repeated in the text the more familiar they should be and the correlation ought to be positive. In fact the average number of repeats was too small in our texts (only 1.35) and the nouns had not become familiar. Obviously nouns are repeated more often in texts of unfamiliar topics. The vocabulary used in speaking about familiar topics tends to be richer. The same tendency could be noted when texts on physics were studied (Kukemelk & Mikk 1991: 75).



4. **Abstractness of nouns.** The texts on unfamiliar topics contained more abstract nouns. The texts on unfamiliar topics contained 34 per cent of nouns of Abstractness Class No 3, while the texts on familiar topics had only 12 per cent of the nouns in Class 3.

5. **Number of terms in the text.** Scientific terms, which are not found in everyday speech, are present in texts of lower subject familiarity. There were 23 percent scientific substantival terms in the texts on unfamiliar topics, and only 5 percent in the texts on familiar topics.

We must note that the texts on familiar topics are only relatively familiar with their 13.6 percent on familiarity index in our experiment. The index was not high, there ought to be texts the subject matter of which the students should be more familiar with. The authors of the texts know the topic much better. And yet the order topics are arranged as to their familiarity is nearly the same for experts and learners.

Although we have excluded characteristics of the same kind from Table 1 there are high intercorrelations in Table 3. That allows us to group the characteristics into

1) characteristics of sentence and word length (No 11-91);
2) frequency characteristics of nouns (No 103, 107);
3) characteristics of abstractness and "terminologicality" of nouns (No 233-242).

There are some characteristics that cannot be grouped with these groups, for example No 93. It is interesting to note that the correlation between the mean frequency value of all the words in the text (No 97) and that of the nouns (No 103) is not significant ($r = 0.18$).

High intercorrelation of most of the characters and above-mentioned groups indicates that there is no need for all of them to be used in the formulas for the evaluation of subject familiarity levels. A characteristic accounts for the effect of others that have high correlation with it. For example, the word length and the sentence length correlation was 0.71. Only one of these characters will add a statistically significant increase in the exactness of the formula. Our formulas for predicting the subject familiarity level of texts have 2 and 4 arguments respectively.

The first formula contains only the highly valid predictors of subject matter familiarity in a text: the indices of sentence length and word frequency. Texts

on familiar topics contain more nouns of high frequency rate and fewer long sentences. The formula accounts for 47% of the subject familiarity in the text. The standard error of the estimate is not very high, its absolute value being 4 %, yet it should be considered high if compared with the mean value of subject matter familiarity, which is 5.5% of the correct answers given in testing.

The second formula contains two more arguments. The coefficients by those arguments are statistically significant ($t > 2.6$). The constant is not significant, and yet it cannot be excluded as that would cause a systematic bias. Formula (2) contains the argument of word length instead of sentence length. The added arguments are the modification ratio (No 93) and the percentage of nouns of Abstractness Class 2 (No 235). The modification ratio was calculated as the proportion of the percentage of adjectives and adverbs to the percentage of nouns and verbs. The index was introduced by O.A.Wiio (1968: 43). We may suppose that a greater number of adjectives and adverbs makes sentences longer and the subject matter less familiar. We can see in Table 1 and Formula (2) that the modification ratio is higher in texts on more familiar topics. It may be caused by the largest group of parts of speech in the modification ratio. There were 35% nouns in the text and their mean frequency in speech was low (26) when compared with the mean frequency value of all words (969) (Table 1). The more nouns there were in the text the less known the subject matter was and the lower was the modification ratio. So we can see a positive correlation between the modification ratio and text familiarity in Formula (2).

The coefficient of the percentage of nouns of Abstractness Class 2 in Formula (2) has the lowest t value - 2.6. The positive correlation indicates that the nouns designating phenomena and processes perceivable by senses are more frequent in texts on familiar topics.

The formulas can be used to evaluate subject familiarity level in study texts for eight-grades. The optimal level would be about 20 per cent. Only one text in our experiment met the requirement.

We have analyzed texts but we have not analyzed questions on them. The influence of question characteristics on the correctness of answers was studied earlier (Mikk 1988). The most important factor - the question form ($r = 0.28$) - was held constant in the present investigation. Other factors - the number of terms in the questions and others - are obviously different in different texts. As questions have to be based on the text in our experiment, they reflect the subject familiarity level of the text.

Our experiment was based on the methodology of readability investigation (Klare 1963; Kukemelk & Mikk 1991; Tuldava 1975). The results are in good correlation. The percentage of correct answers given after reading a text depends on the same text characteristics as does the percentage of correct answers given without reading the text. That suggests that the indices depend on the degree of subject matter familiarity for the given group of people.

There are many rules of clear writing (Baumann, Geiling & Nestler 1987; Flesch 1946; Groeben 1982) which have repeatedly been proven to be effective. And yet there were cases when the rules failed. The results of our investigation indicate that for a text to be clearly written the subject matter must be familiar. It is not enough to write in short sentences and short words. A text is easy to comprehend if it contains frequently occurring words, few terms and few abstract nouns (Elts 1992b).



It must be noted that our investigation reflects statistical tendencies in the language. That means that there are always exceptions to prove the next rule. Very clear texts on difficult things testify to the mastery of their authors. They should serve as examples for textbook compilers.

## Conclusion

Texts with familiar/unfamiliar subject matter differ in many text characteristics. Texts on familiar topics are made up of short words and sentences, and they contain words of higher frequency of occurrence, few terms and few abstract nouns. There may be other differences in the characteristics between these two types of texts. The values of some of these characteristics may be used to calculate the required familiarity level of a study text. Texts ought to be written at the optimal familiarity level to make for the maximal information gain.

## References

**Anderson R. C., Davison A.** (1988). Conceptual and empirical bases of readability formulas. In: Anderson, R.C., Davison, A. (eds.), *Linguistic complexity and text comprehension.* Hillsdale, NJ, Erlbaum: 23-54.

**Automatizacia analyza naučnogo texta** (Automatization of the analysis of scientific text) (1984). Kiev, Naukova dumka.

**Baumann M., Geiling U., Nestler K.** (1987). Katalog verständnishemmender Textmarkmale ("Storstellenkatalog"). *Informationen zu Schulbuchfragen 56, 36-55.*

**Elts, J.** (1992a). A readability formula for texts on biology. Psychological problems of reading. *Theses of papers for the international scientific conference, Vilnius, 6-7 May.* Vilnius: 42-44.

**Elts, J.** (1992b). Bioloogiatekstide loetavusvalem (Comprehensibility of biology texts). *Haridus 12, 23-26.*

**Elts J., Mikk J.** (1991). Millest sôltub bioloogiatekstide omandamine (On what does the learning of biology texts depend). *Oppekirjandus kooliuuenduses* (School literature in reforms of school), Tartu, 5-12.

**Flesch R.F.** (1946). *The art of plain talk.* New York-London, Harper and Brothers Publishers.

**Flesch R.F.** (1950). Measuring the level of abstraction. *Journal of Applied Psychology 34, 384-390.*

**Gillie P.J.** (1957). A simplified formula for measuring abstraction in writing. *Journal of Applied Psychology 41, 214-217.*

**Groeben N.** (1982). *Leserpsychologie: Textverständnis - Textverständlichkeit.* Aschendorff, Münster Westfalen.

**Klare G.R.** (1963). *The measurement of readability.* Iowa, Iowa State University.

**Kukemelk H., Mikk J.** (1991). The prognosticating effectivity of learning a text in physics. *Acta et Commentationes Universitatis Tartuensis 926, 51-81.*

**Mikk J.** (1974). Metodika razrabotki formul čitabelnosti (Methods of elaborating of readability formulae). *Sovetskaja pedagogika i škola 9, 78-163.*

**Mikk J.** (1988). O proverke dostupnosti učebnika s pomoščju postanovki voprosov (Measuring difficulty of textbooks with asking questions). *Sovetskaja pedagogika i škola 20, 80-98.*

**Sachs L.** (1982). *Applied statistics. A Handbook of statistics.* New York-Heidelberg-Berlin, Springer.

**Tuldava J.** (1975). Ob izmerenii trudnosti teksta (On measuring the difficulty of texts). *Acta et Commentationes Universitatis Tartuensis 345, 102-120.*

**Wiio O.A.** (1968). Readability comprehension and readership. *Acta Universitatis Tamperensis (Tampere) ser. A, 22.*

# Measuring Text Difficulty in Japanese
## Different Tools - Same Results?

*Undine Roos, Essen/Bochum*

### 1. Introduction

In this article, four tools for measuring text difficulty will be compared, using the Japanese language as an example. The four tools are

(I) Type-Token-Ratio,
(II) the list of Jōyō-Kanji,
(III) Fog-Index,
(IV) Tuldava's Index.

Text difficulty is here defined as the difficulty for the reader of a text. Four neutral texts (i.e. texts for which no special knowledge is necessary) were used, with the intention of ranking them from the most easy to the most difficult one. The texts have the following topics: Fire department, mice, kindergartens and realism. They are translations from English texts which were used by Raatz & Klein-Braley (1983) as completion tests for measuring language proficiency in English. (The texts are shown in the appendix in the Japanese and the English version).

The purpose of this study is to show whether the "traditional" formulae Type-Token-Ratio, Fog and Tuldava can also be adapted to Japanese texts, or whether there are more valid means for measuring text difficulty in Japanese, as e.g. with the list of Jōyō-Kanji. Empirical research with native speakers of Japanese in the form of a C-test will demonstrate this.

### 2. Type-Token-Ratio

One possibility for measuring text difficulty is the Type-Token-Ratio, which is used to show the relationship between the number of different words (or characters) in a text and the length of the text, i.e. the total number of words (or characters). The more rapid the increase of formula (1), the more new information is transmitted in the text and the more the recipient has to decode; i.e.: ***b*** is a measure of the flow of information in the text and perhaps a measure of text difficulty for the recipient, because a quick flow of Type-Token-Information does not necessarily mean high text difficulty.

The most simple form of Type-Token dependence was derived by Herdan (1960) and can be shown as

$$Type = Token^b . \quad (1)$$

After logarithmic linearisation we get

$$ln\ Type = b\ ln\ Token$$

i.e. ***b*** is the coefficient of linear regression through the origin. As we do not have an unambiguous definition of the concept "word" for Japanese, we get - with respect to the peculiarities of this language, among others the property of agglutination - three possibilities for defining types and tokens:

(a) we take the characters as a basis;

(b) we try to define "word", regarding postpositions as bound forms (cf. Lewin 1975:40ff);

(c) we try to define "word", regarding postpositions as free markers (cf. Hayashi et al. 1982:582ff).

For the texts examined by these methods we get Table 1, where the ***b***'s were arrived at by means of the method of least squares.

To prove whether the flow of information is different in the four texts we have to compare the parameters ***b***. The following formula was used for this purpose.

$$t = \frac{b_1 - b_2}{s\sqrt{\frac{1}{\sum_i (x_{i1} - \bar{x}_1)^2} + \frac{1}{\sum_i (x_{i2} - \bar{x}_2)^2}}}$$

with

$$s^2 = \frac{\sum_i (y_{i1} - \hat{y}_{i1})^2 + \sum_i (y_{i2} - \hat{y}_{i2})^2}{n_1 + n_2 - 2}$$

where x = ln Token und y = ln Type.
Testing the difference of the lowest b (in "Fire department") and the highest b (in "Mice") we get a value of

$$t = 0.67;\ df = 217.$$



Table 1
Types, Tokens and b for the four different Japanese texts

| | Fire department | Mice | Kindergartens | Realism |
|---|---|---|---|---|
| Types | 65 | 122 | 104 | 100 |
| Tokens | 160 | 268 | 216 | 205 |
| ***b*** | .8832 | .8936 | .9065 | .9055 |
| Rank | 1 | 2 | 4 | 3 |

As this value is not higher than the critical value of $t_{217;.01} = 2.58$, it can be stated that these two extreme ***b***'s do not differ from each other significantly.

Consequently, a comparison of the remaining ***b***-values is not necessary. Nonetheless, the four texts can be ranked according to the values of ***b*** in the way shown in Table 1 in the "rank" row.

### 3. Measuring text difficulty with the list of Jōyō-Kanji

A further possibility of measuring text difficulty is the list of Jōyō-Kanji.[1] In Table 2 all Kanji of texts 1 to 4 were classified according to this list, where A1 to A6 represent the elementary 996 characters which have to be learnt in the six years of primary school, *chuu* for the characters of middle school, *koo* for the characters of high school, *Z* for additional characters and *S* for the characters not belonging to the list of Jōyō-Kanji. The first sentence of "Kindergartens" may serve as an example.

Example:

1 6 4 6 2 2 2 2 2 3 2 4
子供達は幼年学校に入学した後も、家庭にお

1 5 2 5 5 6
いて正確な知識のストックを増やし得るよう

3 4 3 6 中 4
な場に直面する状況が続く。しかしながら、

[1] Jōyō-Kanji means: Characters for common use. The list, dating from 1981, includes 1945 characters which are - according to the Japanese Ministry of Education - the most important and the most frequently-used ones. The list can be found in Kenbo et al. (1982:1260ff).

中 2 5 6 5 中 5 3 5

彼らの知識は未だ無秩序で取り留めがない。

Table 2 shows that the four texts may differ in difficulty. This proposition can be proved by a Chi-square-test (cf. Siegel 1976:42). We get $X^2 = 53.74$ with 27 df. As this result is higher than the critical value of $\chi^2_{.01(27)} = 47.00$ we can say that the distribution of Kanji in the four texts is not homogenous; in other words: the difficulty of the texts is significantly different - as measured by the characters.

Table 2
Classification of Kanji according to the list of Jōyō-Kanji

| | Fire department | Mice | Kindergartens | Realism |
|---|---|---|---|---|
| A1 | 8 | 6 | 9 | 14 |
| A2 | 7 | 12 | 14 | 12 |
| A3 | 15 | 22 | 9 | 19 |
| A4 | 16 | 19 | 14 | 21 |
| A5 | 9 | 5 | 20 | 11 |
| A6 | 4 | 3 | 12 | 5 |
| *Chuu* | 0 | 12 | 7 | 4 |
| *Koo* | 0 | 0 | 0 | 1 |
| *Z* | 0 | 1 | 0 | 1 |
| *S* | 0 | 6 | 1 | 1 |

The order of the texts can be determined by the means of these distributions. We get the following values:

Text 1: 3.38 Text 2: 4.28 Text 3: 4.08 Text 4: 3.58

So the order of the texts, in increasing difficulty, runs:

(easy) (diff.)
T1/Fire engine T4/Realism T3/Kindergarten T2/Mice

Measuring text difficulty by the list of Jōyō-Kanji leads to an entirely different result from the measurement with the Type-Token-Ratio.



## 4. Empirical research with Japanese natives

In order to determine the real order of the four texts, they were given in the form of completion texts to native speakers of Japanese. The selection of the spaces was governed by the C-test-Principle (cf. e.g. Grotjahn 1987, Raatz 1985, Raatz & Klein-Braley 1985). The C-test-Principle is based on that of the Cloze-Test, which was developed by W.C. Taylor (1953) for measuring the readability of texts. The C-test-Principle runs as follows: Beginning with the second word of the second sentence the second half of every second word is deleted, until you have created 25 blanks in each text. Natives have to solve at least 95% of the text correctly. As there is no unambiguous definition of the concept "word" for Japanese, the C-test-Principle has been modified so that for Japanese the lower part of every second character is deleted (cf. Roos 1990, Roos 1992). The second sentence of text "Realism" may serve as an example.[2]

Example:

しゃし小説に関して言いたいことは、

普通、日常生活の精密な描写を意味す。

The texts were presented to the testees in the following order, which was obtained by chance:

Kindergartens - Realism - Mice - Fire department

The tested persons were 49 students from the University of Ōsaka (Japan). The mean of the right answers, measured in percent (cf. Andersen 1985:138), was:

| | |
|---|---|
| Text Kindergarten | 99.6% |
| Text Realism | 99.7% |
| Text Mice | 97.3% |
| Text Fire service | 99.8% |

These values show that the texts fulfil the criterion of at least 95% correct answers from natives; furthermore they can be ordered the same way as was determined by the list of Jōyō-Kanji; a correspondence with the order arrived at by Type-Token-Ratio is not given.

[2] The first sentence of the text is never distorted by deletions.

### 4.1. Further Readability Formulae

The Type-Token-Ratio is only one of several means for measuring text difficulty. In this chapter two additional formulae will be applied to the four texts to determine whether it is necessary to use the list of Jōyō-Kanji, or if the same conclusion can be reached with a "traditional" readability formula.

First, there is the Fog-Index (cf. Alderson & Urquhart 1986:xxii) which runs as follows:

$$R = 0.4\ (k + j)$$

where k = percentage of words with three or more syllables and
j = mean sentence length in words.

The coefficient R can be interpreted as follows:

| | |
|---|---|
| 0-12 | Easy |
| 13-16 | Undergraduate |
| >16 | Postgraduate. |

For the four texts we get:[3]

| | Fire dept. | Mice | Kindergartens | Realism |
|---|---|---|---|---|
| R | 8.00 | 17.86 | 21.60 | 16.40 |

The Fog-Index shows that there is a difference concerning the level of the texts (Text 1: easy, Text 4: Undergraduate, Text 2 and Text 3: Postgraduate), but the order is completely different from the one arrived at by the C-test and the list of Jōyō-Kanji.

Finally, Tuldava's Index was calculated (cf. Tuldava 1975). The formula is:

$$R = i \log j$$

with i = mean word length in syllables and
j = mean sentence length in words

[3] For the calculation of this and the following index "word" was defined as "character" for achieving comparability to the list of Jōyō-Kanji.



Table 3
Summary of the calculations

| | Fire dept. | Mice | Realism | Kindergarten |
|---|---|---|---|---|
| C-test | 99.8% | 97.3% | 99.7% | 99.6% |
| Rank | *(1)* | *(4)* | *(2)* | *(3)* |
| Type-Token | .86 | .89 | .905 | .906 |
| Rank | *(1)* | *(2)* | *(3)* | *(4)* |
| Jōyō-list | *(1)* | *(4)* | *(2)* | *(3)* |
| Fog R | 8.0 | 17.86 | 16.40 | 21.60 |
| Rank | *(1)* | *(3)* | *(2)* | *(4)* |
| Tuldava R | 1.30 | 1.65 | 1.61 | 1.73 |
| Rank | *(1)* | *(3)* | *(2)* | *(4)* |

The values obtained are:

| | Fire dept. | Mice | Kindergartens | Realism |
|---|---|---|---|---|
| R | 1.30 | 1.65 | 1.73 | 1.61 |

This order also differs from that of the C-test and the list of Jōyō-Kanji, as does the Fog-Index.

Table 3 summarizes all the calculations. The numbers in brackets show the rank of the texts.

### 5. Conclusions

The calculation of different indices for measuring text difficulty shows that the order of the four Japanese texts could be predicted exactly only by the list of Jōyō-Kanji. As mentioned briefly in the introduction, there are theoretically two possibilities for division into words for Japanese. These two possibilities were also taken into account and tested. The results are summarized in Tables 4 and 5. The tables show the empirically obtained C-test-values, the order of the texts according to the list of Jōyō-Kanji, the Type-Token-Ratio, the Fog-Index and Tuldava' Index.

It becomes evident from the tables that the C-test-Principle for Japanese does

not use words as the basis for deletion but characters, as native speakers of Japanese were not able to reconstruct texts, based on words, more than 95% correctly. The fact that the order could not be predicted by any coefficient is therefore irrelevant for the two versions summarized in Tables 4 and 5.

Table 4
Division into words with Postposition as a free form

| | Fire department | Mice | Realism | Kindergartens |
|---|---|---|---|---|
| C-test | 72.4% | 80.9% | 64.9% | 64.2% |
| Rank | *(2)* | *(1)* | *(3)* | *(4)* |
| Type-Token | .88 | .934 | .933 | .935 |
| Rank | *(1)* | *(3)* | *(2)* | *(4)* |
| Jōyō-list | *(1)* | *(4)* | *(2)* | *(3)* |
| Fog R | 15.2 | 20.6 | 19.34 | 17.56 |
| Rank | *(1)* | *(4)* | *(3)* | *(2)* |

Table 5
Division into words with Postposition as a bound form

| | Fire department | Mice | Realism | Kindergartens |
|---|---|---|---|---|
| C-test | 53.3% | 58.0% | 51.0% | 54.5% |
| Rank | *(2)* | *(4)* | *(3)* | *(1)* |
| Type-Token | .975 | .9977 | .9976 | .986 |
| Rank | *(1)* | *(4)* | *(3)* | *(2)* |
| Jōyō-list | *(1)* | *(4)* | *(2)* | *(3)* |
| Fog R | 29.1 | 34.41 | 35.3 | 37.24 |
| Rank | *(1)* | *(2)* | *(3)* | *(4)* |
| Tuldava R | 5.36 | 6.20 | 5.76 | 6.18 |
| Rank | *(1)* | *(4)* | *(2)* | *(3)* |



Finally, Kendall's coefficient of concordance was computed for the values of Table 3 (cf. Siegel 1976) to determine whether the orders differ significantly from each other. The following formula was used:

$$W = \frac{s}{k^2(N^3 - N)/12}$$

where s = squared deviations of the ranges from the mean range
k = number of columns
N = number of rows.

The denominator of W is the maximum sum of squared deviations, i.e. the value in case of perfect identity of the rows. For Table 3 we get the following values

s = 101, W = 0.808.

Transforming s to a Chi-square results in

$$X^2_{3;.05} = 12.12$$

which is greater than the critical value of 7.72. So one can conclude that the orders ascertained by the C-test, the list of Jōyō-Kanji, the Type-Token-Ratio, the Fog-Index and Tuldava's Index differ significantly. It becomes obvious from Table 3 that only the values of the C-test and the list of Jōyō-Kanji are identical. For this reason Tau was computed a second time, without the values of the list of Jōyō-Kanji, and a third time without the C-test and the list of Jōyō-Kanji. For the second computation we get:

s = 62, W = 0.775, $X^2 = 9.3$

and for the third computation:

s = 41, W = 0.911, $X^2 = 8.17$.

As all Chi-square values are higher than the critical value, one can conclude that the ranges obtained by the different indices differ significantly from that obtained by the C-test. It was only by using the list of Jōyō-Kanji that an exact prediction could be made.

## 6. Further Perspectives

The investigations reported on here have made it clear that the list of Jōyō-Kanji seems to be a more valid means for ordering Japanese texts according to their difficulty than are "traditional" coefficients, as, e.g. the Type-Token-Ratio, Tuldava' Index and the Fog-Index, under the assumption that the Japanese language is a character-based language, which - as Tables 4 and 5 show - seems to be permissible and in fact sensible. This hypothesis of course needs to be verified on a larger number of texts and especially using longer texts.

If the Chinese and Korean languages do exactly the same thing as Japanese, one might think about a more exact index for measuring text difficulty in these languages, because it seems obvious that text difficulty depends not only on the characters but also on the number and length of sentences, the number of subclauses and so on (cf. e.g. Davison 1986). Further research should deal with these questions.

## References

**Alderson, C. & Urquhart, A.H.** (1986). *Reading in a Foreign Language*. London, New York: Longman.

**Andersen, S.** (1985). *Sprachliche Verständlichkeit und Wahrscheinlichkeit.* Bochum: Brockmeyer.

**Davison, A.** (1986). *Readability - the situation today*. Reading Education Report No. 70. University of Illinois.

**Grotjahn, R.** (1987). How to Construct and Evaluate a C-test. A Discussion of some Problems and some Statistical Analyses. In: Grotjahn, R., Klein-Braley, C. & Stevenson, D.K. (eds.), *Taking Their Measure. The Validity and Validation of Language Tests, 219-253*. Bochum: Brockmeyer.

**Hayashi, O., Miyajima, T., Nomura, M., Ekawa, K., Nakano, H. Sanada, S. & Kitake, H.** (eds.) (1982). *Zusetsu Nihongo.* Japan: Kakusen shoten.

**Herdan, G.** (1960). *Type-Token Mathematics.* The Hague: Mouton.

**Kenbō, H., Kindaichi, H., Kindaichi, K. & Shibata, T.** (eds.) (1982). *Sanseido Kokugo Jiten.* Tokyo: Sanseido.

**Lewin, B.** (1975). *Abriss der japanischen Grammatik.* Wiesbaden: Harrassowitz.

**Maas, H.H.** (1972). Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift für Literaturwissenschaft und Linguistik 2/8, 73-96.*

**Raatz, U.** (1985). Better theory for better tests? *Language Testing 2, 60-75.*

**Raatz, U., Klein-Braley, C.** (1983). Ein neuer Ansatz zur Messung der Sprachleistung. Der C-Test: Theorie und Praxis. In: Horn, R., Ingenkamp, K. & Jäger, R.S. (Hrsg). *Tests und Trends 3. Jahrbuch der Pädagogischen Diagnostik.* Weinheim/Basel: Beltz.



**Raatz, U., Klein-Braley, C.** (1985). How to develop a C-test. In: Klein-Braley, C., & Raatz, U. (Eds.). *Fremdsprachen und Hochschule (FuH). AKS-Rundbrief 13/14, 20-22.* Bochum: AKS.

**Roos, U.** (1990). *Das C-Test-Prinzip und das japanische Schriftsystem. Eine empirische Untersuchung zur sprachspezifischen Anwendungsproblematik.* Diss., Ruhr-Universität Bochum.

**Roos, U.** (1992) Einige Probleme bei der Anwendung des C-Test-Prinzips auf das Japanische. In: Grotjahn, R. (Hrsg.). *Der C-Test. Theoretische Grundlagen und praktische Anwendungen.* Bochum: Brockmeyer (to appear).

**Siegel, S.** (1976). *Nichtparametrische statistische Methoden.* Frankfurt: Fachbuchhandlung für Psychologie.

**Taylor, W.C.** (1953). Cloze procedure: a new tool for measuring readability. *Journalism Quarterly 30, 415-433.*

**Tuldava, J.** (1975). Readability measurement. *Transactions of Tartu University 345, 102-119.*

## Appendix

**Text 1: Kindergarten**

After children have entered the infant school they will continue to meet situations in their home lives which will increase their store of mathematical knowledge; but this knowledge will still be unordered and vague. The first function of the teacher in the infant school is to make this knowledge more ordered and more precise and to help children to build up a vocabulary in which the words have a greater precision or meaning. Some children will need to have experiences which will give to them that background knowledge which others have gained from their home environment.

子供達は幼年学校に入学した後も、家庭において正確な知識のストックを増やし得るような場に直面する状況が続く。しかしながら、彼らの知識は未だ無秩序で取り留めがない。幼年学校の教師にとって最初の職務は、こうした知識をより秩序づけ、より正確にさせる

ことであり、また子共達の語彙を増やし、言葉がより正確な意味をもつように手伝うことである。子供達のうち幾人かは、ほかの子供達が家庭環境を通じて得るような知識的バックグラウンドを与え得るような経験を必要とするであろう。

**Text 2: Realism**

'Realism', a much abused word, has at least four current meanings, but when applied to novels it normally means a photographic imitation of everyday life. A 'realistic' novel is one in which the dialogue is colloquial and physical objects are described in such a way that you can visualise them. In this sense almost all modern novels are more 'realistic' than those of the past, because the describing of everyday scenes and the construction of natural-sounding dialogues are largely a matter of technical tricks which are passed on from one generation to another, gradually improving in the process.

大変濫用される言葉である、‘リアリズム’には少なくとも四つの基本的意味がある。しかし小説に関して用いられるときは、普通、日常生活の精密な模写を意味する。リアリズム小説と言えば、会話は口語体で読者が登場人物等を思い浮かべることが出来るものを言う。この意味では、現代小説の殆どすべてが過去の小説よりも‘リアリズム’的である。



何故なら、日常生活の措写と自然な調子の会話の構成と言うものは世代を重ね徐々に進歩する技術的な点にかかっているからである。

**Text 3: Mice**

Mice need not enjoy all that they destroy. They have to keep gnawing to file down the fast-growing incisor teeth that project beyond their lips. Otherwise, they would die of starvation, their teeth ingrown on jaws locked tight. Since electric wire is such a convenient jawful, mice amuse themselves by ringing doorbells and by fusing appliances. More seriously, many "mysterious" fires are thought to have been started by mice stripping insulation from wires. Aircraft managers are so alarmed at the damage mice might do on board that the sight of just one dropping can lead to an aircraft being taken out of service for fumigation.

鼠たちは、何も好んで破壊主義者である訳ではない。彼らは、唇の上に突き出て、急速に伸びる門歯を削るために、かじり続ける必要があるのである。さもなくば、歯が肉に食い込むか、あるいはあごが開かなくなり、餓死してしまうであろう。それで、丁度口にある電線をかじっては、戸口のチャイムを鳴らしたり、電気器具のヒューズを飛ばしたりするのである。更に悪いことには、不審火の多くは鼠が電線の絶縁体を食いちったことに端を発していると考えられている。飛行場管理者

が機上における鼠の害に気を使うことは並大抵ではなく、もし鼠の糞がひとつでも見つかろうものなら、飛行機の運行を停止し、消毒するほどである。

**Text 4: Fire department**
There are usually five men in the crew of a fire engine. One of them drives the engine. The Leader sits besides the driver, the other firemen sit inside the cab on the fire engine. The leader has usually been in the Fire Service for many years. He will know how to fight different sorts of fires. So, when the firemen arrive at a fire it is always the leader who decides how to fight the fire. He tells each fireman what to do.

いつもは五人が消防自動車にのっている。一人は消防自動車を運転します。指導者はドライバーのよこに座っています。他の消防士は車の中に座っています。長い間は指導者はいつも消防署にいた。いろいろな種類の火をどのように処置するかしっている。そのまま消防士は火事場に着くときに、どんなふうに火と戦うかいつも指導者は決定します。何をするか消防士に話す。

# On Classifying Texts with the Help of Cluster Analysis

*Heino Liiv and Juhan Tuldava, Tartu*

The paper deals with a comparatively new method of cluster analysis which has been modified to make it applicable to partially overlapping clusters. The method is illustrated by an analysis of twenty English texts.

## Principles of cluster analysis

Cluster analysis is a useful tool in computerized classification of objects[1]. The task by classification is to break up a corpus of objects into more or less homogeneous subgroups or clusters. The method is used to classify objects of many characteristics. Each object in cluster analysis is determined by a fixed set of parameters, and the classification is built up on the fixed parameters and their relationships.

There are several kinds of cluster analysis. According to the approach, methods of cluster analysis are applicable to fixed sets or fuzzy sets. By fixed sets a system of isolated clusters is formed (an example of the use of this method on linguistic material is to be found in Tuldava 1981). By fuzzy sets certain overlapping of clusters is allowed. Fuzzy sets of indistinct set outlines in which a set element is gradually fused into a subsystem (see Zadeh 1973) call for a method of cluster analysis admitting certain cluster overlapping.

We have applied cluster analysis to formal quantitative linguistic characteristic of natural languages. Cluster analysis of fuzzy sets seems to suit the purposes of linguistic analysis rather well, as most linguistic objects are of indistinct overlapping quality and can be treated as probabilistic systems. A computer machine programme produced in the Computation Center of Tartu University (Ääremaa 1980) for the so-called k-cluster formation (cf. Jardine & Sibson 1971) was used. The programme provided for the computation of overlapping clusters within the limits of $k - 1$ objects with the clause of $1 \leq k \leq n - 2$, where $n$ is the number of objects under study. The computer will automatically select the classification level ($h$) and the permitted number of overlapping elements in the cluster ($k$). The programme will produce a system of clusters characterizing the given set of objects somewhat more precisely than the

[1] The cluster analysis of the present study was conducted on computer EC-1060 with the use of special programmes produced by the Computation Centre of Tartu University.

programme for fixed sets would do.

**Initial data**

Twenty texts of modern literary English (each excerpt of about 2,000 words) were analysed. The data were excerpted from the book by H. Kučera and W.N. Francis (1967). The texts were of the following genres (see the full list of the books excerpted at the end of the paper):

Texts 1 - 10 Fiction
Texts 11 - 12 Press
Texts 13 - 14 Skills and hobbies
Texts 15 - 16 Learned and scientific writings
Texts 17 - 18 Popular Lore
Texts 19 - 20 Belles Lettres.

To avoid the impact of the subject matter discussed in the texts and to demonstrate as much as possible the inherent peculiarities of different styles, ten quantitative text characteristics widely used in stylometry served as the attributes in the analysis. They are the following:

1. Mean word (word-form) frequency: N/V with N being the number of running words ("tokens") in the text and V the number of various words ("types") in the vocabulary of the text.
2. Variation coefficient of the mean word frequency: $V = \sigma/(N/V)$ in which $\sigma$ is standard deviation.
3. Yule index K (see Yule 1944).
4. The ratio of words occurring only once in the whole vocabulary: $V_1/V$, in which $V_1$ is the absolute number of words occurring only once in the text.
5. The ratio of words occurring only once in the text: $V_1/N$.
6. The concentration of words of low frequency in the text: $V^*_5/N$, where $V^*_5$ is the number of words occurring 1-5 times in the text.
7. The concentration of words of low and medium frequency in the text: $V^*_{10}/N$, where $V^*_{10}$ is the number of words of 1-10 times of occurrence in the text.
8. The relative frequency of the most frequent word in the text: $p_1 = F_1/N$, where $F_1$ is the absolute number of occurrences of the most frequent word in the text.
9. The concentration of the most frequent words in the text: $p^*_{10} = F^*_{10}/N$, where $F^*_{10}$ is the frequency of occurrence of the ten most frequent words in the text.
10. Index of the stereotype level (see Mistrík 1967): $S = (N - V)/(V - V_1)$.



The initial data are presented in Table 1.

Table 1
Initial data: Distribution of values of characteristics of twenty objects (texts)

| Characteristic → / Object ↓ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.01 | 2.35 | 89.8 | 0.618 | 0.212 | 0.487 | 0.642 | 0.055 | 0.242 | 5.01 |
| 2 | 2.55 | 2.64 | 96.2 | 0.686 | 0.268 | 0.534 | 0.638 | 0.062 | 0.253 | 4.95 |
| 3 | 2.55 | 3.33 | 146.5 | 0.722 | 0.284 | 0.503 | 0.599 | 0.094 | 0.285 | 5.57 |
| 4 | 3.11 | 1.92 | 66.9 | 0.598 | 0.192 | 0.445 | 0.617 | 0.031 | 0.206 | 5.15 |
| 5 | 3.03 | 2.61 | 113.6 | 0.623 | 0.205 | 0.477 | 0.596 | 0.056 | 0.282 | 5.39 |
| 6 | 2.64 | 2.43 | 86.3 | 0.680 | 0.257 | 0.500 | 0.607 | 0.061 | 0.222 | 5.21 |
| 7 | 2.47 | 3.03 | 119.5 | 0.791 | 0.301 | 0.521 | 0.624 | 0.061 | 0.299 | 5.66 |
| 8 | 2.98 | 2.36 | 93.0 | 0.636 | 0.214 | 0.456 | 0.574 | 0.043 | 0.250 | 5.44 |
| 9 | 3.02 | 2.83 | 131.9 | 0.669 | 0.222 | 0.448 | 0.545 | 0.067 | 0.292 | 6.11 |
| 10 | 2.22 | 2.48 | 74.4 | 0.748 | 0.337 | 0.592 | 0.654 | 0.044 | 0.223 | 4.82 |
| 11 | 2.39 | 3.03 | 115.2 | 0.712 | 0.297 | 0.553 | 0.637 | 0.077 | 0.259 | 4.81 |
| 12 | 2.29 | 2.94 | 104.3 | 0.673 | 0.294 | 0.621 | 0.726 | 0.075 | 0.245 | 3.93 |
| 13 | 2.47 | 2.76 | 101.5 | 0.699 | 0.283 | 0.550 | 0.648 | 0.064 | 0.260 | 4.90 |
| 14 | 2.97 | 2.59 | 108.5 | 0.656 | 0.221 | 0.460 | 0.563 | 0.061 | 0.270 | 5.73 |
| 15 | 3.32 | 3.50 | 201.5 | 0.564 | 0.170 | 0.449 | 0.570 | 0.115 | 0.313 | 5.33 |
| 16 | 2.81 | 2.96 | 103.9 | 0.670 | 0.239 | 0.485 | 0.564 | 0.067 | 0.304 | 5.48 |
| 17 | 2.19 | 3.13 | 112.5 | 0.792 | 0.362 | 0.561 | 0.616 | 0.059 | 0.294 | 5.71 |
| 18 | 2.31 | 3.12 | 118.6 | 0.727 | 0.315 | 0.560 | 0.648 | 0.075 | 0.273 | 4.78 |
| 19 | 2.64 | 3.38 | 156.5 | 0.699 | 0.264 | 0.497 | 0.608 | 0.091 | 0.300 | 5.47 |
| 20 | 2.92 | 2.60 | 106.4 | 0.649 | 0.223 | 0.466 | 0.581 | 0.064 | 0.266 | 5.46 |
| Mean | 2.69 | 2.80 | 112.4 | 0.678 | 0.258 | 0.508 | 0.613 | 0.066 | 0.267 | 5.25 |
| SD | 0.33 | 0.40 | 30.1 | 0.055 | 0.051 | 0.051 | 0.042 | 0.019 | 0.030 | 0.47 |

**Affinity matrix**

Cluster analysis comprises two stages of investigation. The first stage is spent on affinity measurement (similarity or difference) of the objects on the basis of a given set of characteristics. The second stage is devoted to establishing a cluster system which would be able to group the objects under study at different levels of affinity.

The mathematical basis for classifying objects with the help of cluster analysis is the calculation of paired-object functions by their values. The first stage of investigation results in an affinity matrix between objects. In the present study the measure of affinity is the ordinary Euclidean distance applied to previously normalized values of the characteristics (by the traditional way of calculating the mean value and dividing it by the standard deviation). The Euclidean distance (d) is calculated according to the formula:

$$d(X_s X_t) = \left[\sum_{j=1}^{k} (x_{js} - x_{jt})^2\right]^{0.5}$$

**Table 2**

Matrix of difference of twenty texts (on the basis of Euclidean distances between the normalized values of characteristics)

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 2.37 | | | | | | | | | | | | | | | | | |
| 3 | 5.02 | 3.64 | | | | | | | | | | | | | | | | |
| 4 | 2.57 | 4.53 | 7.15 | | | | | | | | | | | | | | | |
| 5 | 2.21 | 3.07 | 4.08 | 3.86 | | | | | | | | | | | | | | |
| 6 | 2.06 | 1.70 | 4.26 | 3.46 | 3.01 | | | | | | | | | | | | | |
| 7 | 4.46 | 2.80 | 2.31 | 6.45 | 3.75 | 3.69 | | | | | | | | | | | | |
| 8 | 2.11 | 3.30 | 5.01 | 2.61 | 1.75 | 2.43 | 4.41 | | | | | | | | | | | |
| 9 | 4.35 | 4.47 | 3.52 | 5.56 | 2.45 | 4.19 | 3.73 | 3.10 | | | | | | | | | | |
| 10 | 4.63 | 2.87 | 5.59 | 5.92 | 5.69 | 3.43 | 4.26 | 5.37 | 6.80 | | | | | | | | | |
| 11 | 3.91 | 1.74 | 2.81 | 6.16 | 4.15 | 3.04 | 2.56 | 4.70 | 4.99 | 3.22 | | | | | | | | |
| 12 | 5.16 | 3.74 | 5.79 | 7.24 | 6.25 | 5.07 | 3.38 | 6.79 | 7.84 | 3.88 | 3.27 | | | | | | | |
| 13 | 3.04 | 0.75 | 3.45 | 5.24 | 3.57 | 2.36 | 2.54 | 3.97 | 4.81 | 2.69 | 1.20 | 3.30 | | | | | | |
| 14 | 2.90 | 3.35 | 3.79 | 4.10 | 1.41 | 2.81 | 3.60 | 1.63 | 1.60 | 5.73 | 4.31 | 6.90 | 3.87 | | | | | |
| 15 | 6.71 | 7.00 | 5.11 | 8.43 | 5.25 | 7.27 | 6.69 | 6.64 | 4.88 | 9.68 | 6.79 | 8.68 | 7.12 | 5.54 | | | | |
| 16 | 3.60 | 3.16 | 2.19 | 5.37 | 2.00 | 3.41 | 2.70 | 2.93 | 2.03 | 5.57 | 3.60 | 3.35 | 3.41 | 1.79 | 5.25 | | | |
| 17 | 5.95 | 3.89 | 3.42 | 7.75 | 5.53 | 4.74 | 1.95 | 5.89 | 5.33 | 4.02 | 3.21 | 5.70 | 3.44 | 5.22 | 8.28 | 4.33 | | |
| 18 | 4.52 | 2.23 | 2.93 | 6.77 | 4.64 | 3.71 | 2.45 | 5.27 | 5.38 | 3.34 | 0.80 | 3.24 | 1.63 | 4.83 | 7.10 | 3.94 | 2.84 | |
| 19 | 4.95 | 3.83 | 0.95 | 7.16 | 3.79 | 4.57 | 2.57 | 4.98 | 3.36 | 6.04 | 3.13 | 5.87 | 3.67 | 3.75 | 4.42 | 2.76 | 3.96 | 3.22 |
| 20 | 2.26 | 2.75 | 3.63 | 3.84 | 1.09 | 2.32 | 3.45 | 1.53 | 2.19 | 5.30 | 3.78 | 6.22 | 3.30 | 0.78 | 5.48 | 1.77 | 5.12 | 4.34 | 3.57 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |

where $x_{js}$ and $x_{jt}$ are the normalized values of characteristics; k is the number of measurements effected. The value $d(X_sX_t)$ for the vectors $X_s$ and $X_t$ is considered equivalent to the distance between the objects (texts) $T_s$ and $T_t$ by the chosen set of characteristics. It is supposed that the text affinity indicates closeness in the individual styles of the authors by revealing certain covert characteristics which are reflected in the integrated interdependent quantitative text characteristics.

Affinity measurements between objects are concisely expressed in the matrices of similarity or difference. The analysis based on Euclidean distances will result in the matrix of difference (see Table 2).

As seen from Table 2, the texts that come closest judged by the given set of characteristics are Text No. 2 (Fiction) and Text No. 13 (Skills and hobbies). The Euclidean distance (d) between these texts is 0.75. And, indeed, there is a good convergence of the values of the corresponding vectors of measurement (see table 1) as can be attested to by the following extract:



| | Text No. 2 | Text No. 13 |
|---|---|---|
| N/V | 2.55 | 2.47 |
| $V_1/V$ | 0.686 | 0.699 |
| $V_1/N$ | 0.268 | 0.283 |
| $p_1$ | 0.062 | 0.064 |
| $p^*_{10}$ | 0.253 | 0.260 |
| S | 4.95 | 4.90 |
| K | 96.2 | 101.5 |

On the other hand the most distanced are Text No. 10 (Fiction) and Text No. 15 (Scientific prose on micrometeorites). Their difference coefficient d = 9.68. The initial data of these texts are distributed as follows:

| | Text No. 10 | Text No. 15 |
|---|---|---|
| N/V | 2.22 | 3.32 |
| $V_1/V$ | 0.748 | 0.564 |
| $V_1/N$ | 0.337 | 0.170 |
| $p_1$ | 0.044 | 0.115 |
| $p^*_{10}$ | 0.223 | 0.313 |
| S | 4.82 | 5.33 |
| K | 74.4 | 201.5 |

Text No. 15 compared to Text No. 10 has a more varied (rich) vocabulary (low N/V value, high $V_1/N$ and $V_1/V$ values), a low concentration of the most frequent words ($p_1$ and $p^*_{10}$) and a lower stereotype level (S). By more detailed analysis these two texts appear as the opposing extremes on the periphery of the cluster system. They prove to be isolated (one-element) clusters in the text corpus under study.

## The building of a cluster system

One can proceed to build up a cluster system from clustering the data of the affinity matrix (similarity or difference) between the objects. With the help of a cluster-analysis algorithm it has been established that the most reliable clustering is achieved by partially overlapping clusters if there are no more than two overlapping elements in the cluster. The classification level was automatically set at h = 2.06; i.e. clusters were formed out of pairs of objects whose difference coefficient (d) was lower than 2.06. It is true that this level and the extent of permitted overlapping is most appropriate (optimal) only in the formal sense of description. The researcher is always free to decide if the outcome of the

analysis in its content matter meets the requirements set to the classification being established. It is possible to feed new initial conditions (level index h, and the number of k) into the computer and repeat the whole procedure. This will make the analysis a little more subjective, but there are cases when the approach is justified. The point is that on principle the quality of a classification may be assessed, first and foremost, by the level of interpretation, in particular the correspondence of the classification to the theoretical considerations of the investigation.

To see the algorithm at work, let us consider the clustering of paired objects (texts under study) in the increasing order of difference between the objects (see Table 3). At the level of h = 2.06, the total number of paired objects is twenty according to the intial matrix of difference.

The outcome of sequential pairing of texts at the given level of difference is a cluster system of eight multiple-element clusters:

1. (3, 19)
2. (7, 17)
3. (1, 6)
4. (2, 6)
5. (2, 11, 13)
6. (11, 13, 18)
7. (9, 14, 16)
8. (5, 8, 14, 16, 20).

Table 3
Clustering of paired objects

| No | Difference (d) | Objects (Nos of texts) | No | Difference (d) | Objects (Nos of texts) |
|---|---|---|---|---|---|
| 1 | 0.75 | (2, 13) | 11 | 1.63 | (13, 18) |
| 2 | 0.78 | (14, 20) | 12 | 1.70 | (2, 6) |
| 3 | 0.80 | (11, 18) | 13 | 1.74 | (2, 11) |
| 4 | 0.95 | (3, 19) | 14 | 1.75 | (5, 8) |
| 5 | 1.09 | (5, 20) | 15 | 1.77 | (16, 20) |
| 6 | 1.20 | (11, 13) | 16 | 1.79 | (14, 16) |
| 7 | 1.42 | (5, 14) | 17 | 1.95 | (7, 17) |
| 8 | 1.53 | (8, 20) | 18 | 2.00 | (5, 16) |
| 9 | 1.60 | (9, 14) | 19 | 2.03 | (9, 16) |
| 10 | 1.63 | (8, 14) | 20 | 2.06 | (1, 6) |

As can be seen there are several partially overlapping clusters. For example, Clusters Nos. 3 and 4 share Text 6 (or in other words Text 6 belongs to two clusters). Clusters Nos. 5 and 6 share two elements: Texts 11 and 13. There are some more clusters with two overlapping elements. Clusters No. 1 (Texts 3 and 19) and No. 2 (Texts 7 and 17) do not have any overlapping elements. Texts 4, 10, 12, 15 also stand isolated due to the peculiarities they manifest at the given level of difference. They may be said to form one-element clusters within the



given cluster system. The whole cluster system has been graphically represented in Figure 1.

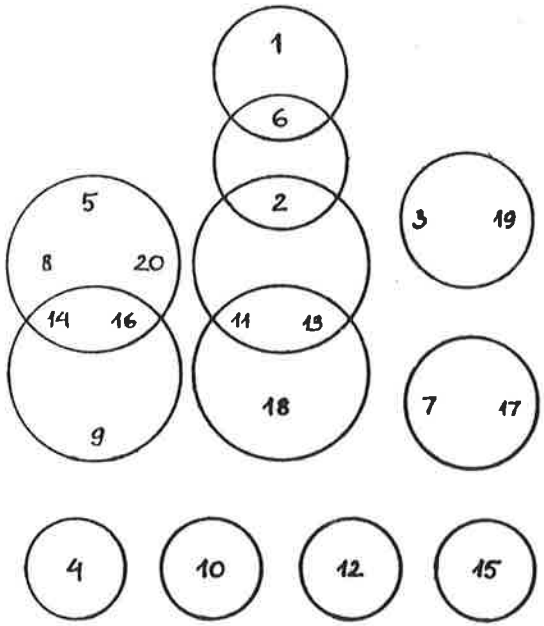



Figure 1. Cluster system of twenty texts

To illustrate cluster formation let us consider the process on the cluster including Texts 5, 8, 14, 16, and 20. The stages of the procedure are to be seen in Table 3. First Texts 14 and 20 are paired off as their quantitative linguistic characteristics are very much alike (see Table 1). Texts 5 and 8 can be joined to the two previous ones as all four texts are closely related (coefficient of difference d < 2.06). Text 16 can easily be paired off with Text 20, and it is also related to the other texts in the cluster except Text 8 (the coefficients of difference between Text 16 and Text 8 is equal to 2.93, i.e. there is a distorted relationship between Texts 8 and 16 which will to some extent weaken the uniformity of the cluster).

Text 9 proves to go well with texts 14 and 16, but it fails to meet the affinity requirements of the given level with the other texts in the cluster (i.e. Texts 5, 8 and 20). So a new cluster of Texts 9, 14, 16 has to be formed which shares two elements (14, 16) with the cluster made of Texts 5, 8, 14, 16, 20. These two clusters are considered to be partially overlapping. In the given example Text 9 (by J. Ford) is a piece of fiction but by some of its stylistic characteristics it comes close to the texts of skills and hobbies and scientific

prose (Texts 14 and 16). It is the same with two other pieces of fiction analyzed but they are similar to skills and hobbies and scientific writing according to some other of their characteristics.

## Conclusion

The above analysis proves that the cluster method yielding partially overlapping clusters is a more adequate and flexible way of classifying and analyzing natural-language texts than the traditional one, which operates with isolated clusters. The algorithm used has been adapted to the fuzziness of the outlines of the analyzed objects (texts of literary English). It can be concluded that by cluster analysis of texts the clusters should be treated as fuzzy sets.

The analysis has also proved that the interrelationships of texts belonging to different literary genres may be very complicated (texts of different genres fell into the same group by our experiment). It is most likely that the set of formal quantitative characteristics used in our example are to a large extent indicative of individual peculiarities of different authors' manner of writing. The clusters obtained and the interrelationships revealed reflect certain deep (covert) typological characteristics of texts. The method can be made use of at text attribution and typification and also for some other purposes. However, the analysis will benefit, if the formal quantitative methods are supplemented by some qualitative ones.

This paper is basically a description of a method based on modified cluster analysis. Further analysis will be useful in revealing intracluster and intercluster relationships. The aim of clustering, as of any other scientific classification, is to reveal the nature and covert regularities within a set of objects under study.

## List of the Texts Excerpted

Fiction

1. Francis Pollini, *Night.* Boston, Houghton Mifflin 1961: P. 246-252
2. Leon Uris, *Mila 8.* New York, Doubleday 1961: P. 324-329.
3. E. Luas Myers, *The Vindication of Dr. Nestor.* Sewance Review 69:2, 1961, 290-295.
4. Dell Shannon, *The Ace of Spades.* New York, Morrow 1961: P 184-191.
5. George Harmon Coxe, *Error of Judgement.* New York, Knopf 1961: P. 24-29.
6. Jim Harmon, *The Planet with No Nightmare.* "IF" Magazine July 11:3, 1961, 7-12.



7. Mary Savage, *Just for Tonight.* New York, Dodd & Mead 1961: P. 114-120.
8. Peter Bains, *With Women...Education Pays Off.* "Monsieur" Magazine Feb. 4:2, 1961, 77-78.
9. Jesse Hill Ford, *Mountains of Gilead.* Boston, Little & Brown 1961: P. 128-133.
10. Robert Carson, *My Hero.* New York, McGraw-Hill 1961: P. 170-174.

Press

11. Press. Editorial: *The New York Times Oct. 17, 1961, p.38.*
12. Press reportage: *The Sun, Baltimore March 18, 1961, p. 1,18; Dec. 10, sec. E, 1961, p.8.*

Skills and Hobbies

13. Joseph E. Choate, *The American Boating Scene.* Rudder Magazine Jan. 77:1, 1961, p. 31-32, 35, 82, 84.
14. Ethel Norling, *Renting a Car in Europe.* Playbill Magazine March 13, 5:11, 1961, p. 5-11.

Learned and Scientific Writings

15. J.F. Vedder, *Micrometeorites. Satellite Environment Handbook.* Stanford, Stanford University Press 1961, p. 92-97.
16. J.H. Hexter, *Thomas Moore: On the Margins of Modernity.* Journal of British Studies Nov. 1961, 1:1, p. 28-32.

Popular Lore

17. Kenneth Allsop, *The Bootleggers and Their Era.* Garden City, New York, Doubleday 1961, p. 68-72.
18. Frederic A. Birmingham, *The Ivy League Today.* New York, Crowell 1961, p. 142-149.

Belles Lettres

19. Arthur S. Miller, *Toward a Concept of National Responsibility.* The Yale Review Dec. 1961, LI:2, p. 186-191.
20. Tom F. Driver, *Beckett by the Madelaine.* Columbia University Forum Summer 1961, 4:3, p. 21-24.

**References**

**Ääremaa, R.** (1980). Klassificirovanie naučnych dannych s polučeniem častično pokryvaemych klassov (Scientific data classification yielding partially overlapping groups). *Papers of the Computational Center of Tartu University 44, 55-73.*

**Jardine, N., Sibson, R.** (1971). *Mathematical Taxonomy*. London, Wiley.

**Kučera, H., Francis, W.N.** (1967). *Computational Analysis of Present-Day American English.* Providence (R.I.), Brown University Press.

**Mistrík, J.** (1967). Matematiko-statističeskie metody v stilistike (Mathematical-statistical methods in stylistics). *Voprosy jazykoznanija No. 3, 42-52.*

**Tuldava, J.** (1981). Opyt klassifikacii tekstov s pomoščju klaster-analiza (Experiment in text classification with the help of cluster analysis). *Acta et Commentationes Universitatis Tartuensis 591, 136-157.*

**Yule, G.U.** (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

**Zadeh, L.A.** (1973). Outline of a new Approach to ther Analysis of Complex Systems and Decision Processes. *IEEE Transactions on Systems, Man and Cybernetics SMC-3, 28-44.*

# Parts of the Sentence:
## Evidence of their Communicative Significance in Text Structure

*Ludmila Uhlířová, Prague*

Every text which is coherent is highly organized. The so-called "development of communication", to use J. Firbas' term, or "flow of information", to use M.A. K. Halliday's term, is not - as has been repeatedly stressed by many authors - an unstructured flow. It is "perspectived", "oriented", "hierarchized", with a periodicity of "peaks of prominence" and "troughs of non-prominence", with elements with a "higher degree of communicative dynamism" and a "lower degree of communicative dynamism", with "foundation-laying" and "core-constituting" elements, with points of departure and focuses, themes and rhemes, "given" and "new", etc. (see Berry 1989; Firbas 1992; Givón 1979; Halliday 1985; Hawkins & Siewierska 1991; Mathiessen 1992 and others). Generally speaking, communicatively (or pragmatically) prominent are those semantic elements in a text (in the "semantic space" of a text) which are either bearers of the text continuity, or, on the contrary, which signal the text discontinuity, or change, alteration.

The principles of text organization are language independent in their most general essence, but their specific implementation is always subject to various limitations imposed on it by the grammatical and semantic structures in the system of a particular language. Thus, e.g., at a general typological level it is possible to classify languages either as "topic prominent", or as "subject prominent" (Li 1976). Any more detailed contrastive inquiry into two or more languages, be they typologically akin or remote in this respect, will map the specific interplay of the particular means of expressing textual coherence in any one of them (see, e.g., for ten Slavic languages, Běličová & Uhlířová 1993).

In the present study it is hypothesized that there exists a significant relationship between the communicative prominence of a semantic element (semantic component, or semantic unit, i.e., a meaningful item of any kind) in a sentence (utterance) and its structural position in the sentence pattern, expressed in terms of parts of the sentence. Some of the structural positions are supposed to be preferred as points of prominence over the others. This hypothesis is based on the opinion, widely accepted by functional linguists, that the syntactic structure of a sentence is not independent of text strategy; on the contrary, it is subject to it and mirrored by it. The textuality of a particular sentence is (a) created by its function in the text strategy and therefore, from another point of view, is (b) determined by the function in the strategy of the text that it has to fulfil.

To use Enkvist's wording, "the discourse is the father of the sentence, not vice versa" (Enkvist 1992: 16). To be more concrete: It is well-known that, e. g., in many languages, such as in English, in all Slavic languages, as well as in many other Indo-European and non-Indo-European languages the grammatical subject serves as the "first candidate" to function as the theme, in other words, to function as the "unmarked theme"[1], in contrast to verbal "specifications" which are natural candidates to serve as rhemes, unless the factor of immediate context acts counter it (see J. Firbas' scales of communicative dynamism; Firbas 1992). Clauses with certain syntactic structures seem to be predetermined to be used in a certain way in the act of communication, and vice versa; perhaps the most common examples treated in linguistic literature from this point of view are various presentative, existential and locative sentences (not only those with the copula).

If this hypothesis about the relationship between syntactic positions (labelled as sentence parts) and their different communicative prominence (or, non-prominence) in text is plausible, it is probable that it should be manifested somehow at the level of linguistic expression. In Czech, which is the subject of interest in this paper, there are various lexical, grammatical and intonation means of expression of prominence. In this contribution, only one class of means of expression is dealt with, namely the most common definite (demonstrative) and indefinite pronouns, which function as determiners in nominal groups. They represent a relatively small and non-productive, closed class of elements, which, nevertheless, are very frequent in texts.

It is not claimed here that any occurrence of a determiner in a text signals a prominence of the lexical item (nominal group) with which it is associated. Prominence is a communicative (pragmatic) function and determiners are lexical means which may take on different functions in utterance, that of prominence being one of them. No one-to-one correspondence between the two should be presupposed or expected. Prominence and determinedness are in principle two different functions.[2] The thematic "point of reference" cannot be defined as a nominal group with definite status, and similarly, the rhematic "point of reference" cannot be defined as a nominal group with indefinite status. Neither is it possible to ascribe a definite status to a nominal group only on the basis of the fact that it functions as the theme in the sentence, and, on the contrary, any nominal group cannot be considered indefinite only because it is the rheme of the sentence. Generally, it may be only stated that in Czech in unmarked cases - which are the most frequent ones in texts - theme is, as a rule, definite,

[1] See e.g., Halliday (1985: 869) for English: "The Subject is the element that is chosen as Theme unless there is a good reason for choosing something else".

[2] In Czech, nouns are not determined obligatorily, neither definite nor indefinite articles exist, and the category of determinedness/indeterminedness is not a grammatical category.



whereas rhemes display an evident tendency to be indefinite. (Even in Czech, however, it is easy to give examples of utterances with the opposite distribution of these features.)

The category of determinedness (and indeterminedness) is overtly expressed lexically in Czech only if there arises a special semantic need to do it. The definiteness or indefiniteness of a nominal group must be expressed overtly if it is necessary to characterize the point of reference uniquely, i.e., to point out that the point of reference is just this and not that NP, in other words, if the feature of definiteness or indefiniteness is the bearer of a referential distinction. E.g., a thematic NP regularly has a marker of explicit indefiniteness if it represents a new and previously not mentioned theme, and, consequently, if it is not co-referential with another, lexically identical NP in the previous sentence. Similarly, an overt definiteness marker will very probably stand before a rhematic NP, if it should be interpreted as co-referential with another, lexically identical NP in the immediately preceding text. In all other cases the presence or absence of a lexical determiner is more or less optional, the choice depending on the interplay of a number of contextual, lexical, syntactic, stylistic, rhythmical-intonational and other factors. The specific choices manifest themselves as more or less evident tendencies rather than strict rules. To give several examples: There is a clear tendency not to use an anaphoric determiner in cases when the NP is - at least in the given context - semantically unique (specific), or, e.g., if the person's naming is constant in the given text; in such cases the determiner seems usually redundant. On the contrary, the determiner is quite regularly present, if the head of the anaphoric NP is a noun of a categorial, very broad meaning, typically in the thematic section of the utterance. Also, the presence of a determiner may often seem desirable, if a subsequent NP differs from its antecedent NP by the presence or absence of a qualitative (evaluative) attribute; it is so because an attribute is capable of modifying the meaning of the NP, and the determiner contributes to the unambiguousness of the co-referential relation. The same reasons call for the presence of a determiner if the subsequent NP is metaphorical, or expressive, and as such it has a (sometimes very strong) rhematizing effect which would possibly make the coreferentiality of the NP fuzzy. Further, a determiner is usually present, if it has - in addition to its determinative function - a highlighting, topicalizing effect (in the sense of 'namely', 'just', 'only', 'it is the NP that...', etc.), which may influence the understanding of the respective text tie. Also, if the antecedent is a VP, especially a VP denoting a concrete activity or a concrete state, or if it is a VP with copula and predicative complement, the subsequent nominalization is often supplied by a determiner.

And last but not least, there is the influence of syntactic factors on the presence or absence of the determiner, which is the topic of the present study. Taking into account all the reservations made above concerning the relationship between the prominence and determinedness, it is now possible to formulate the hypothesis above more precisely and to present statistical data which will either

confirm or deny it. The question to be answered is: In the Czech sentence do syntactic positions (expressed in terms of sentence parts) exist which, providing the sentence is used in a text, become preferred bearers of explicitly lexically marked determinedness or indeterminedness, and which as such are supposed to be either probable points of thematic continuity, or probable points of thematic alteration in text?

Statistics offers one possible way to answer the question. It may be expected from what is known about the frequencies of occurrence of word classes in Czech (Těšitelová & others 1985) that there will be significant frequency differences in usage of determiners between various styles, genres, or registers. Nevertheless, the explanation of the basic correlations between the syntactic and communicative (pragmatic) structures are expected to outweigh the differences at the level of style and to reflect more underlying pragmatic principles of text coherence.

The statistical study presented in this article has been performed on a Czech corpus of 540 000 orthographic words consisting of 180 samples (3 000 words each) of scientific expository prose, newspapers (both narrative report and analytical editorial comment) and administrative texts. The corpus, representing the written and spoken Czech Literary Standard of the seventies, was compiled in the Institute of the Czech Language, Czechoslovak Academy of Sciences, Prague.[3] Recently it has been transferred to diskettes and the data relevant for the present study has been counted on the IBM PC. The data is arranged in Table 1 below, together with commentary.

Before presentation of the data it may prove useful to know which of the Czech determiners are most frequent in texts. Let us briefly mention that the statistical counting has confirmed, as expected, a high prevalence of definite determiners over indefinite ones in texts. The total number of the definite determiners *ten, tento* and *onen* 'this/that' together come to 13168, whereas the total number of the indefinite determiners *jeden* 'one', *některý* 'some/any/a', *nějaký* 'some/any/a', jakýsi 'some/any/a' and *jistý* 'certain' amount to only 2477 occurrences in the corpus. This prevalence has natural pragmatic reasons: If a text is coherent, then the primary means of expressing its thematic continuity can reasonably be expected to be more frequent than those expressing the introduction of a "new" element of content.

All the above listed determiners, both definite and indefinite, in *dependent* (= attributive) positions in NPs, and only those, have been statistically processed. The frequencies of occurrences of all NPs containing at least one of the determiners have been counted, separately for each syntactic position (i.e. separately for each nominal sentence part). The statistical distribution over the sentence parts with determiners has then been compared with the **total** frequency distribution of the nominal sentence parts in the whole corpus. The hypothesis

[3] A detailed description of the corpus is given in Těšitelová et al. (1985).



is proved valid if the differences between the total distribution of the nominal sentence parts and the distributions of the NPs with a determiner are significant.[4]

Table 1 below is arranged in the following way. Each row begins with the *total* absolute and relative frequencies of a particular sentence part in the corpus as a whole; it is followed by the absolute and relative frequencies of the same sentence part in cases where the sentence part is the headword of one of the most common Czech determiners, as listed above. The values in the column 'total frequency of NP' are compared with those in the corresponding columns. First, the comparison shows that *ten*NP occupies the position of grammatical subject in 25.6 % of all occurrences, the position of grammatical object in 23.7 %, the position of attribute in 19.2 %, etc. (as can be seen from the fourth column) and that this frequency distribution differs significantly from the total distribution of the corresponding parts of the sentence in the corpus: It is only in 17.1 % thatan NP occupies the position of subject, in 17.6 % the position of object, and in 39.0 % the position of attribute (see the second column). The differences between the two distributions are clearest in the *subject* position: This is the syntactic position that shows the highest preference of the explicit determiner of definiteness *ten*. Similar differences, too, are found with *tento*: In 27.0 % the *tento*NP occupies the subject position, in 20.9 % it occupies the object position, and only in 25.9 % does it occupy the position of an attribute (see the sixth column for the details). These values, too, differ significantly from the corresponding values of the total distribution of syntactic functions of NPs in the corpus (see below).

Quite another picture of prominent positions shows when NPs with indefinite determiners are considered. The comparison of the respective values shows that with *jeden, některý, nějaký* and *určitý* it is the *object* position that is highly prominent. *Jeden*NP occupies the object position in 31.4 % of all occurrences, *některý*NP occupies the same position in 28.6 %, *určitý*NP in 30.5 % and the

[4] It should be noted that the sizes of the subparts of the corpus representing different styles differ from each other. The size, given in number of words, is the following:

| | |
|---|---|
| written newspapers | 52000 of word forms |
| academic writing | 204000 of word forms |
| administrative texts | 45000 of word forms |
| spoken journalistics | 24000 of word forms |
| spoken academic genres | 96000 of word forms |
| spoken administration | 15000 of word forms. |

Due to the difference in size, the level of precision of the counted relative frequencies is not the same in all cases. Nevertheless, the sizes of all the subparts of the corpus are large enough so that the interpretative power of the data is not affected in any way.

Table 1
Frequency distribution of sentence parts expressed by an NP.
Total distribution as compared with the NPs with determiners
*ten, tento, takový, jeden, některý, nějaký, určitý.*

| Syntactic position | NPsubst. | | Frequency of NP with a determiner → | | | |
|---|---|---|---|---|---|---|
| | Total frequency | | Definite | | | |
| | | | *tenNP* | | *tentoNP* | |
| | abs. | rel. | abs. | rel. | abs. | rel. |
| subject | 29963 | 17.1 | 678 | 25.6 | 1123 | 27.0 |
| object | 30950 | 17.6 | 628 | 23.7 | 869 | 20.9 |
| attribute | 68472 | 39.0 | 507 | 19.2 | 1081 | 25.9 |
| Adv place | 12585 | 7.2 | 244 | 19.2 | 370 | 8.9 |
| Adv time | 5693 | 3.2 | 130 | 4.9 | 196 | 4.7 |
| Adv modus | 9257 | 5.3 | 129 | 4.9 | 321 | 7.7 |
| Adv caus. | 2606 | 1.5 | 66 | 2.5 | 100 | 2.4 |
| other | 16058 | 9.1 | 264 | 10.0 | 106 | 2.5 |
| total | 175604 | 100.0 | 2646 | 100.0 | 4166 | 100.0 |

| → Frequency of NP with a determiner | | | | | | | |
|---|---|---|---|---|---|---|---|
| Indefinite | | | | | | | |
| *jedenNP* | | *některýNP* | | *nějakýNP* | | *určitýNP* | |
| abs. | rel. | abs. | rel. | abs. | rel. | abs. | rel. |
| 38 | 20.2 | 171 | 26.3 | 65 | 18.3 | 68 | 11.7 |
| 59 | 31.4 | 186 | 28.6 | 126 | 35.5 | 177 | 30.5 |
| 34 | 18.1 | 134 | 20.6 | 52 | 14.6 | 137 | 23.6 |
| 25 | 13.3 | 94 | 14.5 | 24 | 6.8 | 36 | 6.2 |
| 12 | 6.4 | 14 | 2.2 | 6 | 1.7 | 33 | 5.7 |
| 6 | 3.2 | 15 | 2.3 | 43 | 12.1 | 65 | 11.2 |
| 2 | 1.1 | 10 | 1.5 | 4 | 1.1 | 13 | 2.2 |
| 12 | 6.3 | 26 | 4.0 | 35 | 9.9 | 52 | 8.9 |
| 188 | 100.0 | 650 | 100.0 | 355 | 100.0 | 581 | 100.0 |



highest preference in the object position is shown by *nějaký*, the frequency of occurrence amounting to 35.5 %. The object position proves to be the prominent position to express new, unknown, previously not mentioned and at the same time indefinite pieces of information. In addition, some of the adverbial positions (when expressed by an NP) show similar prominence as well; e.g., *jeden*NP and *nějaký*NP show relatively high frequencies in the position of place adverbial, amounting to 13.3% and 14.5% respectively. The subject position shows as prominent only with those indefinite determiners which are typically used referentially (not, e.g., predicatively), namely with *jeden*NP (20.2 %), and *některý*NP (26.3 %).

In order to test the hypotheses that the conditional frequencies of lexical determiners are different from their overall frequencies two procedures can be used:

(1) It can be tested whether the distribution of conditional frequencies is homogeneous with that of total frequencies. The familiar chi-square or 2I-test would be adequate here. It can easily be shown that each of the six distributions is significantly different from the overall distribution.

(2) However, we are more interested in the conditioning strength of the individual positions/functions (subject, object, etc.) exerted on the particular determiner. To this end we can test the deviation of the relative frequency of a determiner in a given position/function from its expected relative frequency represented by that of the total count, using asymptotically the normal distribution according to the formula

$$\frac{p - P}{\sqrt{\frac{PQ}{N}}} = u$$

where P is the expected relative frequency, Q = 1 - P and N = sample size. In order to preserve the direction of deviation we use the one-sided test. Thus, for example, for *tenNP* in the subject position we obtain from Table 1

$$\frac{0.2560 - 0.1710}{\sqrt{0.1710\ (0.8290)\ /\ 2646}} = 11.61$$

This value is highly significant ($P < 10^{-10}$) and signals that *tenNP* is preferably used as subject. The results of testing are shown in Table 2. Taking $\alpha = 0.001$ all values of $|u| > 3.09$ are considered significant.

Thus the hypothesis that the usage or non-usage of lexical determiners accompanying NPs in **different** syntactic positions is communicatively conditioned is very strongly supported by the statistical data.

Table 2
Test for the preference of Czech determiners in syntactic positions

| Position | ten | tento | jeden | některý | nějaký | určitý |
|---|---|---|---|---|---|---|
| subject | 11.61 | 16.97 | 1.13 | 6.23 | 0.60 | -3.46 |
| object | 8.24 | 5.59 | 4.97 | 7.36 | 8.86 | 8.17 |
| attribute | -20.88 | -17.34 | -5.88 | -9.62 | -9.43 | -7.61 |
| Adv place | 3.86 | 4.11 | 3.14 | 6.98 | -0.28 | -0.90 |
| Adv time | 4.97 | 5.50 | 2.49 | -1.45 | -1.61 | 3.42 |
| Adv mod | -0.92 | 6.91 | -1.29 | -3.41 | 5.72 | 6.35 |
| Adv caus | 4.23 | 4.78 | -0.45 | 0.00 | -0.62 | 1.39 |
| other | 1.61 | -14.81 | -1.33 | -4.52 | 0.52 | -0.17 |

## References


**Běličová, H., Uhlířová, L.** (1993). Slovanská věta (Slavic sentence). Prague, Euroslavica (in press).

**Berry, M.** (1989). *An introduction to systemic linguistics I, 2nd ed.* Nottingham, University of Nottingham.

**Davies, M., Ravelli, L.** (eds.) (1992). *Advances in systemic linguistics. Recent theory and practice.* London and New York, Pinter.

**Enkvist, N.E.** (1991). Discourse strategies and discourse types. In: Ventola (ed.), 3-22.

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

**Givón, T.** (1979). *On understanding grammar.* New York, Academic Press.

**Halliday, M.A.K.** (1985). *An introduction to functional grammar.* London, Arnold.

**Hawkins, J.A., Siewierska, A.** (eds.) (1991). *Performance principles of word order.* Eurotyp working papers 2. Amsterdam, European Science Foundation.

**Li C.N.** (ed.) (1976). *Subject and Topic.* New York, Academic Press.

**Matthiessen, Ch.** (1992). Interpreting the textual metafunction. In: Davies, Ravelli (eds.): 37-81.

**Těšitelová, M.** and others (1985). *Kvantitativní charakteristiky současné češtiny* (Quantitative characteristics of Present-day Czech). Prague, Academia.

**Ventola, E.** (ed.) (1991). *Functional and systemic linguistics.* Berlin-New York, Mouton-de Gruyter.


# On Quantitative Analysis of Dialogue and Monologue

*Marie Těšitelová, Prague*

## Introduction

Since the beginning of this century, linguists have been interested in the problems posed by dialogue and monologue (cf. Mukařovský 1948). Mukařovský's work in this area remains inspiring to this day. Much can be gained in applying his ideas to quantitative linguistics. Initially, stylistics was typically concerned with dialogue as contrasted with monologue (cf. Bečka 1970, 1992; Mistrík 1979, 1985; etc.). Since the advent of communication theory (cf. Müllerová 1979; Hoffmannová 1989; etc.), dialogue and monologue have occupied an important place in the field of linguistics.

The aim of this article is not to resolve the problem of basic concepts and terms regarding dialogue and monologue (cf. Hoffmannová & Müllerová 1992). This article focuses rather on quantitative analysis of language phenomena and endeavours to enhance our knowledge of dialogue and monologue from a quantitative point of view. We shall allow for the classical distinction of dialogical concepts, where the drama occupies a prominent place (cf. Mukařovský 1948; Bečka 1970); the dramatic dialogue - as is well known - has an artistic character and a written form. The dramatic dialogue is marked by the meaningful features of dialogue, i.e., the situation with regard to the speaker and the penetration of some dialogical replies, of some contexts, by the situation with regard to the problems of sequential structures, of stereotyped expressions, etc. Even though a precise boundary line between dialogue and monologue is a difficult assignment, we are comparing their quantitative characteristics in an attempt to determine their distinctive and coincident features.

The choice of the dramatic dialogue and the poetry for representation of dialogue and monologue depends on the application of quantitative methods on the basis of material contained in the frequency dictionary of Czech (Jelínek, Bečka & Těšitelová 1961, further FDC); FDC includes some quantitative characteristics of dramatic dialogue (in style group D) and monologue (in style group B). For the purpose of this article, it was necessary to elaborate, reclassify and reassess this material from the point of view of contemporary linguistics. To simplify the terminological problem I am using the term *dialogue* in the sense of dramatic dialogue and *monologue* in the sense of poetic monologue.

To corroborate the results of the FDC based on the material from the years 1942-1948, I analyzed the following texts of present-day Czech: V. Havel's *Asanace* (Slum clearance, Prague 1990) and the poetic text of J. Žáček *Text-appeal* (Prague 1990). The results of the quantitative analysis of this text were to a certain extent specific, so that it was necessary to verify them on the basis of another text (control text of the same size chosen by cluster sampling). This third text is *Rodné číslo Homéra* (Homer's number of birth, Prague 1986) by K. Sýs.

As to quantitative characteristics of the pair dialogue - monologue, references are not numerous (see above). For example C. Muller (1962) studied the frequency of pronouns in French dramatic dialogue, J. Mistrík (1971) analyzed the vocabulary of dialogue in Slovak. The work by M. Těšitelová et al. (1983, 1983a) brings the quantitative characteristics of spoken scientific (non-fiction) Czech, for details of which see the monograph by M. Těšitelová (1992).

## Description of Materials

This article is based on two statistical populations of FDC:

*drama D* (10 texts published in Prague in the years 1942-1948)
corpus size $N_D$ = 140,225 words,
vocabulary $V_D$ = 11,034 different words ;

*poetry B* (10 poetic texts published in Prague in the years 1942-1947)
corpus size $N_B$ = 61,150 words,
vocabulary $V_B$ = 10,461 different words.

The differences in the size of N - in comparison with FDC - are not significant and can be handled by some necessary technical modifications of materials. In contrast to this fact the differences in the size of vocabularies V are of great importance, cf.:

$V_D$ in FDC = 24,323 different words,
$V_B$ in FDC = 20,122 different words;

in both cases the size of vocabulary in modified populations is lower by nearly 50% (in dialogue about 2.2 times, in monologue twice). These differences are caused by the fact that in contrast to the FDC, the vocabulary V in both populations is not based on the cumulative number of V in the different texts making up the populations D and B, but on the first occurrence of a different word (accompanied by the cumulative frequency in both populations). These elementary characteristics of the vocabulary in both populations D and B show



one of the basic differences between the vocabularies of dialogue and monologue.

The chi-square test used as a test of homogeneity to examine the distribution of different words in $V_D$ and $V_B$ rejected at the 0.05 level the null hypothesis that the differences between the vocabularies of dialogue and monologue are not significant, cf. $X^2 = 87.2668 > \chi^2_{0.05;8} = 27.9$.

These results were verified by the quantitative analysis of dialogue and monologue samples from present-day language, cf.:

*drama* by V. Havel:
size $N_H$ = 9,550 words, i.e.
vocabulary $V_H$ = 2,134 different words;

*two poetic texts:*

by J. Žáček
size $N_Ž$ = 4,324 words, i.e.
vocabulary $V_Ž$ = 1,562 different words;
by K. Sýs
size $N_S$ = 4,324 words, i.e.
vocabulary $V_S$ = 2,053 different words.

The vocabularies of the two texts whose authors are contemporaries with similar treatment of poetic problems are "richer" than the materials quoted above; the differences in the results of the analysis $V_Ž$ and $V_S$ between the two texts are significant ($X^2 = 40.914 > \chi^2_{0.05;8} = 27.9$); therefore we are in need of at least two texts.

Detailed analysis of both vocabularies $V_D$ and $V_B$ and those of control texts $V_H$, $V_Ž$ and $V_S$ shows that in dialogue the words with frequency 6 and higher are more frequent (20.30%) than that in monologue (13.82%). The number of different words in the dialogue vocabulary is greater than that in monologue. Hence it follows that in the dialogue the number of words with the frequency 5-1 is lower (79.70%) than in the monologue (86.18%). This fact is caused by the number of words with the frequency 1 which is higher in the monologue vocabulary ($V_B$ = 55.31%) than in that of dialogue ($V_D$ = 46.76%). It means that a new word is used more often in the monologue; the word is repeated more in the dialogue (cf. Těšitelová 1968). It depends on the grammatical or semantic character of words; see the value of $\chi^2$ calculated for the vocabularies $V_D$ and $V_B$ above.

On the basis of the quantitative analysis of the material contained in the FDC and of the above-mentioned three control texts, we shall devote attention to the language of dialogue and monologue from the point of view of the parts of speech in the vocabulary, as well as of the parts of speech in the text (size).

**The Parts of Speech in the Vocabulary**

In accordance with FDC and with some of my works (cf. Těšitelová 1974), we shall examine the distribution of the parts of speech divided in three groups:

(1) *nominal* (nouns, adjectives, prepositions),
(2) *verbal* (verbs, pronouns, adverbs, conjunctions) and
(3) *neutral* (numerals, interjections, particles).

We obtained significant differences in results among the three groups mentioned ($X^2 = 45.5366 > \chi^2_{0.05;2} = 15.2$):

- in the dialogue vocabulary the verbal group prevails with 36.49% (in that of monologue with 34.86%);

- in the monologue vocabulary the nominal group prevails with 64.13% (in that of dialogue with 61.49%);

- in dialogue, nouns occupy the dominant position within the nominal group at 70.26% (followed by adjectives with 29.20%);

- in monologue we find adjectives in first place with 32.32% (followed by nouns with 67.16% and by prepositions with 0.52%); the differences are confirmed by the chi-square test, $X^2 = 15.4606 > \chi^2_{0.05;2} = 15.2$.

These relations are caused above all by the fact that the number of nouns and adjectives with frequency 1 is higher in monologue than in dialogue, cf. above, the value of $\chi^2$ calculated for the distribution of different words in $V_D$ and $V_B$:

| Parts of speech | Vocabulary of dialogue ($V_D$ in %) | Vocabulary of monologue ($V_B$ in %) |
|---|---|---|
| Nouns | 46.30 | 52.72 |
| Adjectives | 56.94 | 63.84 |

Within the framework of the verbal group in dialogue, adverbs occur with 20.20% (followed by verbs with 76.65%, by pronouns with 1.81% and by conjunctions with 1.34%).

In monologue we find verbs with 78.86% (followed by adverbs with 17.88%, pronouns with 1.86% and by conjunctions with 1.40%). These differences are not significant ($X^2 = 6.6504 < \chi^2_{0.05;3} = 17.7$).

The number of different verbs in the vocabulary of dialogue and monologue depends on the higher number of lexemes with the frequency 1 (in monologue 56.12%, in dialogue 44.40%). The same tendency holds for the vocabulary of adverbs with the frequency 1 (in monologue 52.30%, in dialogue 41.94%). This means that adverbs and verbs are more often repeated in dialogue than in monologue where the choice of a new verb or adverb prevails.



The size of the vocabulary of parts of speech with the limited repertory of lexemes - in the verbal group the pronouns and conjunctions - is in mutual agreement (cf. the differences between the observed and expected frequencies regarding the chi-square value calculated for the distribution of different words in $V_D$ and $V_B$).

In dialogue, the most frequent parts of speech in the neutral group are interjections (59.19%), in second place numerals (40.81%); in monologue, we find numerals (68.87%) followed by interjections (31.13%); these differences are significant ($X^2 = 22.6296 > \chi^2_{0.05;1} = 12.1$).

The prevailing position of the interjections in the vocabulary of dialogue depends on their open number in the system and on their stylistic role in text; but in the vocabulary of monologue we should expect a more meaningful role for this part of speech. The relatively high number of numerals in the monologue vocabulary can be interpreted by the dependency on the selected themes of texts in FDC.

The quantitative characteristics $V_H$ of the vocabulary of V. Havel in the drama - compared with M.V. Kratochvíl's play *České jaro* (Czech spring, Prague 1948) of approximately the same size and vocabulary ($N_K$ = 12,386 words, $V_K$ = 2,156 different words) have corroborated the following features of dialogue:

(1) The verbal group occupies a special position in dialogue; the differences among the respective parts of speech are not significant (Havel and Kratochvíl, $X^2 = 0.6018 < \chi^2_{0.05;2} = 15.2$).

(2) The number of adjectives in the nominal group and of adverbs in the verbal group depend on the individual use of these words by the authors, even if the higher number of adverbs can be conceived as a significant feature of the dialogue vocabulary ($X^2 = 20.8850 > \chi^2_{0.05;2} = 15.2$).

The quantitative characteristics of the vocabulary in the two present-day poetic texts confirm the individual marked features of monologue vocabularies:

(1) Within the framework of the nominal group, the number of nouns and adjectives is lower than in the dialogue vocabulary, namely due to the choice of new words or to the repetition of words (Žáček and Sýs, $X^2 = 21.2112 > \chi^2_{0.05;2} = 15.2$; the differences are significant).

(2) Within the framework of the verbal group, the number of verbs is higher than that of adverbs. The deviations in the number of pronouns and conjunctions indicate the individual features of the style; the differences are not significant (Žáček and Sýs, $X^2 = 0.7653 < \chi^2_{0.05;3} = 17.7$).

### The Parts of Speech in the Text (Size)

Compared with the FDC, where the ratio of the nominal group to the verbal group is 1:1, the dialogue text is characterized by a higher verbal group (63.27%) and by a lower nominal group (34.47%) due to a specific repetition of words. In contrast to these characteristics, in the monologue text the nominal group is higher (53.74%) than the verbal group (44.58%). The differences are statistically significant ($X^2 = 6575.68 > \chi^2_{0.05;2} = 15.2$). In the dialogue text, there are the features of spoken language, whereas in the monologue text we find the features of written language (cf. Těšitelová et. al. 1983).

Within the framework of the nominal group in dialogue as well as in monologue texts, the number of nouns is higher than we find in the FDC (56.60%). As to the adjectives, their number in dialogue and in monologue is obviously lower than that in the FDC (22.76%). The prepositions are repeated more often in dialogue and even monologue texts than in the FDC (20.64%), although the nouns are repeated less often in dialogue than in monologue. It depends evidently on the spoken character of dialogue texts. The above-mentioned differences are statistically significant ($X^2 = 116.65 > \chi^2_{0.05;2} = 15.2$).

The prevailing verbal group in the dialogue texts $N_D$ is above all caused by the frequency of pronouns surpassing the respective amount in the FDC (22.21%) - about 7%; the frequency of pronouns in the monologue texts $N_B$ differs significantly from that for dialogue texts (29.01%). It depends obviously on the position of pronouns in spoken texts. As to the frequency of adverbs, it is in agreement with the FDC (20.94%). The dialogue text favours an asyndeton of sentences and words, a loose juxtaposition of clauses which is a feature of spoken texts (cf. Těšitelová et al. 1983, esp. p.115). In contrast to the dialogue texts, the higher number of conjunctions in monologue can be conceived as a feature of written texts with complicated sentence structure. The above-mentioned differences are statistically significant ($X^2 = 385.15 > \chi^2_{0.05;3} = 17.7$).

In the neutral group, the parts of speech are represented by numerals and interjections as in the FDC. The interjections have the highest frequency in dialogue (29.82%); in monologue this figure is lower (21.57%). The frequency of interjections bears out the characteristic feature of dialogue (cf. with $V_D$), especially the feature of spoken text, and above all of everyday language; the differences are statistically significant ($X^2 = 26.3391 > \chi^2_{0.05;1} = 12.1$).

The individual style of authors and the themes of texts are obviously reflected in the relations of three groups of parts of speech and their components; according to the character of different parts of speech the frequency of these components are complemented and conditioned by one another. The dialogue and monologue texts are distinguished by the above-mentioned quantitative characteristics. Above all these results have been corroborated by those obtained in the quantitative analysis of the three texts of present-day language.



### Conclusions

On the basis of the specially prepared analyses of the material FDC (groups D and B) and of the three control texts of present-day language (the drama H and the two poetic texts Ž and S), other phenomena marked for dialogue and monologue were also investigated. The results of these analyses can be summarized as follows:

In the **vocabulary** of dialogue $V_D$ the verbal group of parts of speech occupies an insignificant place (due to the special relation of different verbs and adverbs); in the vocabulary of monologue $V_B$ the nominal group plays a significant role (due to the higher occurrence of different adjectives). In both cases the predominance of verbs or nouns is caused by the different parts of speech modifying their meanings in context.

As to the **text (size)**, the verbal group is most frequent in the dialogue text $N_D$ and the nominal group in monologue. The verbal group of $N_D$ is dominated by the pronouns in relation to the verbs and the adverbs. In monologue text $N_B$ the nouns are the most frequent part of speech in the nominal group in relation to the adjectives and the prepositions. In dialogue and monologue texts, the three groups of parts of speech are used individually in accordance with the semantic features of the respective parts of speech, with the characteristics of individual style of different authors, etc. The word is repeated more often in dialogue; a new word is chosen more often in monologue.

In the **semantic analysis** of the most frequent words, esp. of the three groups of parts of speech, the detailed results of which will be published elsewhere, the following differences between dialogue and monologue appeared:

(1) In the nominal group, the semantics of nouns and adjectives are different in dialogue from those in monologue.
(2) In the verbal group, the pronouns play a dominant role in dialogue, namely in combination with the polysemic verbs and the pronominal adverbs.
(3) In the neutral group, the distinctions in the semantics of interjections are insignificant for dialogue as well as for monologue.

The quantitative characteristics of **morphological categories** concerning
(a) nominal (nouns, adjectives, pronouns) and
(b) verbs distinguish dialogue from monologue as follows (the detailed analysis will be published elsewhere):

(a)

1. Case and number with *nouns*. The most significant differences in the frequency of cases consist in the cases as a rule with lower and lowest frequency: In dialogue there is, e.g., the vocative sg., i.e. we address most often in this case in sg. This is a feature of spoken text. In monologue the typical cases are the nominative sg. and the genitive sg. and pl., i.e. the subject and the incongruent attribute play a significant role in the constructions of this type of style; that is a feature of written text.

2. Case, number and gender with *adjectives*. In dialogue and monologue, the nominative sg. and pl. and the accusative sg. represent the most frequent cases with adjectives with regard to their gender too. The significant differences between both types of style consist in the frequency of the cases with lower frequency, e.g., in that of the genitive and the instrumental case, i.e. in that of the congruent attribute and of the nominal predicate.

3. Case (number and gender) with *pronouns*. In the use above all of the personal pronouns whose forms lack grammatical gender there are the most significant differences between dialogue and monologue. E.g., the dative is the most frequent case in dialogue with the pronouns *my, ty, vy,* in monologue with those of *já* and *my*.

(b)

Frequency of the morphological categories with *verbs*. The present tense (in the indicative) is most frequent in dialogue as well as in monologue. The frequency of the past tense (in the indicative) in dialogue is in agreement with that in monologue. The future tense is more frequent in monologue than in dialogue. The conditional (of the present tense) occurs more often in dialogue than in monologue. As to the category of person and number, the 1st person sg. is more frequent in dialogue, the 3rd person pl. in monologue, in both cases next to the 3rd person sg. These characteristics are in agreement with those distinguishing the spoken texts from the written ones.

The above results characterizing the constructions of dialogue and those of monologue can be generalized. It holds for the basic features predicated by J. Mukařovský (1948). The quantitative characteristics of dialogue are in agreement with the features of spoken texts, those of monologue with those of written texts. On this quantitative basis dialogue can be distinguished from monologue, - let us say, the dramatical text from the poetic one.



## References

**Bečka, J.V.** (1970). *Stylistická syntax a kompozice projevu* (Stylistic syntax and composition of utterance). Prague, Státní pedagogické nakladatelství.

**Bečka, J.V.** (1992). *Česká stylistika* (Czech stylistics). Prague, Academia.

**Hoffmannová, J.** (1989). Práce českého teatrologa o (dramatickém) textu a (divadelní) komunikaci (The work by a Czech theatricologist on a dramatic text and theatre communication). *Slovo a slovesnost 50, 331-339.*

**Hoffmannová J. & Müllerová, O.** (1992). Vývoj a současné akcenty analýzy dialogu (Development and present-day accents of the analysis of a dialogue). *Slovo a slovesnost 53, 111-122.*

**Jelínek, J., Bečka, J.V. & Těšitelová, M.** (1961). *Frekvence slov, slovních druhů a tvarů v českém jazyce* (Frequency of words, parts of speech and word-forms in Czech). Prague, Státní pedagogické nakladatelství.

**Mistrík, J.** (1971). Kvantitatívna analýza lexiky dialógu (Quantitative analysis of dramatic lexicon). *Jazykovedné štúdie (Bratislava) 11, 5-19.*

**Mistrík, J.** (1979). *Dramatický text* (Dramatic text). Bratislava, Slovenské pedagogické nakladatel´stvo.

**Mistrík J.** (1985). *Štylistika* (Stylistics). Bratislava, Slovenské pedagogické nakladatel´stvo.

**Mukařovský, J.** (1948). *Kapitoly z české poetiky I* (Chapters from Czech poetics). [Cf. the chapter `Dvě studie o dialogu` (Two studies on dialogue), pp. 129-156]. Prague, Svoboda.

**Muller, C.** (1962). Les "pronoms de dialogue": interprétation stylistique d'une statistique de mots grammaticaux en français. In: Straka, G. (ed.), *Actes du X^e congrès international de linguistique et philologie romanes.* Paris, Klincskieck 1965: 605-612.

**Müllerová, O.** (1979). *Komunikativní složky výstavby dialogického textu* (Communicative components of construction of dialogical text). Prague, Universita Karlova.

**Těšitelová, M.** (1968). O básnickém jazyce z hlediska statistického (On poetic language from the statistical viewpoint). *Slovo a slovesnost 29, 362-368.*

**Těšitelová, M.** (1974). *Otázky lexikální statistiky* (Problems of lexical statistics). Prague, Academia.

**Těšitelová, M.** (1992). *Quantitative linguistics.* Prague-Amsterdam-Philadelphia, Benjamins.

**Těšitelová, M. et al.** (1983). Psaná a mluvená odborná čeština z kvantitativního hlediska (v rámci věcného stylu). (Written and spoken scientific Czech - with non-fiction style) (*Linguistica IV).* Prague, Institute of the Czech Language (internal publication).

**Těšitelová, M. et al.** (1983a). *Frekvenční slovník češtiny věcného stylu* (Frequency dictionary of the non-fiction style in Czech). Prague, Institute of the Czech Language (internal publication).

# The Significance of the Statistical Method for an Interdisciplinary Analysis of Professionalism of Texts

*Klaus-Dieter Baumann, Leipzig*

## I. Introduction

There is no question that rapidly developing technological progress belongs to those phenomena which to a very high degree are determining the present development of human society.

The changes in the fields of science and technology can be regarded as a substantial revolution which started in most industrialised countries by the middle of the 1950's. The mastery of this scientific-technological revolution demanded a modification of specialised communication in the following respects. Languages for Specific Purposes (LSP) should:

1. guarantee effective communication among all those people working together in this or that way, in this or that profession, industry etc.;

2. enable an exact exchange of knowledge (e.g. in special books, dictionaries, encyclopaedias, computer assisted data banks...);

3. support intellectual activities by means of abstractions, generalisations etc.;

4. be carefully analysed and confronted with properties of general language;

5. be analysed according to specific features of specialised subject fields (mathematics, physics, chemistry, biology... - history, linguistics, philosophy, economy...) and (foreign) languages (English, German, Russian, French, Spanish...);

6. stimulate improved job-specific language training methods at universities and colleges.

Due to rapid scientific-technological progress changes could also be observed in the usage of language means. Obviously, the significance of pro-



fessionalism has been increasing recently to a very large extent. In this context the extensive development of LSP research - especially since the 1960's - is easily understood.

In LSP research there have always existed a considerable number of linguistic schools and traditions (Hoffmann 1984: 21-71). It goes without saying that not all these theories have played the same role in the practical analyses of LSP phenomena.

In the beginning, LSP research dealt with lexicological items because subject specialists and LSP experts soon agreed with each other that it is the vocabulary which is very characteristic of LSP. So the notions 'LSP' and 'Vocabulary of LSP' have been used as synonyms for a certain time (especially in the 1960's). Later, various investigations in the field of LSP showed bluntly that the essence of LSP couldn't be explained exhaustively only by lexical means. At the end of the 1960's LSP research was shifting more and more to syntax.

After that, there were quite a lot of attempts to deal with syntactic peculiarities of LSP on three levels: the level of syntagma(s), phrase(s) and sentence(s) (Hoffmann 1976: 339). But what soon became evident, however, was that a more comprehensive description of all those language means used in the process of specialised communication could only be realised sufficiently on the *text* level. And so, at the beginning of the 1970's, the so-called 'communicative-pragmatic change' of linguistics took place (Helbig 1986: 13).

Consequently, the main interest of linguistics moved from the structural view of language system (especially on the syntactic and semantic levels) to a complex functional view of all levels of communication. The integration of the language system into the analysis of the communicative process and social interaction has made it possible to consider a greater number of features determining the complexity of communication. So language is defined as a network of language signs which has no end in itself but is an instrument for realising extralinguistic purposes. That's why it is also determined by extralinguistic factors.

In the following years this new interpretation promptly led to an extension of linguistics. Step by step linguistic research took up the text as the basic unity of communication - and a lot of non-linguistic phenomena surrounding text-production and text-reception were also taken into account. It is a fact that this new trend manifested itself in the development of further branches of linguistics: e.g. text-linguistics, sociolinguistics, functional and communicative description of language, etc. (Crystal 1982: 248). In each case, all these linguistic disciplines analyse one and the same subject, i.e. *language*. A cardinal principle underlying this linguistic approach is that language is not an isolated phenomenon. It is a part of society, social interaction, social and/or individual behaviour, etc. That's why all the above-mentioned linguistic disciplines are obviously interconnected and it seems impossible to clearly separate them from each other.

So the point is that these branches can be regarded as *integrative* directions of linguistics, approaching language phenomena under one dominating aspect of

communicative activity from different directions. This insight into language complexity to a large extent determined our methodical approach to the analysis of LSP texts.

After the communicative-pragmatic change in linguistics at the beginning of the 1970's there was already by the mid 1970's a considerable rise in interest in LSP research. Thus more and more LSP-oriented analyses were dealing with texts as ensembles of various morphological, semantic and syntactic constituents and text-organising principles respectively. Texts were regarded as units realising specific communicative functions in different fields of knowledge.

Nevertheless, the idea that the extension of the area of interest in LSP research was the automatic and logical consequence of the (complex) phenomenon 'LSP', would be too simple. As is generally known, objects are phenomena of the surrounding objective reality always existing independently from consciousness and only being reflected by cognitive processes. The individual intentions being pursued by the proper human process of cognition make possible various approaches to those objects of reality which are conveyed by particular phenomena. And these approaches can - in principle - be regarded as the starting point of scientific disciplines with their specific area of interest.

The hitherto existing history of LSP research has illustrated that we always have had to do with one and the same object, i.e. LSP as a means of communication being used among those people working in various fields; but the predominant object of interest has often changed in the history of LSP research (semantics, stylistics, functions of texts, etc.).

Though the scientific illustrations of LSP strive for a totality, the interpretation of LSP conveyed by only one specific approach must remain rather limited.

When analysing LSP it seems to be very problematic to concentrate on only one aspect and to understand this as *the whole* of LSP research. An exaggeration of one/some aspect(s) has always led to a one-sided theoretical orientation and to corresponding consequences for the methodical approach to LSP.

The results of our analyses show that such a restricted conception (i.e. a conception considering only a few properties of LSP) isn't able (and cannot be able) to draw a complete picture of LSP texts. That's why we have tried to work out an integrated concept of investigation referring to the complexity of LSP.

Before approaching the structural and functional totality of LSP texts it was necessary to develop a complex methodology. At this point it was appropriate to take into account the interrelation between language system and communicative activity.

In our text analyses we have started from the following three premises:



A. LSP texts are regarded as texts-in-function. By this we understand complex units consisting of social, situational, and thematic factors as well as of structural, stylistic and formal features. LSP texts-in-function are the result of an interaction between the so-called text internals and text externals. In our conception we have analysed LSP texts-in-function with the help of the categories of the *functional-communicative description of language.*

B. LSP texts are characterised with respect to

   a. the social relationship between the author(s) of the text and its addressee(s);
   b. the object of communication and
   c. the degree of professionalism of communication (text externals).

   We have analysed all these aspects with the help of the categories of *sociolinguistics, psycholinguistics, language theory, cultural studies and corresponding branches of science.*

C. LSP texts are distinguished by a specific language structure (text internals). This text structure can best be investigated by categories of *lexical semantics, functional stylistics and textlinguistics* (Baumann 1992).

So altogether we have integrated eight various approaches into our interdisciplinary concept of LSP texts. These eight approaches do correspond to eight hierarchically arranged levels or dimensions of investigation:

1. the intercultural dimension being analysed by categories of *cultural studies* (Clyne 1987, 211-247);

2. the social dimension covered by categories of *sociolinguistics* (Ammon 1977, 1991; Hartung 1976; Wardhaugh 1990);

3. the cognitive dimension analysed by categories of *psycholinguistics* (Goeppert 1973; Leontjew 1987; Lomov 1987; Lompscher 1989; Schwarz 1992; Leont'ev, Leont'ev & Judin 1984);

4. the dimension with regard to the contents being analysed with the help of the corresponding *branch of science;*

5. the functional dimension covered by categories of *functional description of language* (Schmidt 1981; Michel 1986);

6. the textual dimension analysed by categories of *textlinguistics* (Baumann 1992);

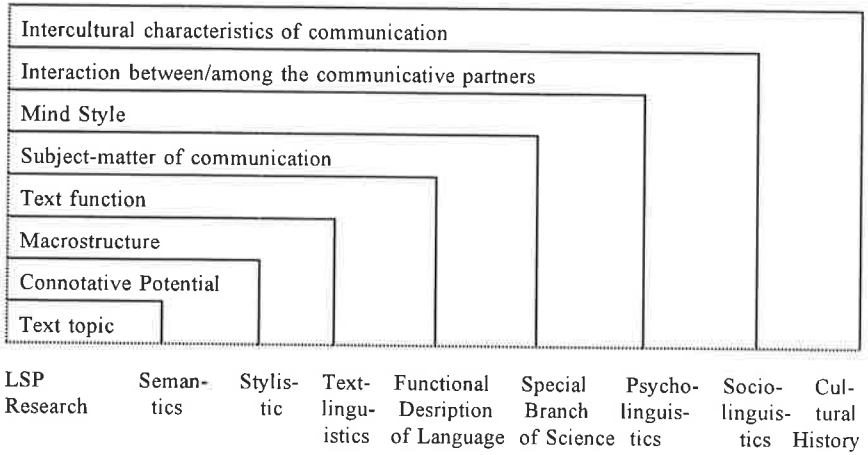
7. the stylistic dimension analysed by categories of *linguostylistics* (Fleischer & Michel 1975; Riesel & Schendels 1975);

8. the semantic dimension covered by categories of *lexical semantics* (Baumann & Kalverkämper 1992).

It is important to be aware that each level or dimension of analysis refers to a specific field of structural and/or functional features, e.g. each level opens a special approach to LSP texts.

By analysing various LSP text forms (monograph, scientific article, textbook, essay) of different sciences (historiography, linguistics, psychology) in English, Russian and German (our text corpus consists of 1772 pages, 5998 paragraphs and 26,226 sentences) we have tried to make evident the complex correlations among the above-mentioned branches of linguistics with the help of some central categories of analysis (Baumann 1992):

Text Forms




Undoubtedly, this multi-level text analysis has opened a new developmental stage of LSP research.

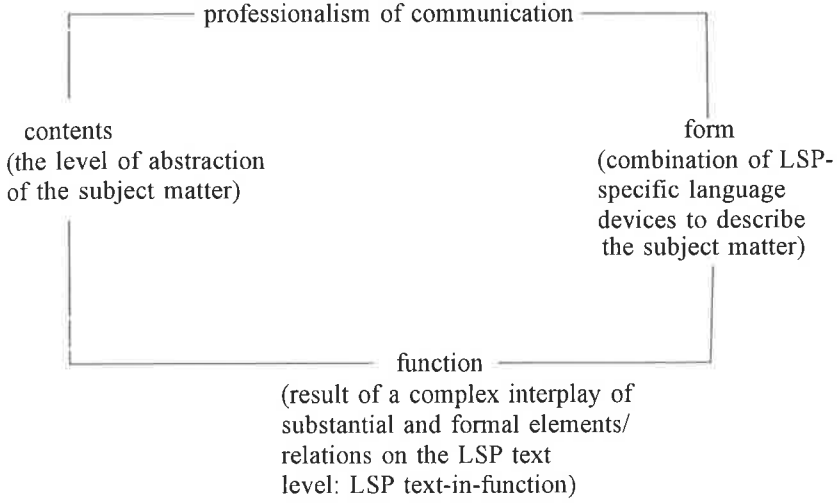


## II. Professionalism of Texts

There are results of our multi-level text analyses pointing out that professionalism is a central category in describing communicative processes (Baumann 1992). By *professionalism* of communication, we understand a complex system of extralinguistic and intralinguistic interrelations. This system of interrelations is being determined by subject-specific, convention-based, and functional elements relations.

It goes without saying that professionalism of communication is a dynamic category.

The significance of professionalism for text analyses results from the following components:




The complex character of the professionalism of texts can only be revealed if we succeed in describing its structural and functional aspects, finding out the hierarchical position of its components, and analysing their interrelations. Based on the description of professionalism, it will become possible to characterize the multidimensionality of communicative processes more deliberately.

Unfortunately, up to now, linguistic and non-linguistic disciplines have not been successful enough in pinning down exact criteria to determine degrees of professionalism of texts.

Our complex LSP text analyses in the fields of historiography, linguistics, psychology and business communication have confirmed our assumption that professionalism can only be analysed by means of an interdisciplinary approach. The interdisciplinary concept of investigation which we have been developing

for some years comprise methods and categories from cultural studies, sociology of language, sociology, psycholinguistics, psychology, pragmalinguistics, LSP text linguistics, functional description of language, linguostylistics, semantics, terminology, comparative studies, statistics, and the particular science.

Of course, there are several ways to systematically describe those elements and relations which contribute to the degree of professionalism of terms.

In our analyses, we have distinguished the following hierarchically arranged levels of professionalism: the intercultural, social, cognitive, subject matter, functional, textual, stylistic and semantic elements.

## III. The Significance of Statistical Methods for Determining Degrees of Professionalism of Texts

The methods of an interdisciplinary analysis of degrees of professionalism of texts form a dynamic system. This means that all those rules, principles and procedures contributing to a comprehensive analysis of professionalism of texts can be regarded as one method of interdisciplinary LSP research. It goes without saying that theoretical premises and practical aspects of interdisciplinary text analyses are always interdependent. In order to analyse the phenomenon of professionalism it is necessary to employ methods of all those disciplines being integrated in the above-mentioned multi-level approach to the degree of professionalism of texts.

Within any level of the interdisciplinary approach to professionalism of texts it is possible to count the different elements and relations that occur, and interrelate the frequencies obtained, to see if there are statistical regularities.

The study of these regularities, and of the factors that constrain them, is the province of *statistical linguistics.* Statistical linguistics is a branch of applied linguistics which tries to find underlying regularities in any large sample of communication. The importance of statistical linguistics became evident in the 40`s. At that time many scientific works on information and probability theory referred to language as a means to transmit and to store information (Muller 1972; Altmann 1972; Alexeev, Kalinin & Piotrowski 1973; Hoffmann & Piotrowski 1979).

Besides, it was shown that language fulfils the function of a code. Subsequently, the interpretation of language as a functional code with specific statistical characteristics definitely opened the way for mathematical methods to be applied in linguistics.

In particular, methods of statistical linguistics concentrate on the following operations:

- the exposition of the frequency of occurence of individual textual elements;



- the demonstration of statistical regularities within a sample of LSP texts;

- the determination of probabilities concerning the occurrence of various elements/relations in an LSP text (form) and

- the demonstration of constant regularities in language used in an LSP text (form).

LSP research and statistical linguistics combine within the framework of an interdisciplinary approach to professionalism of texts in order to describe the variable elements, relations, and functions of LSP texts statistically and to analyse these statistical properties systematically (Hoffmann & Piotrowski 1979).

A complete statistical analysis of the structural *and* functional aspects of various LSP texts would decisively contribute to an overall determination of the degrees of professionalism of texts.

In my extensive LSP text analyses I have found that statistical methods are generally essential to work out various degrees of professionalism. But with respect to the various levels of the interdisciplinary approach to professionalism of texts, the importance of statistical methods is variable.

### 1. The Socio-cultural Dimension of the Analysis of Professionalism

This aspect of text forms has remained unconsidered for many years. Now we are dealing with this side of text manifestation, especially in connection with comparative text analysis. A systematic comparison of English, Russian and German texts drew our attention to the fact that some text forms do occur more frequently in a specific language community than in another.

That is the case with the LSP text form 'essay': it is frequently used in English, but less important in Russian or in German.

M. Clyne has pointed out that in English and German scientific communication there are different organisational structures. He emphasised: "Linearity, symmetry, hierarchy and continuity are examined in 52 texts as are the position of definitions and advance organizer and the integration of data. It is suggested that the difference between the English and German texts may be promoted by the education system and by varying intellectual styles and attitudes to knowledge and contents" (Clyne 1987: 211).

It seems to be a matter of course that communicative structures and text structures are culturally bound. It has become one of our main research fields. But statistical methods haven't become relevant yet for the analysis of the socio-cultural aspects of professionalism.

## 2. The Social Dimension of the Analysis of Professionalism

This dimension covers the social variability of language means. The central category for analysing this level is the *interaction* of the communicative partners. By interaction we understand the mutual interrelation of the communicative partners. This mutual interrelation takes place under the specific conditions of the proper communicative situation. It is a complex of the

- situation of activity (branch of science; number of participants; structure of interaction - monologue, dialogue, polylogue; communicative intentions; object of communication; medium, local and temporal aspects of the communicative situation);

- social situation (social status of the communicative partners, political point-of-view, values, familiarity, publicity);

- situation of the surroundings (expectations, idiolect, intellectual abilities, motivation, sex, age, physical state of the communicative partners, paralingual elements etc.).

At this level statistical methods couldn't have been used yet because at the moment it is still very difficult to analyse social aspects and their effects on a specific degree of professionalism quantitatively.

## 3. The Cognitive Dimension of the Analysis of Professionalism

By analysing various sorts of texts from different sciences we tried to make evident the correlations between specialist thinking and the degree of professionalism of texts. By specialist thinking we understand a form of thinking aimed at the solution of theoretical problems, connected with practice only indirectly.

Undoubtedly, LSP texts are especially suitable for researching the close relationship between professionalism and the way of thinking. In this way we succeeded in demonstrating the outside importance of LSP as the instrument of thinking in the different scientific spheres of communication. From this we could prove how significant the cognitive component of an LSP text is in order to differentiate among various degrees of professionalism.

Our specific way of investigating the very complex correlations between professionalism and specialist thinking in texts is to consider the frequency of meta-communicative text elements (words). Meta-communicative text elements always help to make texts easier to understand by paraphrases, explications,



translations of terms, references, etc. (Baumann 1992: 91 ff). Meta-communication can also be defined as "communication about communication" (Techtmeier 1984: 122). So meta-communicative elements guarantee the adequacy of the communicative activities decisively.

In our investigation we determined that a high portion of meta-communication obviously refers to a lower degree of professionalism. The author of the text has to involve a lot of meta-communicative elements to make its understanding easier for those addressees not familiar with the specific topic of discussion.

The following results of quantitative analyses shed light on the relation of inverse proportionality between meta-communication on the one hand and the degree of professionalism on the other. The proportion of meta-communication in the text is expressed by the formula:

$$\frac{\text{number of meta-communicative elements} \times 100}{\text{number of sentences in the text.}}$$

| ENGLISH | Historiography | | | Linguistics | | | Psychology | | |
|---|---|---|---|---|---|---|---|---|---|
| | FM CE | S | RF | FM CE | S | RF | FM CE | S | RF |
| monograph | 63 | 439 | 14.35% | 221 | 604 | 36.58% | 120 | 400 | 30% |
| scientific article | 125 | 458 | 27.29% | 85 | 310 | 27.41% | 73 | 338 | 21.59% |
| text-book | 312 | 600 | 52% | 167 | 322 | 51.86% | 417 | 436 | 95.64% |
| essay | 425 | 638 | 66.61% | 347 | 600 | 57.83% | 317 | 782 | 40.53% |

| RUSSIAN | Historiography | | | Linguistics | | |
|---|---|---|---|---|---|---|
| | FM CE | S | RF | FM CE | S | RF |
| mono-graph | 57 | 233 | 24.46% | 88 | 265 | 33.2% |
| scientific article | 50 | 242 | 20.66% | 47 | 295 | 15.93% |
| textbook | 193 | 334 | 57.78% | 289 | 434 | 66.58% |

*FMCE - frequency of metacommunicative elements*
*S - sentences; RF - relative frequency*

So, for example, the proper degree of professionalism of scientific articles seems to be higher - because of the small number of meta-communicative elements - than in textbooks. In the latter text form we observe a larger number of meta-communicative elements. Obviously, statistical methods play a useful role in analysing the cognitive aspects of professionalism.

## 4. The Dimension with regard to the Contents of the Analysis of Professionalism

The significance of this component for the degree of professionalism is still being analysed. The object or topic of the text has an influence on the manifestation and the degree of professionalism of the text form. It is important to consider whether the text covers a process, a single item or state of fact, a problem of natural, social or technical sciences.

There is no doubt, that the object of communication determines the thematic base of the text, influences the semantic progression and the macrostructure of the text.

But on that level of analysis statistical methods haven't been used yet.

## 5. The Functional Dimension of the Analysis of Professionalism

In order to establish a homogenous basis for the functionally determined classification of degrees of professionalism of texts we have introduced five complex intellectual procedures (description, narration, exposition, argumentation, instruction) (Werlich 1976), correlating with the basic cognitive processes of human beings. With respect to their linguistic realisation, these complex procedures can be regarded as so-called combined cognitive-communicative procedures. The conception of the communicative procedures introduced by W. Schmidt (Schmidt 1981) and others can be integrated into the system of the five complex procedures without any difficulties. The interrelations between the intention(s) of the author(s) - the textual function - the complex procedures - the communicative procedures and the proper usage of language means open a favourable way to the functional classification of degrees of professionalism of LSP text forms. In this way our quantitative analyses refer to the following trends:

- The complex procedure 'narration' is the dominating one in historiographical LSP texts. Above all, it realises the textual function (intention) 'information' (Baumann 1986: 126-7).



- The complex procedure 'instruction' only exists in the LSP text form 'textbook'.

- The rank of frequency of the complex procedures in English LSP text forms is the following:
  1. exposition (clarification) 36.18%;
  2. description (information) 26.03%;
  3. argumentation (activation) 15.54%.

- In Russian LSP text forms we have found the following range of complex procedures:
  1. exposition (clarification) 52.69%;
  2. description (information) 25.90%;
  3. narration (information) 9.30%;
  4. argumentation (activation) 8.47%.

- The complex procedure 'get into contact' was found particularly in our LSP text forms of linguistics.

Statistical analysis has helped us to find the following correlation: the higher the percentage of 'exposition' and 'argumentation', the higher the degree of professionalism of the text. This is because these two complex procedures obviously tend to be more 'abstract' than the complex procedures 'description' and 'narration' (Baumann 1992).

## 6. The Textual Dimension of the Analysis of Professionalism

As a result of the analysis of our text corpus we can show that the surface structure of a text doesn't reproduce semantic coherence. This means that the semantic elements don't primarily structure the text (Baumann 1992: 75-103). Obviously, other factors - especially pragmatic ones - play a more important role. An exact analysis of the relationship between isotopy and text structuring will only be possible if the notion of the 'text paragraph' as a structural text unit is defined more precisely. We absolutely agree with T. Mage's point of view (he is a representative of the Washington School of Linguistics) characterizing 'text paragraphs' as a 'contents - form relation' (Mage 1978: 154-166).

Analogously, we differentiate between the 'physical' and 'conceptual' paragraphs. A physical paragraph is a group of sentences marked on a page of text by spacing or indentation.

A conceptual paragraph is defined as a group of organizationally related concepts which develop a given generalization in such a way as to form a

coherent (complete) unit of discourse. It can consist of one or more physical paragraphs.

The following results make clear the correspondence of physical/conceptual paragraphs. The degree of correspondence expressed by the formula:

$$\frac{\text{number of conceptual paragraphs} \times 100}{\text{number of physical paragraphs of the proper text.}}$$

| ENGLISH | Historiography | | | Linguistics | | | Psychology | | |
|---|---|---|---|---|---|---|---|---|---|
| | CP | PP | C | CP | PP | C | CP | PP | C |
| monograph | 22 | 93 | 23.65% | 45 | 55 | 81.81% | 26 | 55 | 47.27% |
| scientific article | 83 | 177 | 46.89% | 12 | 77 | 15.58% | 27 | 107 | 25.23% |
| textbook | 116 | 116 | 0 | 30 | 195 | 15.38% | 56 | 217 | 25.80% |
| essay | 10 | 105 | 9.52% | 23 | 86 | 26.74% | 8 | 230 | 3.47% |

| RUSSIAN | historiography | | | Linguistics | | |
|---|---|---|---|---|---|---|
| | CP | PP | C | CP | PP | C |
| monograph | 4 | 66 | 6.06% | 3 | 65 | 4.61% |
| scientific article | 6 | 66 | 9.09% | 9 | 188 | 4.78% |
| textbook | 11 | 71 | 15.49% | 117 | 117 | 100% |

*CP - conceptual paragraph*
*PP - physical paragraph*
*C - correspondence*

Our statistical analyses helped to find the following correlation: the higher the degree of correspondence between physical and conceptual paragraphs, the lower the degree of professionalism of the text.

If the degree of correspondence between physical and conceptual paragraphs is higher, the author tries to give much more structural help to the recipients of the texts (for example textbooks).

## 7. The Stylistic Dimension of the Analysis of Professionalism

LSP research and functional stylistics refer to relevant aspects of language organisation of LSP text forms from different points of view. Especially in the field of natural and social sciences with their large number of different text



forms and their varying stylistic colourings, we can gain new insights into the stylistic characterization of text forms. Structural and functional differences in the stylistically relevant text elements can only be explained when certain principles about the relation of text externals to language means are established. Additionally, the proper stylistical colouring of LSP text forms is determined to a great extent by the various degrees of professionalism.

Hence we have introduced the concept of 'stylistical potential' into our text analyses. It is defined as the specific form of stylistically relevant elements interacting on all levels of the language system and non-language complexes of signs. These contribute together to the constitution of the individual sense - in contrast to the meaning - of the LSP text and they influence the interpretation of the text contents. On the one hand, this stylistic potential represents a qualitative-functional aspect of the LSP text style. On the other hand, the frequency, distribution and combination of stylistically relevant text elements refer to the quantitative-structural part of the style. In our text analyses we tried to determine the connotative quotient according to the following formula:

$$\frac{\text{absolute frequency of stylistically relevant elements of the LSP text}}{\text{number of sentences of the LSP text}}$$

We found the following connotative quotients:

| ENGLISH | Historiography | | | Linguistics | | | Psychology | | |
|---|---|---|---|---|---|---|---|---|---|
| | SE | S | CQ | SE | S | SQ | SE | S | CQ |
| monograph | 339 | 434 | 68.62% | 531 | 604 | 87.9% | 265 | 400 | 66.2% |
| scientific article | 204 | 458 | 59.52% | 53 | 193 | 27.4% | 448 | 860 | 52.1% |
| textbook | 313 | 660 | 47.4% | 295 | 322 | 91.6% | 182 | 435 | 41.9% |
| essay | 492 | 638 | 77.10% | 466 | 600 | 77.7% | 464 | 638 | 72.8% |

| RUSSIAN | Historiography | | | Linguistics | | |
|---|---|---|---|---|---|---|
| | SE | S | CQ | SE | S | CQ |
| monograph | 111 | 233 | 47.6% | 151 | 181 | 83.4% |
| scientific article | 154 | 242 | 63.7% | 321 | 348 | 92.2% |
| textbook | 244 | 334 | 73% | 226 | 434 | 52.1% |

*SE - stylistical element*
*S - sentences*
*CQ - connotative quotient*

Our interdisciplinary LSP analyses demonstrated that stylistically relevant elements support the comprehension of text to a high degree. Stylistical devices function as elements of a partner-related redundancy and contribute to the compensation of the different levels of knowledge between author(s) and addressee(s). If the presuppositions of the communicative partners are comparable the number of stylistical devices is lower. Obviously, they aren't necessary for decoding the contents of the text.

Thus the analysis of professionalism is closely connected with the connotative quotient of text. Quantitative analysis illustrates that the relationship between the stylistically relevant elements of text and the degree of professionalism of the text is inversely proportional: the higher the connotative quotient, the lower the degree of professionalism of this text (Baumann & Kalverkämper 1992: 41-42).

## 8. The Semantic Dimension of the Analysis of Professionalism

The usage of a clearly defined terminology in a text points out that experts or specialists are communicating. Our analyses illustrated that a proper degree of professionalism is dependent upon the qualitative structuring of scientific terminology and the number of terms in a text.

The proportion of terms with respect to the lexical totality of the text:

| ENGLISH | Historiography | | | Linguistics | | | Psychology | | |
|---|---|---|---|---|---|---|---|---|---|
| | FT | NLI | RF | FT | NLI | RF | FT | NLI | RF |
| monograph | 615 | 15048 | 4.08% | 1059 | 15444 | 6.85% | 928 | 12672 | 7.32% |
| scientific article | 515 | 9108 | 5.65% | 292 | 3564 | 8.19% | 1344 | 10692 | 12.57% |
| textbook | 453 | 11088 | 4.08% | 2611 | 34452 | 7.57% | 2024 | 18612 | 10.87% |
| essay | 1085 | 16236 | 6.68% | 1606 | 14256 | 11.26% | 1631 | 16236 | 10.04% |

| RUSSIAN | Historiography | | | Linguistics | | |
|---|---|---|---|---|---|---|
| | FT | NLI | RF | FT | NLI | RF |
| monograph | 267 | 6732 | 3.96% | 420 | 5148 | 8.15% |
| scientific article | 404 | 5940 | 6.8 % | 490 | 6732 | 7.27% |
| textbook | 381 | 7920 | 4.81% | 329 | 8316 | 3.95% |

*FT - frequency of terms*
*NLI - number of lexical items*
*RF - relative frequency*



Thus it can be seen that the higher the number of terms, the higher the degree of professionalism of the text. Quantitative analyses play a great role in the interdisciplinary consideration of the multi-level phenomenon 'professionalism of texts'. They are one among other procedures to show tendencies as far as interdependencies of structural and/or functional elements and relations of professionalism are concerned. The interdisciplinary approach to the professionalism of texts opened a qualitatively new epoch of LSP research (Baumann 1992). But statistical analyses of various aspects of professionalism showed that the usual linguo-statistical methods (i.e. chi-square-test and others; Hoffmann & Piotrowski 1979: 107-114) aren't appropriate for the description of *complete* LSP texts as a whole often comprising *more* than 100 printed pages.

Our textual experiments based on chi-square-tests didn't produce any useful results. It made clear that linguo-statistical methods lag behind the complex approach to LSP text forms. So we agree with L. Hoffmann's point-of-view that linguo-statistical analyses are only successful with respect to small random tests (Hoffmann & Piotrowski 1979: 107).

At the moment we are trying to develop a corpus of comparable quantities of LSP text forms in order to check the potentialities of the traditional linguo-statistical methods for an interdisciplinary description of professionalism of texts.

## References

**Altmann, G.** (1972) Status und Ziele der quantitativen Sprachwissenschaft. In: Jäger, S. (ed.), *Linguistik und Statistik*. Braunschweig, Vieweg 1972: *1-9.*

**Alexeev, P.M., Kalinin, W.M. & Piotrowski, R.G.** (1973). *Sprachstatistik.* Berlin, Akademie-Verlag (tr. L.Hoffmann).

**Ammon, U.** (1977). *Probleme der Soziolinguistik.* Tübingen, Niemeyer.

**Baumann, K.-D.** (1986). *Ein integrativer Ansatz zur Analyse von Fachkommunikation unter besonderer Berücksichtigung des kommunizierenden Subjekts in ausgewählten Fachtextsorten der Gesellschaftswissenschaften im Englischen und Russischen.* (Habilitationsschrift). Leipzig 1986.

**Baumann, K.-D.** (1992). *Integrative Fachtextlinguistik.* Tübingen, Narr.

**Baumann, K.-D. & Kalverkämper, H.** (eds.) (1992). *Kontrastive Fachsprachenforschung.* Tübingen, Narr.

**Clyne, M.** (1987). Cultural Differences in the Organisation of Academic Texts. *Journal of Pragmatics 11, 211-247.*

**Crystal, D.** (1982). *Linguistics.* Harmondsworth, Penguin 1982

**Fleischer, W. & Michel, G.** (eds.) (1975). *Stilistik der deutschen Gegenwartssprache.* Leipzig, VEB Bibliographisches Institut.

**Goeppert, S. & Goeppert, H.C.** (1973). *Sprache und Psychoanalyse.* Reinbek bei Hamburg, Rowohlt.

**Hartung, W.** (ed.). *Sprachliche Kommunikation und Gesellschaft.* Berlin, Akademie Verlag.
**Helbig, G.** (1986). *Entwicklung der Sprachwissenschaft seit 1970.* Leipzig, VEB Bibliographisches Institut.
**Hoffmann, L.** (1976). *Kommunikationsmittel Fachsprache. Eine Einführung.* Berlin, Akademie Verlag.
**Hoffmann, L.** (1984). *Kommunikationsmittel Fachsprache. Eine Einführung.* Berlin, Akademie Verlag (2., überarbeitete Auflage).
**Hoffmann, L. & Piotrowski, R.G.** (1979). *Beiträge zur Sprachstatistik.* Leipzig, Verlag Enzyklopädie.
**Leont'ev, A.A., Leont'ev A.N. & Judin, E.G.** (1984). *Grundfragen einer Theorie der sprachlichen Tätigkeit.* (Hrsg. D.Viehweger) Berlin, Kohlhammer.
**Leontjew, A.** (1987). *Tätigkeit, Bewußtsein, Persönlichkeit.* Berlin, Volk und Wissen.
**Lomov, B.** (1987). *Methodologische und theoretische Probleme der Psychologie.* Berlin, Volk und Wissen.
**Lompscher, J.** (1989). *Psychologische Analysen der Lerntätigkeit.* Berlin, Volk und Wissen.
**Mage, T.** (1978). Contrastive Discourse Analysis. In: Trimble, M.T., Trimble, L. & Drobnic, K. (eds.), *English for Specific Purposes: Science and Technology.* Oregon State University, English Language Institute: 154-166.
**Michel, G.** (ed.) (1986). *Sprachliche Kommunikation.* Leipzig, Bibliographisches Institut.
**Muller, Ch.** (1972). *Einführung in die Sprachstatistik.* Berlin, Akademie Verlag (Hrsg. von L. Hoffmann).
**Riesel, E. & Schendels, E.J.** (1975). *Deutsche Stilistik.* Moskau, Verlag Hochschule.
**Schmidt, W.** (ed.). *Funktional-kommunikative Sprachbetrachtung.* Leipzig, Bibliographisches Institut.
**Schwarz, M.** (1992). *Einführung in die kognitive Linguistik.* Tübingen, Francke.
**Wardhaugh, R.** (1990). *An Introduction to Sociolinguistics.* Oxford, Blackwell.
**Werlich, E.** (1976). *A Text Grammar of English.* Heidelberg, Quelle & Meyer.

# INTEXT
## A Program System for the Analysis of Texts

*Harald Klein, Lengerich*

INTEXT consists of thirteen programs performing different analyses of texts. The first version for MS-DOS was written in 1987, and the program language is C. A supervisor controls the programs, although each program can be executed without the supervisor. Versions for other operating systems (OS/2, UNIX and Apple MacIntosh) are planned. Currently there are four versions available that all have the same features, the same menus and produce the same results, but make use of different hardware and therefore have different limitations in some programs:

- The INTEXT/PC version. It runs on every PC under MS-DOS and needs a minimum of 384 KB free RAM.

- INTEXT/386 version. It does not run on simple XT computers and 80286 CPUs but on all others and can make use of EMS RAM up to 32 MB in the programs that perform indices, word combination lists, word permutation lists, content analysis and style analysis. Like all the versions described below it runs in protected mode under MS-DOS and uses a dos extender technique without requiring an external dos extender. The programs execute faster and work better with huge amounts of texts.

- INTEXT/486 version. The same advantages as INTEXT/386 also apply for this version, but the special hardware of an 486 CPU (cache) is supported.

The IRM (Intext Results Manager) program supports the extended modes of VGA cards, so that the context of search patterns can be displayed on a screen with 132 columns and 60 lines per page if the hardware allows it.

An English and a German version are currently available; the manual consists of over 100 pages including a glossary and an index. Versions in other languages can easily be generated due to the fact that all menu texts and information messages are in separate files that can be edited. The whole system requires less than 600 KB free disk space including example files. INTEXT was rewarded with a certificate for appreciated attainment and participation in the

German Software University Competition 1990 (out of nine programs for the humanities and social sciences two got this certificate).

## 1. Applications in the humanities

The following list contains the most used features in the humanities and the social sciences:

- index of all types (strings) that occur in a text; also reverse.
- word combinations with different number of words.
- word permutations.
- searching in indices, word combinations and word permutation lists using search entries (go words).
- excluding types in indices, word combinations and word permutation lists (stop words, file provided for English and German).
- excluding types in indices, word combinations and word permutation lists due to frequency or number of characters.
- comparison of two indices.
- search entries can be any part of a string; the wildcard characters ? and * are allowed.
- sort program that allows multiple characters sets, definition of sort orders and ignoring differences in upper-/lowercase (case folding).
- index can be sorted by alphabet or by frequency of types, ascending or descending.
- coding of open-ended questions in questionnaires.
- search entries in context (SIC) allowing outputs similiar to KWIC (Key-Words-In-Context) and KWOC (KeyWords-out-of-Context) with variable line length.
- crossreference lists, interactive and batch mode.
- content analysis in the social sciences with handling of ambiguous and negated search entries.
- readability analysis with eight different formulas (six for English, two for German texts).
- result managing program that allows printing in multiple columns for indices, header, page numbering, single sheets, configuring the printer allowing different characters per inch and lines per inch, all margins in millimeters, output can be directed to disk or to the screen, where VGA extended text modes are supported. The interface to $T_EX$ and its macro package $L^AT_EX$ makes it easy to produce high quality printing results.
- supervisor that makes entering file names nearly unnecessary.
- lemmatisation of German texts with LEMMA2 is supported.
- statistical analysis of content analysis with the PC-versions of SAS and SPSS are supported.



## 2. Structure of INTEXT: the supervisor IS

INTEXT consists of thirteen programs that generate files that are used by other programs of INTEXT or by third party programs. Some applications require the use of several programs in a specific order. The supervisor calls the needed programs in the required order and generates the necessary file names. With the cursor you choose the desired application (e.g. index, cross reference list, coding, printing), and execute the necessary programs. File names are generated using a system of file extensions described in the manual. This system works with a project name which consists of a file name and may contain the drive and the (sub)directory specification. Due to the name of the project the file names for input and output files are generated, so the tedious task of entering file names seldom occurs. Each program writes to a log files so that a control of file names and results is always possible.

All statistics at the end of a program are written to a log file that can be used with other text processing software. The name of the log file is derived from the project name that consists of the drive, directory and file name. The project name is used to derive file names from it, so that entering file names is almost eliminated, although one can change the suggested file name. INTEXT has a system of file extensions the programs work with. The log file contains all statistical data of the applications that were used working with the data of a project.

At the start of an application the file name of the program called first is presented in a line editor, a simple hit of the return key accepts the file name and starts the application. Output file names are generated and presented in the same manner, so just hitting the return key will do. Of course you can use your own system of file names, but this will not be supported by the supervisor. If the file does not exists or you forgot its name, you can leave the supervisor at this part of the program and exit to DOS, and return to where you stopped later.

Other programs are integrated into the supervisor. Currently the statistical packages SAS and SPSS/PC are supported, LEMMA2 can be used for the lemmatisation of German texts. There is also an interface to ConClus (Con-

strained Cluster analysis) that allows a fast and repeatable cluster analysis of the categories.

## 3. Text unit

As many other programs for text analyses INTEXT works with a system file. This file consists of the text and three external variables. The length of a text unit is limited to 32500 characters (ca. 16 pages of text), so that even chapters of books can be one text unit. External variables can be the name of an author, the number of a chapter or a page and so on. Three variables are supported without using tricks, they are all numeric and may have up to six digits. With tricks one can use 18 variables with one digit each.

The values of the external variables are to be defined by the researcher. An example: the different books of several authors are to be compared. External variables like name of the author, name of the book and page number make sense. The text unit is one page in this hierarchical system. If one wants to analyse the books of one author, the external variables book, chapter, and page or book, page and line are meaningful. The first solution requires a page to be a text unit, the second one the line to be a text unit. The second solution e.g. is not suitable for analysing contexts.

The values of the external variables are changed with control sequences. They may be inserted anywhere in the text and can define the values of the variables both relative (e.g. increase by one) and absolute. It is also possible to use new and/or blank lines for incrementing the values of external variables. The input modes are code line mode, paragraph mode and page mode. In line mode, every line is a text unit, and this mode makes it possible for one to process nearly every file without any editing. In paragraph mode every paragraph is a text unit (up to 32500 characters), and paragraphs are separated by at least one blank line. Page mode is a variation of the line mode: the line is the text unit, and every x lines - the x can be defined - the values of the second external variable is incremented. This feature is useful for scanned texts that came from OCR (optical character recognition) software.

## 4. Search entries

Search entries can be a string or any part of it, regardless where it occurs: at the beginning (in prefix position), at the end (in suffix position) or elsewhere (in infix position) of the string. Also differences in upper-/lowercase may be ignored (including umlauts and characters with diacritics). Also wildcard characters like ? (represents one character, regardless which one) and * (represents any number of characters, regardless of which one) are supported. The forms of



search entries described can be used searching in indices and for coding.

When performing a content analysis or style analysis search entries can be defined as word root chains. Up to six co-occurences of strings occuring one after the other can be defined, this is extremly useful for the German language.

If strings like 'guter Politiker' or 'gute Finanzpolitik' are to be found by one search entry, a word root chain like 'gut> <politik>' has to be defined. The parameter field must contain a U (ignoring differences in upper-/lowercase).

Word root chains can have three modes: the direct mode, the following mode and the simultaneous mode. In direct mode all word roots must follow each other without any other words between them. The order of the roots must be the same as the one in the text. In following mode the order of the word root is still important, but there may be other words between the word roots. Order and distance between the roots do not matter in simultaneous mode, so this mode is the most powerful and dangerous one. An example shows the differences between the three modes. If the word root is 'gut> <politik>' than in direct mode only text like *guten Politikern* or *gute Finanzpolitik* is found but not part of the text like *guten und schlechten Politikern* or *gute und solide Finanzpolitik.* These will be found in following mode. Texts like Politiker redete gut or die Finanzpolitik wird gut are only found in simultaneous mode.

## 5. Exploration of texts

### 5.1 Indices

Three different forms of indices can be generated with INTEXT:

- the normal form containing the frequency and the word
- a reverse index
- a lemmatised index (German only)
- an index of word combinations
- an index of word permutations

The reverse index can be converted into the normal form and formatted right justified, the right margin is variable. Multiple column printing of the reverse index is possible and dependent on the length of the types.

### 5.2 Word combinations

Word combinations can be performed with two options: the number of words in a combination and the position. The position determines the sort order of the combination and defines whether the first word of a combination is the sort key or the last word. For example, a word combination list of the string *This is a test* with two words with after position looks like this (before sorting):

This is
is a
a test

whereas with two words in before position it looks like this (before sorting):

is This
a is
test a

For better readability these examples are unsorted. With three words in after position the word combination list looks like this:

This is a
is a test

Word combination lists are smaller than indices, and therefore do not need much time to execute. They are based on a text unit.

### 5.3 Word permutations

Word permutation means that every word of a text unit is combined with any other word of the text unit. The result of the example string mentioned above is:

This is
This a
This test
is a
is test
a test

Word permutations are very time- and storage expensive, depending on the length of the text units. Tests show that especially long text units may terminate



execution because 8 MB RAM is not enough. One reason is that not only are the word combinations stored, but also some administration information like the frequency of the word combination.

### 5.4 Cross reference lists

Cross reference lists contain more information than an index. The values of the external variables are output as well as the position of the type within a text unit. The external variables may be separated by dashes, the types can be printed in boldface, in italics or underlined. Also the number of cross references per line can be specified. It is also possible to select all strings interactively, so that only the wanted cross references are generated.

### 5.5 SICs - Search Units in Context

Two very popular applications in the humanities are KWICs and KWOCs. INTEXT's equivalent functions are SIC in short format (similar to KWIC lines) and SIC in long format (similar to KWOC). The reason is that it is possible to specify the length of a line, therefore KWOCs containing more than one line are not necessary any more. The short format looks like an ordinary KWIC--output, but external variables can be suppressed. Printing can be up to 20 cpi, so it is possible to have more than 100 characters per line on a normal sheet of paper. Search entries can be printed in boldface, in italics or underlined. If you use the $T_EX$-interface, the output in landscape mode with very large contexts is easy to generate.

The long format prints the search entry centered in one line, a separation line and after that all occurences.

### 5.6 Index of uncoded types

An index of uncoded types is a useful tool while coding with search entries, especially when using wildcard characters or if search entries are in infix or suffix position. The index is compared with a file containing search entries, the result is an index of uncoded types.

### 5.7 Manipulation of indices

Indices can be searched using search entries. These can be read from a file or entered in an interactive way. Also the reduction of indices with stop-words

stored in a file is possible, as well as frequency or number of characters of the type can be reduction criteria. Converting chores are transforming an index into a file of search entries and transforming a reverse index into normal form.

### 5.8 Comparing indices

When comparing indices, you can have a complete comparision in two formats or just an index of types which don't occur in the first index. The results of the program are statistics about inclusive and exclusive types and tokens and the equivalent Type-Token-Ratios. The types are splitt into words, digits and other.

### 5.9 Special techniques processing huge amounts of texts

Processing huge amounts of texts uses a lot of disk space or a lot of RAM. The versions of INTEXT that can make use of EMS storage can handle nearly every file size: generating an index of a file consisting of 2.6 million words (ca. 20 MB) took 40 minutes on a 386 PC with 25 MHz and 4 MB RAM without using a sort order table. Because the TTR decreases when the amount of text increases, there is practically no limit to the size of a text if an index is to be generated. The index itself is held completly in memory.

INTEXT allows search entries to occur at any position within a string. Therefore it is useful to know which strings are coded and which are not. The first step is to reduce the index using the provided stop-word file. Stop-words may be articles, pronouns, adverbs, conjunctions etc. Stop-word files are provided for English and German. Afterwards both the index of uncoded or coded types can be generated. The stop-word file can be altered or extended, because it is a pure ASCII-file.

Two examples show the speed of INTEXT:

An index from a file of 500 KB using the German sort order table took 120 seconds on an AT-386 with 25 MHz. Without using the sort order table it took just 70 seconds.

A content analysis of a 21 MB file with 247 search entries on the same machine took 7 hours writing an output file of over 5 MB of data.

## 6. Content analysis in the social sciences - style analysis in the humanities

Computer aided content analysis and style analysis need a file of search entries. Every time a search entry is found in the text, the code belonging to the search entry is written to an output file that can be used for raw data input of statistical software. Up to 3000 search entries can be processed at one time, depending



on the length of the search entries. But a lot of problems occur when using this technique.

### 6.1 Problems with ambiguous search entries

In nearly every language search entries can be ambiguous. If ambiguous search entries are coded, the coding may be incorrect and bias the results. Ambigious search entries should first be inspected by generating SIC-lines in short format (similar to KWIC-lines).

### 6.2 Problems with negated search entries

Another problem occurs if search entries are negated. The search entry *attractive* describes a positive characteristic, whereas *unattractive* describes a negative one. If the search entry is *attractive,* both strings with positive and negative meaning are coded, and the negative one should not. Negations, even multiple ones, are detected by INTEXT. Single negation does not allow coding, whereas double negation does. Negation works with a list of indicators stored in a file, indicator files are provided for English and German. The indicators are counted in front of the search entry; the number of words to be searched before the search entry can be specified. The open structure allows the adaptation of the algorithm to other languages.

### 6.3 Labels for categories

Using specific features of coding requires a file where the labels of the codes are stored: the file of category labels. These describe the contents of a category. When the coding mode is switched to interactive, the code and its description are displayed. Also the category labels are used for generating variable labels in the setup for the statistical package.

### 6.4 Controlling the results

The results of the coding can be stored in counters for each category or in the order of occurence. They can be controlled by defining up to three control files. There is one file containing all text units where at least one search entry was found (CODED-file), one file containing all text units where no search entry was found (REST-file) and one file containing all text units where at least one negated search entry was found (NEG-File). In the CODED- and the NEG-file,

the search entries can be printed in boldface, in italics or underscored. The search entries are followed by the code and its responding category label. This output allows a good control of the validity of the coding process.

### 6.5 Interactive coding mode

Using interactive mode ambiguous search entries can be coded while the context - the text unit - is displayed. Ambiguous search entries can be marked in the file of search entries, or all search entries can be treated as ambiguous. Unique search entries are coded automatically, ambiguous search entries are displayed on the screen, together with the external variables, the text unit (or part of it if it exceeds 1320 characters). Search entry, code and category label are emphasized. The code can be accepted, rejected or changed. For pretests all search entries can be treated as ambiguous (regardless of whether specified as ambiguous in the file of search entries or not), thus making the coding process complete interactive and transparent.

## 7. Readability analysis

If the text unit is defined as a sentence, a readability analysis can be performed.

The readability program performs a readability analysis on the basis of different formulas that are based on syntactic criteria of the text. Implications of most of the formulas are that they are language and/or text genre specific, so the results have to be treated carefully. The text unit must be a sentence.

The REFO routine contains eight different formulas for readability analysis. In contrast to the literature mentioned further on, there are no samples of 100 words drawn, but the whole text or the part of it specified by the user will be used. The values of the formulas are between 0 and 100: the higher the value is, the better the readability is. If the values are out of range, it is very likely that the formula is used on texts it has not been developed for. All formulas are based on counting sentences and syllables. The syllable count algorithm is language independent and works with an indicator file; these are provided for English and German. It is also possible to test the counting algorithm in the way that all words, together with the number of syllables that they contain, are written to a file. For English and German the goodness of the algorithm and the indicator files are between 95 and 100 %. The manual describes how indicator files for other languages are to be written and tested.



## 8. Personality analysis

Mittenecker (1952) stated that schizophrenic people differ from mentally healthy ones in that they use repetitions nearly twice as often. The PERSANA program cannot tell whether a schizophrenic or a healthy person is the author of some text, but it helps the researcher by writing the words (or parts of it) that occur more than once in a text unit to an output file. This comparision can be done on whole words and also on parts of words. Every repetition is a case, and so the output file can be processed by statistical software.

## 9. Presenting the results

The results of the desired application are often stored in a file; the application itself displays statistics on the screen and writes it to the log file. The files generated are straight ASCII-files with nearly no control codes (but these control codes and their meanings are explained in the manual). A simple copy to the printer is often unsatisfactory; therefore the result managing program is included that allows viewing and printing the result files. The printing features allow specifying all margins in millimeters, a header, page numbering, page number to start with, position of the page numbering on the paper, paper formats, single sheet feeders, and so on. Viewing is currently done sequentially on the screen, extended VGA-modes (e.g. 80 columns and 50 lines) are supported. So one can view the four columns of an index as well as long SIC-lines with 132 characters per line and 25, 43 or 60 lines per screen. Because these features are dependent on the VGA-adapters, not all text modes mentioned above are available on every VGA-adapter. The automatic installation program for VGA- adapters tests it and writes the result to a file that is used by the Intext Result Manager. If you prefer to use your own text processing system, the output can be written to a file too. If you use $T_EX$ with its macro package $L^AT_EX$, the necessary type setting commands will be generated.