ACADEMY OF SCIENCES OF THE USSR
LENINGRAD RESEARCH COMPUTER CENTER

V.V.Alexandrov, A.V.Arsentyeva

DIALOGUE STRUCTURE
(DIALOGUE - IS IT AN ART OR SCIENCE?)

Part 1

Leningrad
1984

Forword

This paper presents the results of an investigation carried out with the objective of examining the role of dialogue in information interaction.

The investigation has shown that to organize an adequate dialogue purporting to effect common understanding in an easy way both logical and linguistic constituents should be incorporated in dialogue structure.

The logical approach to dialogue construction reflects the tendency of specialization and is characterized by the high rate of information interaction pertinent to restricted functional areas.

The linguistic approach, on the other hand, manifests an aspect of universalization and is distinguished by possessing remarkable inherent features of explanation and adaptability.

To reveal dialogue structure we studied a collection of texts of diverse categories ranging from poetry in original and in translations, through non-fiction works to computer programs including commentaries - each of them representing an example of dialogue in specific form.

The structural analysis was carried out on the basis of application of techniques of pattern recognition and cluster analysis.

To characterize structural patterns of texts we used parameters of functions of Zipf, Mandelbrot and our own approximating ranking distribution of word occurrence frequency in texts.

To facilitate physical interpretation of the observed results, particularly those relating to dialogue construction and development, we conceived parameter α . By means of this parameter, which indicates the rate of structural development, we were able to assess the role of linguistic and logical constituents in dialogue construction.

It is also shown, that in contrast to artificial language texts, the tree structure representation of natural language

texts exhibits presence in them of three clear-cut segments which is the characteristic mark of the linguistic approach.

In the way of mathematical description of various tree-like structures we used recursive approach.

As a conclusion let us remark, that if one wishes to create an adequate user-friendly information interaction system he should take into account the following considerations. The first constituent of the dialogue structure should be aimed at explanation. Consequently, the structure, vocabulary and text generation grammar of this constituent should comply with the structure of linguistic approach. The second constituent of dialogue intended for inner computer program realization should be based on the logical approach.

"Plus il ira, plus l'Art sera
scientifique, de même que la
science deviendra artistique.
Tous deux se rejoindront au
sommet après s'être séparés
à la base".

G.Flaubert*

Introduction.

Advances in computer field brougt about emergence and widespread proliferation of an array of computer types - from all-purpose "super" mainframes to dedicated mini and desk-top personal computers.

Nevertheless, this profusion alone was not sufficient to facilitate exploitation of computer inherent capabilities to full extent. There is a strong evidence of a significant disproportion between the rate it takes an average user to master computer, and the rate by which computer performances are increasing (speed, memory, etc.). This disproportion, hindering efficient use of computer, may be contributed to inherent shortcomings of practised communication methods, which introduce elements of disharmony into a process of man-machine communication.

A lot of effort has been and is being spent lately to

---

* "The further we go the more Art will be scientific and science artistic. They will reunite at the top after having been separated at the base". G.Flaubert, Correspondance, 24 april, 1852.

Большинство диалоговых систем используют в той или иной форме тактику меню, в которой функции диалога определяются заранее, и пользователь таким образом не имеет возможности взять на себя активную роль в построении и модификации диалога. Положение пользователя относительно этой тактики сравнимо с положением человека получающего инструкции о том, как следует реагировать на определенную ситуацию.

Хотя цель диалога и инструкции обычно совпадают, - обучение "правильным" адекватным решениям на реакцию окружающей среды, однако, в повседневной жизни мы обучаемся на естественных диалогах в форме бесед и текстов, несмотря на их явную избыточность по сравнению с инструктивным описанием.

Очевидно, что комфортное, дружеское взаимодействие с ЭВМ должно учитывать специфические особенности построения диалога в процессе общения.

Известны попытки создания проблемно-ориентированных диалогов, похожих на естественные /I,2,3/ , ориентированные на активную роль пользователя. В этом случае пользователь обучает систему.

Чтобы учесть эти специфические особенности нужно прежде исследовать структуру естественного диалога.

Так как любой диалог немыслим без текста, независимо от вида, способа и языка его представления, то анализ структуры диалога мы будем проводить на основе исследования структуры различных текстов.

I. Логическая и лингвистическая составляющая диалога.

Слово "диалог" стало синонимом поиска взаимно-однозначного понимания в различных сферах человеческого общества: политике, искусстве, науке и др. Именно в таком широком смысле мы будем использовать слово диалог, которое включает также заочные диалоги, т.е. научные статьи или художественные тексты, посредством которых автор пытается установить взаимопонимание с читателем. Новый всплеск интереса к исследованиям функции структуры и особенностей ведения (построения) диалога-

find a solution to this problem especially in the fields dealing with natural languages, problem-oriented and expert systems, and as a result several approaches appeared. As a rule these approaches are oriented to specific group of users, and consequently are effective in solving local problems only. Real impact on society the computers may make once the methodology is devised of their widespread proliferation and use as a mass communication tool and information medium.

In our opinion, one of the key reasons of unsatisfactory situation in man-machine communication lies in the fact that too little attention has been paid to the role of dialogue in communication.

Majority of dialogue systems use some or other form of menu techniques in which the dialogue functions are predetermined and thus the user has no opportunity to take active part in dialogue construction and modification. Position of a user in relation to such techniques may be compared to a person getting precise instructions of how to behave in a particular situation.

Although the aims of instruction and dialogue correlate closely, i.e. they both aim to teach us rules of how to act properly in response to environment changes, still we in our everyday life learn to react by resorting to natural language dialogues in the form of loose conversations and texts, regardless of their evident redundance as compared to strict formal instructions.

Therefore, it seems hard not to agree that interaction between man and computer, or more precisely, the organization and structure of dialogue in man-machine communication should be based on a comfortable, easy and informal way human beings interact among themselves.

There are some attempts to produce an illusion of "natu-

rality" in dialogue with problem-oriented systems /1,2,3/ which are oriented to user active part. In this case the user "educate" computer.

To construct a friendly dialogue with computer we need to get a deeper understanding of natural language dialogue structures.

Since any dialogue is impossible without text, regardless of its form, mode and language of presentation, we will start our investigation into nature of dialogue structure by analysing structures of various texts.

1. Logical and linguistic features of a dialogue.

The word "dialogue" seems to have become a synonym in various spheres of society (politics, arts, sciences, etc.) of an act of "searching for a common understanding". In such a broad sense we intend to use the word dialogue in this paper. The word dialogue encompasses also dialogues in correspondence, i,e. texts in scientific papers and literary forms, by means of which an author tries to establish a communication with a reader.

In accordance with an inevitable progress due to history evolution a new wave of interest of investigation of dialogue structure and features of dialogue process came about as a result of the advent of a new tool for storing, analysis and proliferation of knowledge. Emergence of computer networks, personal computers, TV-displays, small-size hard copy units radically changed the manner of communication in human society.

Development of computer-based information systems, special purpose programming languages, interactive technique of communicating with computer introduced greater number of questions than it has solved. These questions are connected first of all with the way an end-user is communicating with computer.

This new turn of thinking in relation to end user reflected also in computer terminology bringing to it from everyday

usage such terms as e.g. "friendly dialogue" /4/.

The first significant scientific contribution to the meaning of dialogue in uncovering a sense and truth and achieving common understanding of participants in conversation may be ascribed to Plato.

In accordance with his teachings we have to turn our attention from formal side of human discourse to a sense which is being conveyed or which is maybe hidden in it . Plato created his dialogues using cyclic way of conversation between persons during which they coordinated their thoughts, corrected representation of thoughts in a language, established common aims and finally grasped common sense.

Soon, it became evident that the words do not convey the exact meaning they were meant to, that they do not transmit truly and faithfully one's image of an object, phenomenon, relation, etc. Is language really the cause of ambiguity ? The above question intrigued man from the ancient times. Brilliant contribution to the subject was made by Leibnitz in his famous work /5/.

It became apparent that the level of ambiguity depended almost exclusively on the skill of the writer or speaker in choosing proper words.

This rather intuitive comprehension of ambiguity problem in dialogue resulted in emergence of two different approaches of dialogue organization: logical and linguistic.

Logical approach to dialogue organization gradually metamorphosed into an axiomatic method of looking for a proof to a preformulated statement (calculus theory). Plato in his explorations came very close to discover the basic laws of formal logic. By saying for example in one of his dialogues that "it is impossible to be and not to be one and the same thing" he, as a matter of fact, formulated the law of contradiction. He also stressed the need to follow this law through all the stages of the thinking process.

The general concept of linguistic approach is based on the

wellknown model of "SENSE-TEXT" sets. Each element of the first set represents text expressed in an artificial semantic language subject to reconstruction and an element of the second set represents text in natural language. It is apparent, that an element of the "SENSE" set should correspond to a subset in "TEXT" set and vice virsa /6/.

Chomsky pointed out that the reason of why traditional grammars (both concrete and universal) fail to formulate adequately the regular processes of construction and interpretation of sentences is a common belief in existence of a natural order of thought which could be mapped in mirror image faithfullness to the appropriate order of words. Therefore, the rules of construction of sentences are really not parts of the grammar and should be looked for in a discipline of science which is studying order of thought.

In Grammaire général et raisonable (Lancelot et al, 1660) it has been said that "the order of words follows "natural order", except in metaphoric sayings,which corresponds to the natural expression of our thoughts" /7/.

The ambiguity, i.e. existence of a "one-to-many" correspondence between the "SENSE" and "TEXT" sets is a prime reason why it is so difficult to translate texts from one language into another, especially poetry. When translating texts from one language into another it is of paramount importance to succeeded in conveying tangible sense of a text which embodies all the concepts of another society possessing another language, national character, a mentality of its own as well as all the underlying associations of that society related to time and space. Otherwise, if the words were only empirical signs, without any socio-historical background that produce complex associative relations, then literature would not differ from algebra, and poetry could not exist.

In response to constant enrichment of society with new experiences language is continuously evolving to meet subsequent knowledge accumulation becoming more abstract, capacious and unambigous so as to be able to express more adequately SENSE by means of TEXT. The process of language evolution is being accompanied by creation of text constructs comprising dialogue means for establishing necessary links between the traditional and new language forms.

In such a way language became a vehicle for expression of ideas, a means of transforming thoughts into texts which could be properly understood by the community they were created for.

Texts in broader sense (painting, sculpture, ballet, etc.), that is, any creation of art and sciences having an impact on society should possess form and incorporate all the richness of past and present experiences of contemporaries, subject to some "rules". Consequently, sense and meaning of a TEXT depends exclusively on the experiences of the individual user.

In the way of our first conclusion let us remark that use of logical approach in dialogue constructions had some measure of success in designing artificial systems characterized by one-to-one correspondence between stimulus (signal) and response to it, and in generating artificial forms of communication. Emergence of computers brought about by inevitable progress in science and technology gave a new meaning to the two forms of dialogue construction, and resulted in two different developments, but this time in computer environment.

The first one is mathematical theory of languages, algorithms, calculus, etc. This development however, not only did not help in bringing end-user closer to the computer, but alienated him from it with the result that the researchers turned their attention to the more promising linguistic approach and developed more natural forms of communicating with the system.

The second one-is linguistic approach to problems of man-machine communication.

It is strange to notice that a number of researchers,

although starting with the same objective in mind of developing new language forms for more natural man-machine communication, after a while turned all their interest to exploring the special field of linguistics proper /8,9/.

Emergence of computers and associated with it the need for man-machine communication should not change anything in the basic role of dialogue which is and always was to establish common understanding. All the principles developed in millenia of man-man communication are still valid and should be applied now to finding an efficient way of communicating with computer.

With this in mind we proceed our investigation into problems of dialogue taking into consideration the following postulate: evolutionary progress of human society worked out specific forms of communication by language and developed ways to efficiently store, represent and proliferate knowledge received as cultural heritage. Thus, language structure becomes a repository of all the variety of human experience, or said in a more general way, it reflects human capability to represent knowledge about their surroundings. This capability according to /5/ is set by inherent characteristics of human mind, properties of perception and principles of human brain functioning.

"In the beginning things were in chaos until the mind made order" said Anaxagoras. "The mind through symbolization created a mode of organizing the various inchaote manifestations of experience in manageable terms ... which gave man, the power to concieve his world, not simply to react towards it, as had the animals ... Hence, he built up a number of symbolic modes which enabled him to find a place in the world and move amongst its multivarious features with a certain degree of confidence ... but all such forms are at once expressive and restrictive" /10/.

It seems the right time now, to try to answer the arbitrary question which might have been posed by an imaginary oppo-

nent: why has the inherent capability of man to make order in his world resulted in such a profusion of knowledge representation forms? In the way of example he might futher elaborated, why are there so many different structural forms of representing numbers. All this abundance of forms is due to the fact that it is always possible to express a sense (an idea) in a number of texts. What is, or should be, though, the common property of all the variants of texts is that they all capture structural ordering of those constituents of perceived sense (idea) which characterize its content. Inherent capability of man to make order in his world enables him to put into effect selection criteria, with the result that in the historical run preserved are only those structural forms which comply to the unmerciful laws of selection. The selection criteria were formulated by Zipf in the most general manner in his principles of least effort: to spend minimum number of words and be understood... /11/.

For example, in the problem of selecting most effective calculus system, that is the one in which to represent N tuples we need $\log N$ signs, following Zipf's principles the priority is given to positional calculus system.

From the point of view of knowledge representation it need not be of decisive significance what earthly or unearthly laws human mind is following in its perpetual ordering and reordering of perceived facts about surrounding reality, always resulting in an output in a compact and easy to reproduce and comprehend form. But what is of importance is the fact that all the products of human activities: social institutions, production plants, scientific institutes, etc. take the structural form of hierarchy.

A primitive system, that is a system without hierarchical relationships is not capable of self-growth and evolution.

To promote growth we need to impose hierarchical relations to the system by means of which we obtain dynamic mechanism capable of generating in compact form new abstract concepts

reducing thereby complexity of description of generic facts.

Let us illustrate this proposition by reffering once again to the foregoing example of unambigous identification of N tuples.

Primitive calculus system operates on the bases of I: I, II, III ...

Number of signs in this case is $Z = N$. It is evident, that on the basis of such numeration it is quite impossible to reduce complexity of description of specified objects.

By using positional calculus system, specified by $Z = \log_i N$ we achieve a qualitative jump from $Z = N$ for $i = 1$ to $Z = \log_i N$, $i > 1$.

Positional calculus system is hierarchical in nature which means that more "compact" and/or less "complex" systems may be generated from it,

This rather trivial example (trivial to-day, but not centuries ago) indicates in an indirect way that between any two calculus systems may exist isomorphism.

Let us examine now the role of texts, sentences and words in dialogue construction.

In the chain "$SENSE_1$ - $TEXT_1$ - $TEXT_2$ - ... - $TEXT_k$ - RESULT - $SENSE_2$", the operators SENSE-TEXT represent query action of the end-user and TEXT-RESULT - operation and response of computer. Computer text operator in TEXT-RESULT shall be deemed equivalent to end-user text operator in SENSE-TEXT if the end-user thinking activity proceeds in the same manner as it used to without computer.

Really, no author ever dared to interrogate his readers as to do they understand anything of what he wrote. This is usually prerogative and privilege of sociologists and teachers. Anyway, the sense which reader percieved from reading a book should, in ideal case, correspond to the ideas put by the author in the text of his book.

Now it seems natural that the requirement for TEXT-RESULT isomorphism brought to light the concept of algorithm - an ar-

tificial construct comprising texts in the form of instructions which specify in a strict and unambigous way the sequence and manner of execution of a task.

So, what shall be the dialogue structure like to both embed effectively SENSE into TEXT and generate RESULT equivalent to SENSE ?

If the role of TEXT is to describe (represent) an idea (SENSE)and convey this idea to other members of community, then the role of words is to identify objects and activities performed on the objects, and of sentences to create generic elementary relations between the words (elementary structural entities).

Let us formulate basic proposition regarding structural forms of the "SENSE - TEXT - RESULT - SENSE" description chain.

1. As a result of establishing cause/effect relationships between the objects of the domain under investigation complex structure is imposed on the system.

2. Systems characterized by complex structural relationships should be represented by a set of hierarchical subsystems.

3. Any complex hierarchical structure may be approximated by a self-similar recursive structure (e.g. positional calculus systems, binary trees etc.) in which way complexity of their algorithmic description is reduced.

4. To properly map one structure of any complexity to another the properties of homomorphism should cover besides the object themselves the relationships of the lower order as well.

5. The laws of Zipf, Mandelbrot and their modifications used in analysis of linguistic features of text structures provide us with an effective tool in perception of developing events by controlled infusion of words, concepts and definitions at comfortable speed into description process /12/.

From the last proposition follows that e.g. textbooks are subjected to much slower growth rate of new definitions, concepts etc. than e.g. specialized scientific works.

In other words, any generally applicable texts oriented to a wide variety of users have a much slower structural growth rate than the texts oriented to a narrow circle of specialists.

The investigation into the principles of structural growth of various texts (by the way, the laws of Zipf and Mandelbrot apply to only one aspect of structural growth) should provide us with a facility to not just analyse the texts, but also to create specific problem-oriented texts and modify them of necessity to comply with the requirements of individual applications.

We believe, that by specifying objective and pertinent characteristics of the text construction, approved by its recipients as to be the most effective, and by analysing trends of change in these characteristics in relation to this or other function of the text under investigation, it will become possible to formulate, on the basis of these characteristics, the criteria for evaluation of languages for man-machine communication.

Moreover, the analysis of a representative collection of texts ought to give us a clearer picture of e.g. what shall be the optimum volume of vocabulary, what is the role of special terminology within the vocabulary, what is the meaning of different aspects in literary text translations and of other characteristics of language pertinent to dialogue. In short, the aim of our study was to find out the common properties of textual description and the correlation between form of language in description and the conceptual content which it embraces.

At present, many specialists in various scientific research areas, from artificial intelligence to literary arts are looking for answers to the above mentioned problems with the same objective in mind of how to organize in a most effective way the communication, i.e. information exchange among people in society in general and in a computer environment

in particular.

To organize adequate dialogue in man-machine systems we should take into account all five of the abovementioned propositions regarding its structure. As first four propositions were already discussed in a number of publications /13,14,15/ we shall examine in more detail on examples of different texts the fifth proposition only, which embodies the two following concepts:

1. The process of text formulation in any language should respect all the variety of structural aspects of the way humans perceive, represent and disseminate knowledge.

2. Zipf's principle of least effort reflects man's perennial search for ideal knowledge representation forms, that is those forms which combine optimum speed of dialogue construction with potential to express complex concepts.

The problem just outlined is pretty much the same as the one of choosing between universal language (i.e. the language in the broadest sense of the word) and the special language (prescriptive language or language of instructions).

To analyse different samples of texts we made use of the techniques from the methodology arsenal of various fields, such as cluster analysis, pattern recognition and problem solving. To evaluate and compare different text samples we employed coefficiants of the corresponding function approximating distribution frequency of words in a given text.

Further on we present an explanation for choosing this or other type of texts. However, in selecting poetic texts we followed our intuition, so this part of text collection reflects our subjective tastes and biases.

II. Structure, vocabulary and "naturalness" of languages.

Really, in searching for an adequate user-friendly computer language we cannot overlook an evident candidate - natural language. Let us set aside for a while the question of adequacy of such a solution and concentrate on the point of what we have in mind when we use expression "natural language". Any language is characterized by two basic parameters, the vocabulary and the structural organization of words in texts. In his works Zipf /11,16/ devoted much of his attention to the investigation of text vocabulary and of laws governing process of knowledge evolution. He found out that the growth rate of human knowledge follows specific law (fig.1). He also noticed that at some point in the process of knowledge accumulation starts division of total knowledge into a series of problem-oriented knowledge /17/.

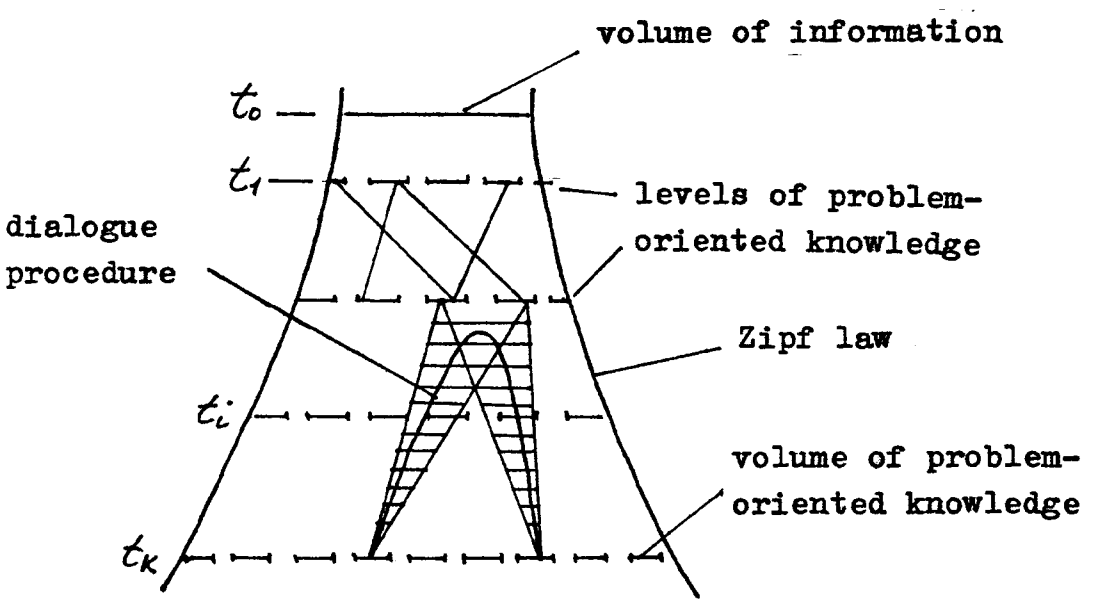

Fig.1. Process of knowledge evolution.

In different realms of human endeavoring, for example within boundaries of an institution, organization or club we are witnessing an emergence and evolution of a jargon, a limited language of specialists, that suits them and satisfies their needs of communication.

In usage, and in this case it means solving of specific problems and communicating about them, the language of specialists evolves in the way all languages evolve.

Specifity of a problem-oriented knowledge is manifested first by the form in which the knowledge of the particular subject area is being represented; the illustration of this are e.g. formulae in mathematics, expedition diaries in geology, experimental data tables in physics, questionnaires in sociology, etc., and second by thesaurus of its own. In a life of a dynamic knowledge system its thesaurus may grow and decrease in volume and shift in information space. The first characteristic of the problem-oriented knowledge contributed to the development of logical approach and is in itself a result of striving to use the most effective formal-logic based apparatus for processing of knowledge of a given subject area. The second characteristic may be related to the linguistic approach, reflecting the need for semantic processing and is of a pragmatic nature. Both characteristics have at their root the urge to speed up information processing which is prerequisite of increasing an efficiancy of the investigation and communication process in a given subject area. That is, the emergence and evolution of a variety of knowledge representation forms, embodied in a concept of problem-oriented knowledge, may be contributed to be an outcome of incessant efforts to speed up interaction within concrete subject areas.

It should be mentioned that fig.1 represents Zipf's law in a rather simplified way. In fact, elements of one level are interacting with elements of another level forming in this way a number of relationships of various character and establishing contacts among different subject areas and hierar-

chies of knowledge.In contrast to the lower level elements, the elements of higher level possess integrative properties which enable them to enter into relationships with any element of the same or lower order exercising in this way the principle of vertical and horizontal ordering.

Scientific activities depend to a great extent on the development of problem-oriented knowledge. However, to solve problems we often have to refer to knowledges of neighbouring areas. To use that knowledge we have to coordinate different concepts which are presented in a variety of specific forms. And to coordinate different representation forms of problem-oriented knowledges and their thesauri we have to refer to the higher levels of knowledge representation (fig.1). At the higher level the concepts of the subject area under investigation and of its neighbouring areas, to which we are referring to for assistance and analogy, have common semantics with a result that a dialogue of specialists is becoming possible and exchange and understanding of concepts from different subject areasis taking place. In such a way the higher level knowledge has a status of an "interpreter" of different representation forms of knowledge from the lower, more developed levels.

In relation to "specialized languages" of the lower levels the high level language may be considered to be a kind of "natural language" to which a role is assigned of ensuring coordination and cooperation among them. Based on such an interpretation of natural language we may refer to its vocabulary as to a sort of abstraction. In fact, in real life we always use some subset of the natural language vocabulary.

The data given in /18/ on the structure of natural vocabulary usage demonstrate in a clear fashion this division into the "general" and a number of "specific" vocabularies.

An examination of structural relationship of words in texts in a given language might be quite helpful in dialogue construction, but this topic has not recieved adequate consideration from the computer specialists. Imparting conventio-

nal descriptions based on grammatical rules to the computer have shown to be an inappropriate method of man-machine communication. At best, conventional descriptions may find their use in analysis of existing texts. In their works, J.Estoup and E.Condon /19/ pointed to the presence in meaningful texts of a definite structure which may be used as a criteria indicator in quantitative evaluation of texts.Later Zipf analysing frequency distribution of words in meaningful texts found out that this distribution can be approximated by some function.

This reasoning may be taken as a confirmation that at the root of our intuitive recognizing of harmony and "beauty" of some texts, of their intrinsic property to be comprehended more easily than others, lies some fundamental law, which although we may not understand completely at the moment, but which nevertheless manifests itself in a definite frequency distribution of words.

Any designer of a computer-based dialogue system is in fact a creator of a language. Hence, he should choose an adequate vocabulary and organize a dialogue on the basis of it.

Computer-based information systems are a fairly new development still without tradition and widely accepted methodology. But it is encouraging to witness penetration of this field with general trend "to humanize" technical systems. There is a recognized tendency to organize information interaction in such a way that participants may communicate among themselves in a most natural way of everyday discourse assisted by computer-based information system. In this way the information system may develop into a genuine user-friendly tool facilitating communication. Nevertheless, that does not mean that user should be competely ignorant of what goes on with his information once it entered into and is being processed by information system. He should certainly be aware of all the modifications his information is subjected to.

At the foundation of a sound scientific methodology should lie principles aimed at achieving common understan-

ding and effective communication among people in a given subject area. We should also take into consideration the following factors. First, that behind all human activities lies motivation to reach an objective and second, that information system is to function within boundaries of a given subject area,which implies its close ties to the specific manner of knowledge representation in the concrete subject area.

In the process of a dialogue its participants are actively reconstructing and evaluating messages. The messages are usually interpreted on the basis of prejudices and previous knowledge, that is, on the basis of expectations, concepts, meanings, knowledge of languages, etc. The aim of any dialogue is to achieve understanding between its participants. But if no ground for understanding existed prior to act of communication the chances are that subsequent dialogue will not result in common understanding of its participants. Common understanding presupposes knowing the language rules and concepts of a given subject area. Therefore, to make appropriate use of computer as an intermediary in dialogue we have to expand communication language by supplementing it with rules and words which could be "understood" by computer.

In this way we obtain two linguistic systems, one that consists of vocabulary and based on it rules for generating texts convenient to be communicated to computer and the other consisting of vocabulary and rules pertinent to specific subject area and existing regardless of computer. If we wish to use computer effectively we need to integrate both systems, which implies that we have to combine logical and linguistic approach to knowledge representation.

But how to combine these approaches in the most effective way so as to achieve adequate man-machine communication, or put differently how to organize dialogue of specialists with computer as an intermediary so that specialists need not change much of the way they communicate in everyday life?

The present state of affair in linguistics and computer

science does not allow us to propose a definite answer to this question. Majority of the existing dialogue systems are based either on an intuitive effort to reduce vocabulary with the result that rigid structure of text generation is achieved or on the concept of providing large vocabularies which give one an illusion of "naturalness" on account of using a lot of synonyms.

As we stated, the aim of our work was not to provide definite answer to the problem of dialogue construction. Nevertheless, the results of experiments we carried out on collection of texts confirmed existence of definite regularities in natural language texts which could be expressed by quantitative parameters. That gave us a means to evaluate various "artificial languages" and compare their characteristics with those of the natural language. By the way, the term "artificial language" seems not to be quite adequate. On account of unsettled terminology we use it as an antonym to the natural language and include into its class all those languages which are characterized by strict rules of text generation. In the way of example we may offer mathematical formulations, descriptions of chemical reactions, instructions of computer programs, which are related by the common characteristics of logical approach prevalence and belong to the group of formalized problem-oriented and programming languages.

In performing experiments on texts of various types we aimed at bringing out the underlying structural and vocabular pattern of an adequate dialogue construction. For this purpose we used texts set in different languages, such as language of poetry, standard programming languages and specialized problem-oriented languages of various subject areas. We tried to use only "good" texts, that is the texts proven in usage to be easily comprehended by man.

III. Collection of texts for analysis.

Since the primary objective of our study was the quantitative analysis of structural aspects of various texts and their comparison and identification we tried to make our collection as diverse as possible.As a consequence we covered various categories and genres ranging from natural language texts of literary arts and nonfiction alike to texts written in different artificial languages.

Enumerated text collection with relevant characteristics is presented in table I.

Table 1.

Enumeration and parameters of investigated texts.

N	Author, text title	Text size (words)	Vocabulary (words)
1	2	3	4
1	А.С.Пушкин. Моя Родословная. (A.S.Pushkin, My genealogy)	365	222
2	А.С.Пушкин. Езерский. (A.S.Pushkin, Ezersky)	920	540
3	А.С.Пушкин. Медный Всадник. (A.S.Pushkin,The Bronze Horseman)	2049	965
4	А.С.Пушкин. Домик в Коломне. (A.S.Pushkin, A hut in Kolomna)	1763	811
5	W.Shakespeare.Sonet 66	91	72
6	W.Shakespeare.Sonet 66,tr.by Marshak	70	54
7	W.Shakespeare.Sonet 66,tr.by Benediktov	77	71
8	W.Shakespeare.Sonet 66, tr. by Pasternak	79	61
9	R.Burns. Is there for all that ...	269	105

1	2	3	3	4
10	R.Burns. Is there for all that ... tr. by Marshak	200		99
11	L.Carroll. Jabberwocky	165		89
12	L.Carroll. Jabberwocky, tr. by Orlovskaya	108		71
13	L.Carroll. Jabberwocky, tr,by Shchepkina-Kupernik	147		101
14	P.Verlaine.Chanson d'Automne	50		41
15	P.Verlaine.Chanson d'Automne, tr. by Sologub	40		37
16	P.Verlaine.Chanson d'Automne, tr. by Geleskul	40		36
17	P.Verlaine.Il pleure dans mon coeur...	80		49
18	P.Verlaine.Il pleure dans mon coeur ..., tr.by Geleskul	55		43
19	P.Verlaine.Il pleure dans mon coeur..., tr. by Ėrenburg	54		28
20	P.Verlaine.Il pleure dans mon coeur,.., tr. by Pasternak	52		38
21	Dialogue with IS "Economy"	991		185
22	Dialogue with IS "DIANA"	1039		311
23	Dialogue with IS "SITO"	2485		578
24	Fortran. Procedure "Mapon"	576		77
25	Algol. IS "SITO" (fragment)	3521		148
26	Basic. IS "DIANA", without commentaries	10247		924
27	Commentaries to IS "DIANA"	1105		253
28	Basic. IS "DIANA" with commentaries	11329		1182
29	Frequency dictionary of Russian language	1056982		39268

1	2	3	4
30	Ternary tree	1216	242
31	Binary tree	2049	256
32	Fibbonacci tree	2421	233
33	Lucas tree	3488	272
34	$f(n)=n-f(f(f(f(n-1))))$	3594	250

As one can see the collection comprises texts from poetry too, which we included with an aim of examining possible relationship between the distinctive feature of condensed form inherent to poetry and the quantitative parameters that may describe pattern underlying its structure.

We also included texts in different natural languages - English, French and Russian with an aim of proving our claim purporting that our methodology of investigation is applicable to any given natural language regardless of its belonging to specific group or category. Or put differently, we tried to confirm a conjecture that structural characteristics of language texts do not reflect innate properties of language but rather universal property of human mind to comprehend information in abstract form. This property may be accounted for the evolution of great variety of natural languages significantly differing in structural form, but all with the same characteristics of being easly comprehended.

It was deemed useful to include Russian translations of some literary texts and compare their structural characteristics with those of originals in French and English. We used not only translations of high literary merit by remarkable translators of outstanding personality, but also translations characterized by almost literal rendering of original. This made it possible for us to draw certain conclusions as to the feasibility of our technique to assess the degree to which some

translation may have approached original.

The chosen texts of literary works (including translations) belong to various epochs, genres and schools but all are generally acclaimed to be masterpieces in their own class, which implies that they are proved to be easily comprehended and that they possess great ease and force in presenting themselves to reader.

As an example of nonliterary natural language text we included into analysis the Frequency Dictionary of Russian Language.

The collection comprises also four programs written in different programming languages. One of the programs we analysed in two versions - with and without appropriate commentaries. The texts of programs differ in size, purpose (computation or information processing) and language of programming.

To illustrate use of problem-oriented formal languages* we included one text of the actual dialogue of a medical expert with computer-based dialogue system for diagnosis of sicknesses of abdominal cavity organs (DIANA).

We also compared vocabularies and structural organization of the natural language literary texts with those of the texts of dialogues in problem-oriented languages. The structural organization was expressed by means of quantitative parameters of functions approximating frequency distribution of words in texts.

Text written in programming languages we investigated with two objectives on mind. Our first goal was to compare

---

* Sometimes, the languages of this class are colled restricted natural languages. We do not think it is an adequate expression since we always use some kind of restricted natural language communicating within our specific knowledge domain.

their quantitative parameters with the parameters of natural texts and at the same time we wished to elucidate the role of commentaries in comprehension of texts. With this aim we analyzed texts written in three different programming languages.

Secondly, we wanted to investigate and quantify a possible correlation between the structural organization of a given program text and the dialogue generated by this program.

As one can see from fig.2 the vocabularies of program text and text of the dialogue are intersecting each other.

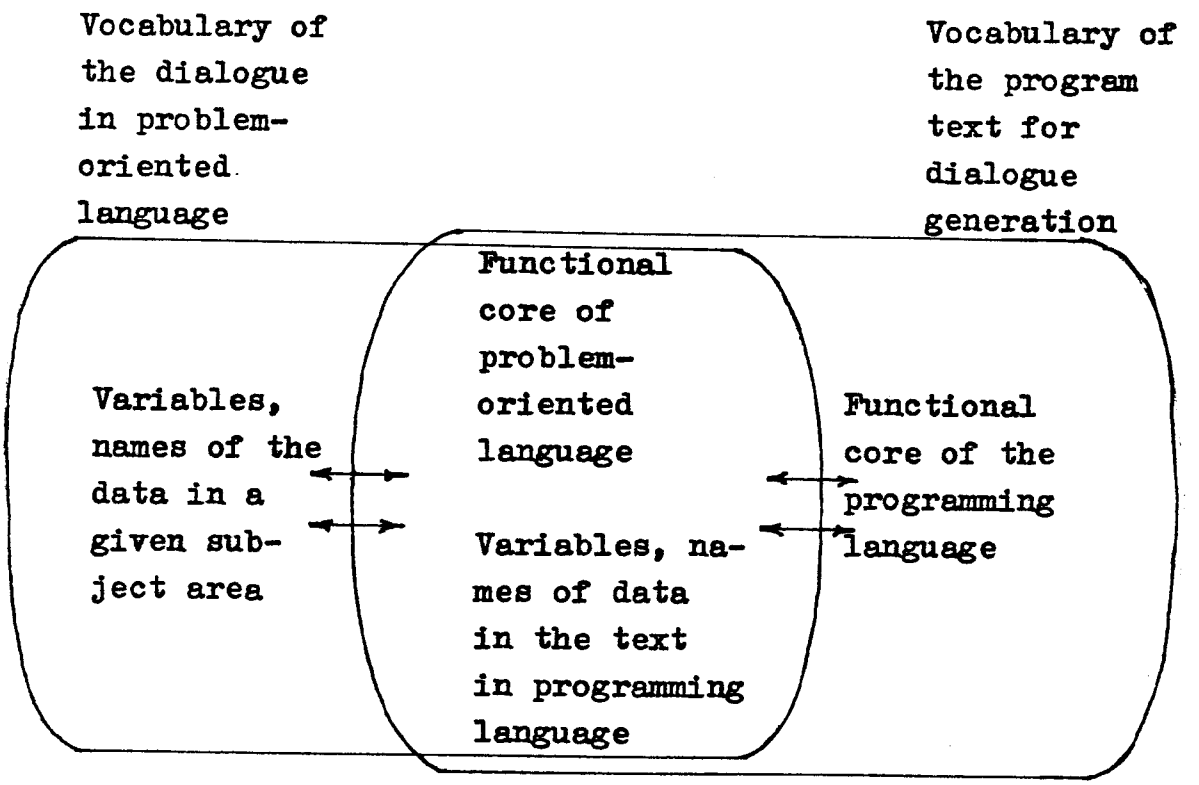

Fig.2. Correlation between the vocabularies of the programming language and the dialogue language.

It is instructive to note that the part of vocabulary of a given subject area comprising its functional core (we use this expression to refer to that part of the vocabulary which is found to be present in all dialogues led in the language of particular subject area) represents at the same time the variable part of the vocabulary of a program written in programming language.

Let us consider now main features of communication in a computer based dialogue system. An expert well versed in his subject matter, let's call him A, developed formalized language of his specific area of endeavor and specified all the relevant, and from his point of view indispensable relationships among words selected from the vocabulary of his knowledge domain. In this way he is actually focussing his attention to the linguistic part of dialogue. His colleague, let's call him B would like to use knowledge of A and would like to do in direct communication. But he is albeit compelled to use an intermediary in the form of computer.

To examine this problem in more detail we analysed operation of the actual computer-based dialogue system of medical diagnosis. In this system all the relationships among symptoms, syndroms, etc. as well as the decisions concerning therapy definition are first specified in medical jargon (subset of natural language) and then this description (in particular, that part of description which concerns logical constituent of the dialogue) is being transformed into the language of predicate logic (artificial language).

The question may be put forth now of what type of language shall we use to organize dialogue between A and B by means of computer based dialogue system and in the process distort as little as possible of the original meaning so as not to impair understanding. Or put more bluntly, does a process of dialogue depend on the type of programming language after all ?

At a first glance, there seems to be no direct connection between the language of dialogue and programming language.

For example, more important appears to be the presence in programming languages of facility to manipulate strings. However, the question posed is not that simple.

Current trend in computer field is application programming without programmers, which is manifested in development of very high level languages characterized by a feature that their functional core is present in them by implication only. In ideal case of computer program statements coinciding with concepts from the application-specific area the user would be working under impression that he is actually communicating in natural language (of course, not natural language but its subset pertinent to his subject area). There is certain analogy between the situation just described and the case of someone reading a book and trying all the time to associate and replace each word of the book with concepts of his own mind. Sometimes the process may have a deeper recursive implication.

Using programming languages of this class we automatically achieve one-to-one correspondence between the vocabularies of the dialogue generator and the generated dialogue which means that their vocabularies have common alphabet.

In real life however, this extreme case would neither be feasible nor possible since due to redundance the number of procedures would greatly increase, their subsequent storing would require excessive number of standard libraries and consequently access to the needed procedure would become rather cumbersome and time consuming.

Therefore, to achieve comfortable, user-friendly dialogue we should first try to balance the size and content of the vocabularies.

The methodology we employed in our exploratory investigation may hopefully be of possible use in supplying some quantitative indicators and bring some more light to different aspects of dialogue organization.

Besides texts enumerated above our text collection includes five items (N° 30 - N° 34) presenting "texts" arti-

ficially synthesized out of hierarchical structures which will be described more in detail in part 2.

IV. Quantitative investigation of language texts

The phenomenon of frequency of word occurrence in language texts is a matter that has recieved considerations from many researchers. Philology for example treats text as a language form through which an author by means of techniques characteristic of his epoch and culture expresses a system of images, concepts, views, feelings, beliefs ... - integral to the degree he mastered his art.

Telephone book, dictionary, random selection are examples of texts which are not semantically meaningful, i.e. they do not possess a definite conceptual content (though they may possess a definite pragmatic sense and meaning).

On the contrary, poetry texts, and that is the reason we included them in our collection for analysis, are immensely saturated with concepts and meanings and are expressive of semantic content.

It is customary to decompose texts into items of different detail and level such as paragraphs, sentences, phrases, words, morphemes, phonemes, etc. In our analysis we used decomposition of texts into lexical units in the sense of Zipf, that is two words are treated as different if they are distinguished in a dictionary.

One of the most important principles in linguistics is that the individual text items may be classified according to frequency of their occuring in usage. Widely used technique of specifying frequency spectrum of word occurrence in texts is the method of ranked distribution which treats frequency of text item occurrence f_r as a function of rank r comprising text items occurring with just this frequency.

Ranked distribution is usually approximated by an empi-

rical formula. In case of text transformation the initial empirical formula is found to be still valid for new versions of text within the order of magnitude of parameters, but the frequencies of occurrences of individual text items may differ significantly.

The basic tenet of linguistics is that semantically significant texts are characterized by possessing definite structure which is manifested in a phenomenon of regular frequency distribution of text item occurrence. This regularity is observed in semantically significant texts, that is texts embodying conceptual content. This empirical law is known as Zipf's law and can be expressed in the most general way as follows.

Consider a collection of Z objects. Each of the objects is identified by an attribute selected from a set. Let $V(f, Z)$ be a number of different attributes, each of which is used f times to identify different objects in collection of Z objects. Then for sufficiently large Z we obtain the following relation:

$$V(f, Z) = \frac{A}{f^{\gamma}} \qquad (1)$$

where A is a constant depending in principle on size of collection Z, $\gamma = 1 + \alpha$ is exponent of the Zipf's law and α is characteristic value (constant).

If all the attributes are ranked in descending order of their occurrence then the value r representing attribute location in this series is called a rank.

Ranked representation of Zipf's law is of the form:

$$fr = \frac{C}{r^{\beta}} \qquad (2)$$

where $\beta = 1/\alpha$, $C = \alpha B^{1/\alpha}$.

Sometimes, the following formula proposed by B. Mandelbrot is used to approximate actual frequency distribution:

$$fr = \frac{C}{(B+r)^{\beta}} \qquad (3)$$

where C, B and β are constants.

Zipf's law has been tested on wide variety of applications from diverse realms of human endeavoring and demonstrated its validity. There is a lot of papers on this subject comprising corroborations of Zipf's law by such prominent statisticians as Yule, Kendall, Lotka and others.

To bring out structural patterns of vocabulary corpus of our text collection we had to estimate quantitative parameters of functions approximating ranked frequency distribution of word occurrence in text. But first we needed the functions, the curves of which would fit as faithfully as possible the plotted calculated data. The functions should be simple enough and comply with the formulae of Zipf and Zipf-Mandelbrot.

The poetry texts from our collection showed an interesting feature which manifested itself in the ranked frequency distribution which could not be adequately approximated by laws of Zipf and Zipf-Mandelbrot over the whole range of distribution. This deviation may be attributed to the restricting applicability of Zipf's law to texts having size less than so called Zipf's size.

The process of analysing various texts from the collection included varying numeric values of parameters of approximating functions (2) and (3). These functions were supplemented arbitrarily with simple rectangular hyperbola of a type $xy = K$, which gave us a facility to assess to what degree the approximation need be true if we wish to obtain subtle differentiation of text types.

To fit the approximation function to the set of points of calculated ranked word frequency distribution plotted in logarithmic coordinates it was first necessary to estimate the parameters of functions and then to check their adequacy at several characteristic points. To improve effectiveness of approximation we additionally varied parameter K.

In the way of approximating function we considered both

wings of rectangular hyperbola $(x-a)(y-b)=K$ /20/. shifting one of the wings in the coordinate plane. The value of parameter $K = 50$ gave in the best fit of approximation curve to the set of points of ranked frequency distribution of word corpus of poetry texts.

To check effectiveness of approximation we chose two characteristic points, which can be seen on the fig.3 . The characteristic points were defined as follows: first point $r=1$ abscissa $\ln r_i = 0$, ordinate $\ln \frac{F_i}{Z} = \ln \frac{F_1}{Z}$, second point corresponding to rank $r = r_{max}$ with coordinates: abscissa $\ln r_i = \ln r_{max}$ ordinate $\ln \frac{F_i}{Z} = \ln \frac{1}{Z}$ since $F_{min} = 1$.

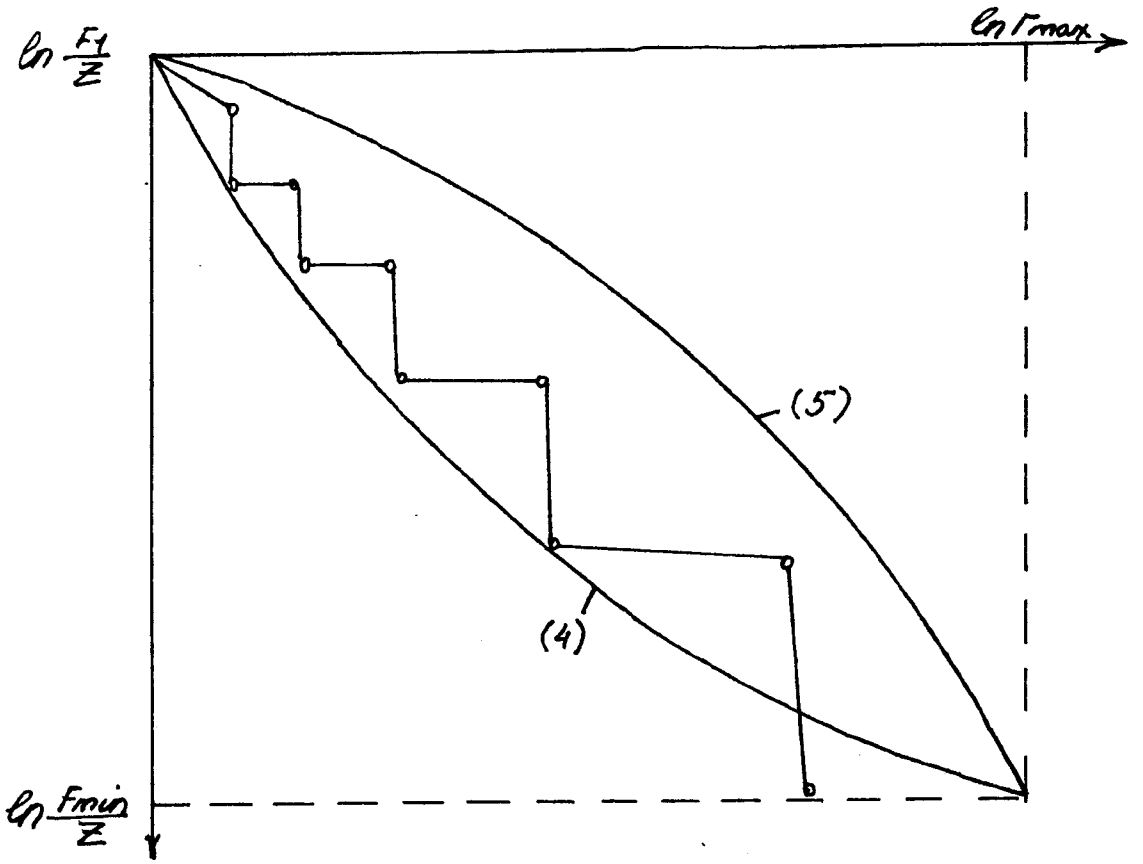

Fig.3. Experimental curve and its approximation by (4) and (5).

Parameters of approximating functions

$$\ln \frac{F_i}{Z} = a + \frac{50}{\ln r_i - b} \tag{4}$$

and

$$\ln \frac{F_i}{Z} = \frac{50}{\ln r_i - b} - a \tag{5}$$

were estimated so that approximation function curve should pass through the calculated characteristic points. For the approximation curve under consideration and given coordinates we have:

$$(x-a)(y-b) = 50$$

and two characteristic points are

$$(x_1, y_1): x_1 = 0,\ y_1 < 0;\quad (x_2, y_2): x_2 > 0,\ y_2 < 0.$$

In this case

$$b_{1,2} = \frac{x_2}{2} \mp \sqrt{\frac{x_2^2}{4} + \frac{100\, x_2}{y_1 - y_2}} \tag{7}$$

and

$$a_{1,2} = \frac{50}{b_{1,2}} + y_1 \tag{8}$$

The set of calculated pairs of parameters derived from the laws of Zipf (2) and Mandelbrot (3) and from the relations (4) and (5) was used as source of data for computer processing. Table comprising experimental data is included in Appendix I. Enumeration of the 34 horizontal lines corresponds to the enumeration of texts in table 1 and numbers in vertical columns apply to the following 11 characteristics:

1. value of $\ln F_1$ (number of word occurrence of rank 1)
2. value of $\ln Z$ (text volume in words)
3. parameter C of function (2)
4. parameter β of function (2)
5. parameter K of function (3)
6. parameter B of function (3)

7. parameter a_1 of function (4)
8. parameter b_1 of function (4)
9. parameter a_2 of function (5)
10. parameter b_2 of function (5)
11. value of $\ln V$ (vocabulary volume in words)

Y. Structural methods of data analysis

At present, there exists a wide variety of application software packages available for use in experimental data analysis but they as a rule comprise statistical methods in the narrow sense of the word. Procedures of pattern recognition and cluster analysis, if included in the package at all, do not form an integral part of the experimental data analysis system.

On the other hand the experimental data tables represent rather restricted mass of data to be analysed and processed on the basis of statistical methods only.

The credibility of the results of processing such experimental data tables may be enchanced to a significant extent by applying approximate methodology of matching the outputs of various programs which were chosen on account of their ability to process source data aggregate in the process of analysis of the investigated area.

Based on such considerations the following requirements regarding experimental data analysis system may by formulated: interactive facilities for source data entry, flexibility of generations and modification of experimental data tables, possibility of program sequence adaptation.

From the available collection of application software packages we chose interactive system SITO (SITO is the Russian word meaning a sieve or to sift) developed in 1980 in Leningrad Research Computer Centre of the USSR Academy of Sciences for use on Cyber 172/6, not only because one of the authors of this study happens to be its developer and the various

practical applications have demonstrated its effectiveness, but rather on account of this system capability to comply in the best of manner to all the requirements imposed on it. And just the features of flexibility in choosing of investigation direction helped to integrate this system into complex strategy (methodology) of experimental data processing which eventually reflected in the authenticity of the results.

The methodology of investigation included the following procedures of source data processing:

1. Creating local data base and storing experimental data tables into it,

2. Determining one-dimensional distribution selected for analysis of parameters over the whole range of objects.

3. Determining principal components.

4. Printout of projections of the totality of objects to the plane, in the order of sequence of the parameters and first two eigen values (1-2, 1-11, 2-11, 3-4, 5-6, 7-8, 9-10 and $\lambda_1 - \lambda_2$).

5. Computing parameter clustering using interrelationship of observed data as a measure of similarity.

6. Applying hierarchical cluster procedure to determine in a stepwise manner the number and formation of classes.

7. Computation of the increment function of interclass distance.

Before we enter upon examination and evaluation of the results of data processing (Appendix) let us, in accordance with previously formulated propositions and conjectures about dialogue organization, make clear once again what we expect to achieve by classifying texts on the basis of the parameters depicting patterns underlying their structures. To this end we shall postulate several hypotheses relying directly on the actual objective grouping of the texts in Table I (selection of the individual texts may be subjective, but collection as a whole has been made to comply with the objectives of the investigation).

The hypotheses we present in the order of increasing

degree of text differentiation:

Hypothesis I.

Totality of 34 texts may be divided into three classes:

- texts 1 through 20 - natural language texts
- texts 21 through 28 - artificial language texts
- text 29 - dictionary, an example of natural language non-fiction text
- texts 30 through 34 - group of tree - structure texts.

Hypothesis II.

The processing divulged texts 14-20 (originals and translations of Verlaine) as a separate distinctive group within the class of natural language texts. The rest of classification remained as presumed.

Hypothesis III.

In the class of the natural language texts it was possible to distinguish those texts of translations which were structurally close to the originals. In the class of artificial language texts it was possible to separate the texts of dialogues from the texts of programs.

Hypothesis IY.

In the process of classification no grouping of texts was observed in respect to the language they were written in.

Hypothesis Y.

The text of Pushkin poem (N° 3) has structural peculiarity in comparison with other natural texts. We could not resist a temptation to check if the classification procedure would reveal any structural specifity of the enigmatic poem "Bronze Horseman" by Pushkin in relation to which a number of researchers /21-24/ are trying to prove the presence of the second conceptual layer in it.

YI. Analysis of the results of data processing

In Appendix 2 listed are in sequential order the projections of all 34 texts to the plane, coordinates of which are parameters 1-2, 1-11, 2-11, 3-4, 5-6, 7-8, 9-10, $\lambda_1 - \lambda_2$, where λ_1, λ_2 stands for the components obtained by the method of principle components for the experimental data table. Visual interpretation of the interrelationships among the totality of texts in the $\lambda_1 - \lambda_2$ plane allow us to make a classification in accordance with the hypothesis proposed. Validity of this conclusion was confirmed by concurrence of the overall patterns of the text distribution in the planes (1-2) and (9-10) respectively, in spite of different position of individual texts in these planes.

On the account of the visual method of classification being of a rather subjective nature we will refrain from drawing final conclusions until we examine the program outputs of the hierarchical cluster analysis procedures.

In Appendix 5 presented are in a sequential order the results of classification obtained on the basis of three different methods of determining distance among the objects.

The results obtained by using the totality of all 11 parameters are in accord with the hypothesis. In addition, the specific structural patterns of various texts were revealed in an unmistakable clear-cut fashion. For example all the Pushkin texts we observed (and they, according to /22/ form a unified cycle) show in refined decomposition a tendency to form separate cluster of their own, which at rough classification breaks down into texts 1,2,4 and blends into class of natural language texts. Text N° 3 however persistently mingles into the class of artificial language texts by which it manifested its particular structural pattern.

The classification revealed the following specific features of the texts of translations:

- text of original 5 and its translation texts fall into the class of natural language texts, but at detailed classi-

fication they cluster into the separate stable group of their own; hence, for these texts may be said that they have similar structural characteristics, Text 7 however, forms separate cluster of his own in spite of being translation of the same original.

- original texts 14 and 17 and their translations 15, 16, 18, 19, 20 form so called Verlaine class, excluding translation texts 19 and 20 which merge into the general class of natural language texts. That is, the hypothesis of existence of separate Verlaine class based on his unorthodox use of words was unequivocally supported by evidence of distinct structural pattern of all the Verlaine texts.

We have no intention of imputing that some translations are more adequate than others.However, the results of classification allow us to observe concordance of structural patterns of some translations with their originals. So, for example, the translation text 13 of the original text 11 judging by the close structural likeness may be included into Verlaine class.

Classification results also confirmed, in an unmistakable fashion, the hypothesis of structural homogeneity of artificial language texts. At more detailed classification the artificial language texts broke down into two classes: dialogue texts 21, 22, 23, 27 and texts of computer programs 25, 26, 28.

As can be seen in Appendix 2 and 3 analysis revealed no evidence that would suggest presence of specific structure in the texts which were written in different natural languages (Russian, English, French). So the hypothesis that structure of the text does not depend on the language it is written in was confirmed beyond doubt.

And to sum up, the results of applying hierarchical cluster analysis procedures allow us to draw sufficiently detailed conclusions as to the structural pattern of each

individual text. These conclusions are, however, reached on the basis of using totality of all 11 parameters, which complicates visual interpretation of obtained results. Thus, the question remained open of what is the actual functional interdependence among the parameters, or put more exactly which parameters should be chosen from the collective of parameters to achieve satisfactory structural portraying and disclosure of individual patterns of each text from the collection.

Besides, it may be of importance to note, that the expressions (4) and (5) we have chosen quite arbitrarily, rather as a counterpoise to the laws of Zipf (2) and Mandelbrot (3) which as a matter of fact necessitate theoretical explanations and modifications as is witnessed by publications of numerous studies devoted to this subject.

To assess the significance of each pair of parameters (1-2, 1-11, 2-11, 3-4, 5-6, 7-8, 9-10) and the role they play in disclosing structural pattern of different texts let us first examine the results of hierarchical classification obtained by use of the following pairs of parameters:

- parameters "1-2" representing basic characteristics of the text, i.e. text size and the number of appearance of a word having rank 1. The results of classification based on application of this pair of parameters confirmed the hypothesis I only.

- parameters "3-4" representing coefficients of Zipf's law. Classification disclosed the following separate groups: the class of natural language texts, the class of artificial language texts including text 29 (dictionary); it was also possible to distinguish "Verlaine class" excluding texts 19 and 20 (translations);

- parameters "5-6" representing coefficients of Mandelbrot's law. The classification on the basis of this pair highlighted Verlain class only, excluding texts 19 and 20 (translations). Natural language texts fused with artificial language texts, including dictionary text, into one large cluster. Confirmed was the specific structural pattern of the

texts 7 and 13.

- parameters "7-8" representing coefficients of the arbitrarily chosen expression (4). Results of classification do not allow us to reach any definite conclusion.

- parameters "9-10" representing coefficients of the expression (5) which is symmetrical in respect to (4). Results of classification confirm the hypothesis I. More detailed classification brought out text N°3 as belonging to the class of artificial language texts. It was possible to distinguish the Verlaine class excluding translation texts 19 and 20. Translation text N°7 was identified as possessing specific structural pattern.

- parameters 1-11 and 2-11 gave similar results.

By matching the results of classification obtained on the basis of different pairs of parameters with that of applying all 11 parameters we were able to reach the following conclusion

The pairs of parameters "1-2" and "9-10" exhibited an inherent property to divulge specific patterns of text structures in respect to hypothesis I. Analysis of the position of objects on their projections pointed to the "9-10" pair as being the most expressive in revealing structural patterns. This was further confirmed during visual interpretation which demonstrated capability of the "9-10" pair to elucidate subtle differentiation of texts in accordance with all the hypotheses. In addition, the visual interpretation coincided with results of classification on the basis of all 11 parameters.

It is noteworthy that the results of classification based on application of hierarchical cluster procedures using "9-10" parameter pair are in some ways inferior to the results of the visual interpretation carried out on the basis of projection of this pair. Besides, the results of applying hierarchical cluster procedure using collective of 11 parameters are in complete concordance with visual interpretation based on the "9-10" pair projection.

Parameters "3-4" (coefficients of Zipf's Law exhibited an inherent property of being able to disclose specific struc-

tural patterns of natural and artificial language texts. Analysis of the position of text numbers on their projections pointed to parameter 4 as being of decisive importance in classification.

Application of the hierarchical cluster procedures on the basis of parameters "5-6" (coefficients of Mandelbrot's law) resulted in the classification which was diametrically opposite to the one obtained by the visual interpretation of their projections to the plane. Visual interpretation coincided however with the classification made on the basis of the collective of 11 parameters. Actually, we were able to distinguish on the projection the class of artificial language texts including text N°3 (that was one more confirmation of this text possessing specific structural pattern) as well as text N°29 (dictionary), and to obtain fine differentiation of the natural language texts. However, the compactness of these classes differed so much, that soon became evident that formal cluster analysis on the basis of "5-6" parameter pair would not be feasible due to inability of this method to compensate for wide ranges of dispersion of values of Mandelbrot coefficients in respect to the natural language texts. This inhomogeneity was displayed in a most obvious way in the case of Verlaine texts which the application of hierarchical cluster procedure brought into focus as a separate class while at the same time all other texts merged into one general class.

The results of matching analysis allow us to suggest that coefficients of Mandelbrot's law possess an inherent property of being able to effect subtle differentiation of semantically significant natural language texts. Let us explain what we have in mind when we use expression "subtle differentiation". Almost all the results of the data processing point to the similarity of structural patterns of the texts 5 and 6; 17, 14, 15, 16, 18 and so on. On the other hand, the significant scattering of these and other text numbers on the projection of Mandelbrot coefficients provide evidence to the contrary. In this way, the coefficients of Mandelbrot's law

may be compared with a microscope with large factor of enlargement with which we aim to observe relationship among objects but which brings into our scope of view the individual properties of objects only. In other words, application of Mandelbrot coefficients presupposes careful and appropriate selection of the objects of investigation.

In Appendix 3 listed are the projections of parameters onto the plane, coordinates of which are the first two components (the method of principal components was applied to 34x11 experimental data table, which was rotated by 90° in relation to the initial position). Observed "closeness" of parameters may be taken as an indication of intensive correlations among parameters 1, 2, 11; 3, 4, 5.

Now the results of text classification and the influence of different parameters could be made more precise by using projection onto the principle components and applying cluster analysis procedures for different subsets of parameters.

By applying such methodology of data processing we were able to solve two tasks at once: to effect stepwise refinement of text structure differentiation, and reveal parameters most appropriate for a particular application.

The methodology which we used for classification of text collection allowed us to differentiate specific patterns of text structures to a much finer scale than it was possible with the method of visual interpretation based either on the empirical ranking distribution of word occurrence in texts or on the plots derived from the laws of Zipf (2) and Mandelbrot (3) (see Appendix 6).

Contents

Forword 3
Introduction 5
I. Logical and linguistic features of a dialogue . . . 7
II. Structure, vocabulary and "naturalness" of language 17
III. Collection of texts for analysis 23
IY. Quantative investigation of language texts 30
Y. Structural methods of data analysis 35
YI. Analysis of the results of data processing 38