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 2

Leningrad
1984

УДК 681.142.4/519.24

Аннотация

В препринте обсуждаются вопросы эффективной организации диалогового взаимодействия с ЭВМ. Работа состоит из двух частей. Первая часть посвящена общетеоретическим вопросам. Во второй части текст рассматривается как сложная иерархическая система, которая может быть описана рекурсивными структурами. Приведена физическая интерпретация структурного развития текста и предложена численная характеристика оценки "естественного" текста. Вторая часть включает также приложения, содержащие результаты экспериментов, список литературы для обеих частей и общие выводы.

Abstract

Questions dealing with effective organization of interaction with computer in dialogue are considered. The paper comprises two parts. The first part is devoted to general aspects of the problem. In the second part text is treated as a complex hierarchical system which can be described by recursive structures. Physical interpretation of structural development of text is given and quantitative characteristics of natural language text evaluation are suggested. The second part includes also Appendixes listing the results of experiments, the list of references for both parts and general conclusions.

Introduction

The technique of analysis presented in the first part of the paper is used here to help reveal specific patterns of various texts structures (poetry, translations, computer programs). The results of applying cluster analysis procedures to natural and artificial language texts indicate that their structural patterns differ significantly. This may be contributed to the difference in manners of constructing dialogues on the basis of them. Dialogue in natural language is based on use of words concepts, attributes, which the human participants can recognize thanks to previous experience and knowledge. Organization of such a dialogue is supported by linguistic approach, which is characterized by a definite regularity exhibited by the size (number of words) of text functional core and the number of auxiliary and specific words.

The artificial dialogue, on the other hand, (mathematical formulae, computer programs, etc.) is based on symbols which can be identified by means of the a priori specified definitions (axioms).

Organization of artificial language dialogues is supported by logical approach, based on sequential presentation of data which should not contradict initial axioms.

On the basis of the analysis of the two approaches to dialogue organization elaborated in the first part of the paper we may conclude that in the process of a dialogue the linguistic approach is oriented mainly to self-education and the role of the logical approach is to facilitate actual realization of the dialogue.

To enable use of the analysis results in synthesis of dialogues for problem-oriented information systems it was necessary to give physical interpretation of text structural development and find its artificial prototype. For example, a positional calculus system defined by the number of positions n, the number of different symbols q and the number of different texts $N = q^n$ may be referred to as an artificial

text. A physical interpretation of such a developing structure (or rather system, because its elements are related hierarchically) is made possible by use of the concept of the rate of its structural development (increase of number of elements as a function of level number). For such a developing structure we can construct a sequence of numbers, in a way similar to the ranking frequency distribution tables, which would reflect the relation between the number of modes in a subtree and the number of subtrees having such a number of nodes (table, part 2).

Eventually, it became evident to us that to solve the problem of synthesis of text structural development one has to devise such artificial constructions whose developing tree structure would have the structural pattern similar to the ranking distribution of natural language texts.

Recursive structures

What should we make out of the results obtained by the investigation of texts and what can we propose in the way of their physical interpretation ? What are the laws that govern structural development of texts ?

The results of classification suggest that the laws of Zipf, Mandelbrot and their modifications are more suitable to be applied in identification of integral interrelationships among size, vocabulary and problem orientation of various texts

than in revealing structural characteristics of text construction. To overcome this restriction the authors of the current study have striven intuitively to estimate the parameters of approximation function in such a way so as to achieve the best concordance between the empirical and theoretical ranking distributions for the observed class of texts. That deficiency is also one of the reasons why graphs based on direct plotting of the empirical and theoretical ranking distribution, in accordance with (2) - (5), part 1 of various texts (Appendix 5) failed to provide us with sufficient information to facilitate classification.

Nonetheless, we were able on the basis of classification experiments to identify the role and the effect the individual parameters (table 1) have in the process of classification.

The methodology we employed in text classification allowed us to ascertain beyond doubt that approximation of ranking distribution on the basis of the collective of parameters enables one to gain insight into the general and specific pattern of text structures.

To be able to carry out physical interpretation of the observed specific patterns of text structures one should synthesize the artificial structural systems and in the process of classification match system parameters defining its development and functioning with parameters of texts (Appendix I). Consequently, the primary objective of an investigation should be the identification of structural interrelations among the elements comprising observed object (text in our case).

We based all our observations on the premise that the structure of a text should coincide with the structure of a dialogue, if we wish to achieve an aim of establishing common understanding between its participants (unambigous comprehension of the concepts being conveyed). That is, for a text to be comprehended easily, it should incorporate explanations of all the concepts, rarely used words, expressions, etc. that are being introduced in the course of a dialogue. The rate of text structure development depends on the

way and degree this explanation facility is embedded.

It seems obvious that the highest rate of interaction should be at the level of generic structural entities (words) which form the basis for the construction of the higher level elements like definitions, etc. At the level of definitions and concepts the rate of interaction should be less intensive on account of the simple fact that these elements are products of the interaction on the lower level.

The process of recursive construction of semantically significant texts proceeds until all the words, concepts, etc. used in the process, are interrelated by means of definitions and associations.

Chomsky indicated that the most evident formal feature of an utterance is a possibility to divide it by means of parenthesis notation into components of various types, i.e. an utterance is associated with tree-structures /25/.

Formally, any text could be represented by a set of interrelated elements forming a system with complex relations.

On the other hand numerous authors /26,27,28/ observed that to predict the behaviour of any complex system of arbitrarily interrelated structure (that is to make it controlled) the system should be of necessity hierarchical.

In /26/ we read "...a step towards hierarchical structure makes the diversity of strategies less but at the same time it reduces indefinity, i.e. it makes it possible to have more effective solution".

However the question remains open of how to represent any complex system with a given set of relations in the form of a set of hierarchical systems.

To bring some light into this matter let us introduce several definitions and theorems /14,15,29/.

Definition 1. Quasihierarchy is an oriented graph, which can be constructed from an oriented tree by fusing some of its terminal - nodes (leaves). If an initial tree is binary, then the resulting quasihierarchy is called the binary quasihierarchy.

Theorem 1. Any n-ary relation R defined over a finite set M can be represented by a quasihierarchy, with terminal-nodes being the elements of the set M.

Consequence. Any system of relations defined over a finite set M may be represented by:

a) a system of quasihierarchies, with the terminal-nodes being the elements of M;

b) a quasihierarchy with the terminal-nodes being the elements of M;

c) a binary hierarchy with the terminal-nodes being the elements of M.

Definition 2. A hierarchy H defined over a set M is a self-similar recursive structure if it can be reduced to a notation:

$$H^0 = M, \quad H^{m+1} = G(H^m),$$

where an operator G does not depend on the level number of the hierarchy m.

A structure produced by means of using G-operator possesses a feature of self-similarity, because every succeeding level is produced by G-transformation of the given level. Information about a number of elements generated at the next level is also hidden in G.

The definition just introduced represents a special case of primitive recursion:

$$f(0) = V, \quad f(m+1) = \psi(m, f(m)),$$

and illustrates the possibility to obtain the developed hierarchy by defining only once the rule of decomposition of the initial set M: $H' = G(H)$ with all the following levels of the hierarchy being derived automatically.

Generally, a self-similar structure is characterized by incorporating an etalon (genotype) and the recursive law of its development.

Constructive technique of generating of self-similar recursive structures in the form of ordered trees /15,30/ consists in superposition of a leaf with the root of an etalon (tree fragment).

This technique facilitates construction of any complex tree configuration by means of etalon compositions.

For a set of self-similar trees analytic formulae can be used instead of constructive technique /15,30/.

Example 1.

Let us interprete a function

$$f(0) = 0, \quad f(n) = n - f(n-1) \tag{1}$$

given in the form of primitive recursion as a tree function. If we use natural numbers sequentially as arguments $n = 1,2,3\ldots$ we shall produce a sequence of values each presenting succeding node for an argument, i.e. a string called canonic representation of a tree structure /31/.

n:	1	2	3	4	5	6	7	8	9	10	...
$f(n)$:	1	1	2	2	3	3	4	4	5	5	...

So (9) presents a binary tree.

Example 2.

$$f(0)=0,\ f(1)=1,\ f(n)=n-f(n-1)-f(n-2) \tag{2}$$

n:	1	2	3	4	5	6	7	8	9	10	...
$f(n)$:	1	1	1	2	2	2	3	3	3	4	...

Function (10) corresponds to a ternary tree.

Example 3.

$$f(0) = 0, \quad f(n) = n - f(f(n-1)) \tag{3}$$

n:	1	2	3	4	5	6	7	8	9	10	...
$f(n)$:	1	1	2	3	3	4	5	5	6	6	...

Function (10) corresponds to a tree enumerated in such a way that numbers of nodes at the right edge of the tree present the sequence of Fibbonacci: 1,1,2,3,5,8,13,21 ...

Fig. 1 a,b,c illustrates binary, ternary and Fibbonacci trees; the dashed lines indicating etalons which according to constructive technique rule can be used for their construction.

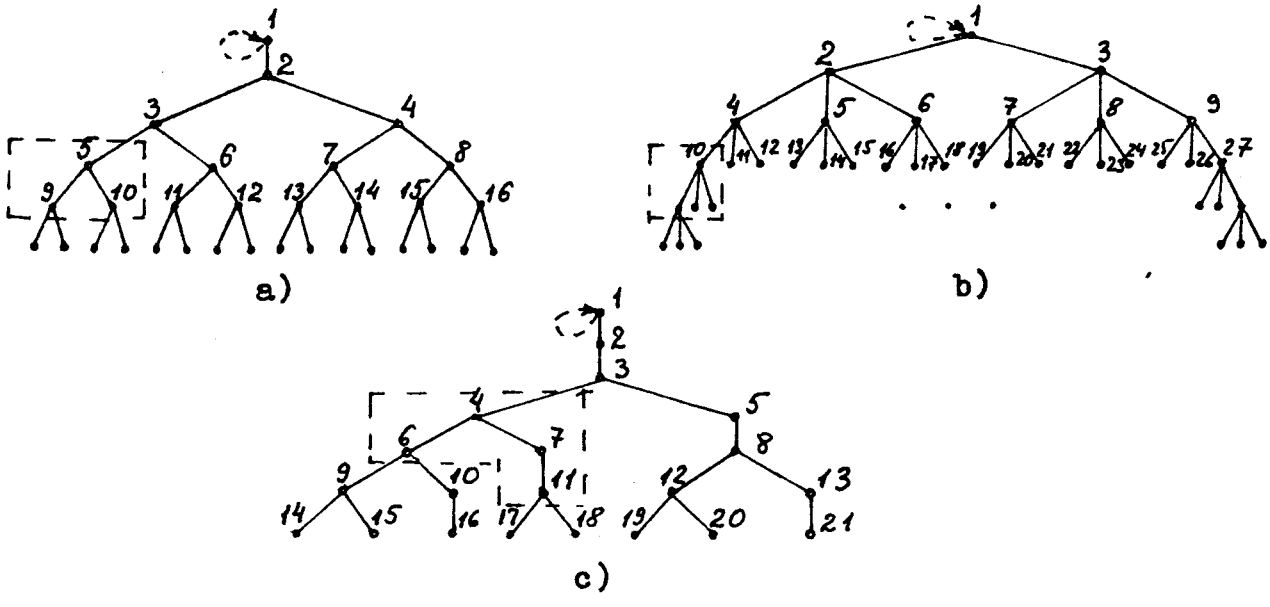


Fig.1. Ordered trees generated by (1) - (3); a) binary tree, b) ternary tree, c) Fibbonacci tree.

Theorem 2. Any quasihierarchy can be presented by self-similar recursive structures (SRS).

Theorem 3. Self-similar recursive structures have the least complexity (in the sense of Kolmogorov /32/) of their algorithmic presentation.

From the analytic representation of SRS it follows that the complexity of generating of dynamic discrete system depends only on a number of steps (tacts) n .

Using a constructive technique the complexity of generating of such structures is determined by a complexity of initial etalon and a number of its applications.

SRS are defined in a unique way by its etalon and a number of levels, which ought to be sufficient to classify them as different types. Recognition of SRS types, a composition of which presents a developing system with complex structural relations, provides the way for its physical interpretation. For example, positional system of calculus - a special case of SRS (fig.1 a,b) - is fully defined by a rate of increasing of elements number from one level to another. Physical interpretation of this parameter - the rate of quantitative changes,

i.e. changes of value of one and the same sign depending on a level number (position).

It is quite clear that in artificial systems where an unambigous identification of an object is predetermined by the rules of object construction (logical approach) the rate of increasing of their number can be much higher than in natural systems, where the redundance is inevitable in identifying an object (linguistic approach).

Table 2 presents structural development of positional systems of calculus with base $q=2$ and base $q=3$ corresponding to SRS in fig.1 a,b and an SRS corresponding to a Fibbonacci tree in fig.1, c.

Table 1.

a) Ternary tree (N°30)

243	121	40	13	4	1
1	2	6	18	54	162

b) Binary tree (N°31)

256	255	127	63	31	15	7	3	1
1	1	2	4	8	16	32	64	128

c) Fibbonacci tree (N°32)

233	232	231	142	88	87	54	53	33	32	20	19	12	11	7
1	1	1	1	1	2	1	3	2	5	3	8	5	13	8
6	4	3	2	1										
21	13	34	21	89										

d) Lucas tree (N°33)

272	271	270	269	186	126	85	82	81	80	59	58	57	40	39
1	1	1	1	1	1	1	1	1	1	1	1	3	1	1
38	27	26	25	22	22	21	20	18	17	16	14	12	11	10
4	1	1	5	1	1	1	1	3	3	8	1	4	4	13
8	7	6	5	4	3	2	1							
5	5	18	9	9	41	32	86							

e) Tree $f(n) = n - f(f(f(f(n-1))))$ (N°34)

250	249	248	247	246	177	127	91	68	67	66	65	49	48	47
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
46	35	34	33	32	25	24	23	22	18	17	16	15	13	12
3	1	1	1	4	1	1	1	5	2	2	2	7	3	3
11	10	9	8	7	6	5	4	3	2	1				
3	10	4	4	4	19	5	12	36	31	69				

f) A.S.Pushkin (N°3)

113	69	57	30	29	27	26	24	16	15	14	13	12	11	10
1	1	1	1	1	1	1	1	1	1	1	1	2	4	2
9	8	7	6	5	4	3	2	1						
5	4	6	13	17	27	54	151	665						

g) P.Verlaine (N°17)

4	3	2	1
4	5	9	31

h) Dialogue with IS SITO (N°23)

187	150	135	134	115	113	91	60	36	30	26	25	24	22	20
1	1	2	1	1	1	1	1	2	1	1	1	1	1	1
19	17	15	14	12	11	10	9	8	7	6	5	4	3	2
1	1	2	2	1	3	1	3	3	2	13	14	14	96	95
1														
250														

i) Algol-60. SITO (N°25)

282	246	238	208	172	165	164	110	97	58	48	47	46	38	35
1	1	1	1	2	2	1	1	6	5	2	2	1	1	2
34	32	30	26	25	24	23	22	21	20	18	17	16	15	14
2	2	3	2	1	3	1	1	4	1	2	1	2	2	4
13	12	11	10	9	8	7	6	5	4	3	2	1		
1	3	3	1	3	5	3	9	7	12	12	21	41		

The first line of the table contains sequence of numbers of nodes in a subtree of the initial tree $F(r_g)$ and the second line contains corresponding numbers of subtrees with the same number of nodes (r_g). The set of all existing subtrees is arranged in decreasing order of numbers in the first line (the number of levels of initial tree being fixed).

Values in this table reflect structural characteristics of artificial systems with clear physical interpretation. We used these strings as additional artificially constructed texts (N°30-34).

We included also in the table 1 the strings of ranking distribution of word frequency describing development of structure in text systems where the first line contains the number of occurrences of a word and the second line contains the corresponding numbers of words with such a number of occurrences.

Let us give an example of construction of structural development of text N°3 and its approximation by means of SRS (fig.2).

Fig.3 presents graphs plotted according to table 1 the physical basis of which is the rate of development of systems with structural relations.

It follows from the fig.3 that structural development of a text type structures possess features of its own. It is possible to recognize three distinct segments I, II, III in the rate of structural development.

The rate of development could be approximated by a composition of three self-similar recursive structures:

$$f(n) = \begin{cases} n-1, & 1 \le m \le \ell/2 \\ n - f(f(n-1)), & \ell/2 < m \le \ell-2 \\ n - \sum_{i=1}^{t} f(n-i), & m = \ell-2, \ell-1, \quad 5 \le t \le 10 \end{cases} \qquad (4)$$

where m is a level number of the tree-structure, ℓ - is the number of the last level.

Let us examine physical sense of this formula.

The segment I of a function in fig.3 is approximated by a unary system which is characterized by an absence of structural relations influencing qualitative changes in the union.

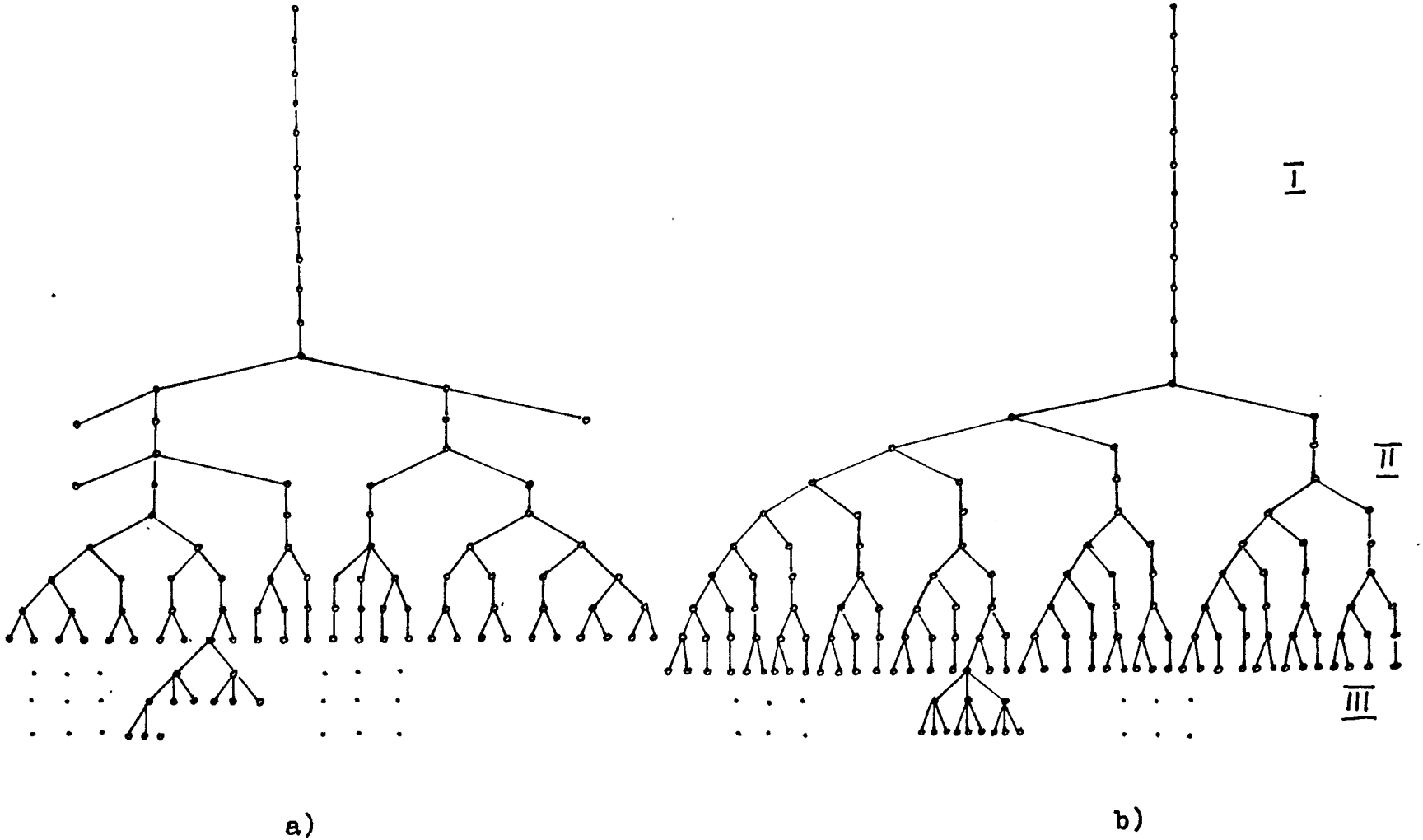


Fig.2. Group rank frequency representation (a) and its approximation by self-similar structure (b).

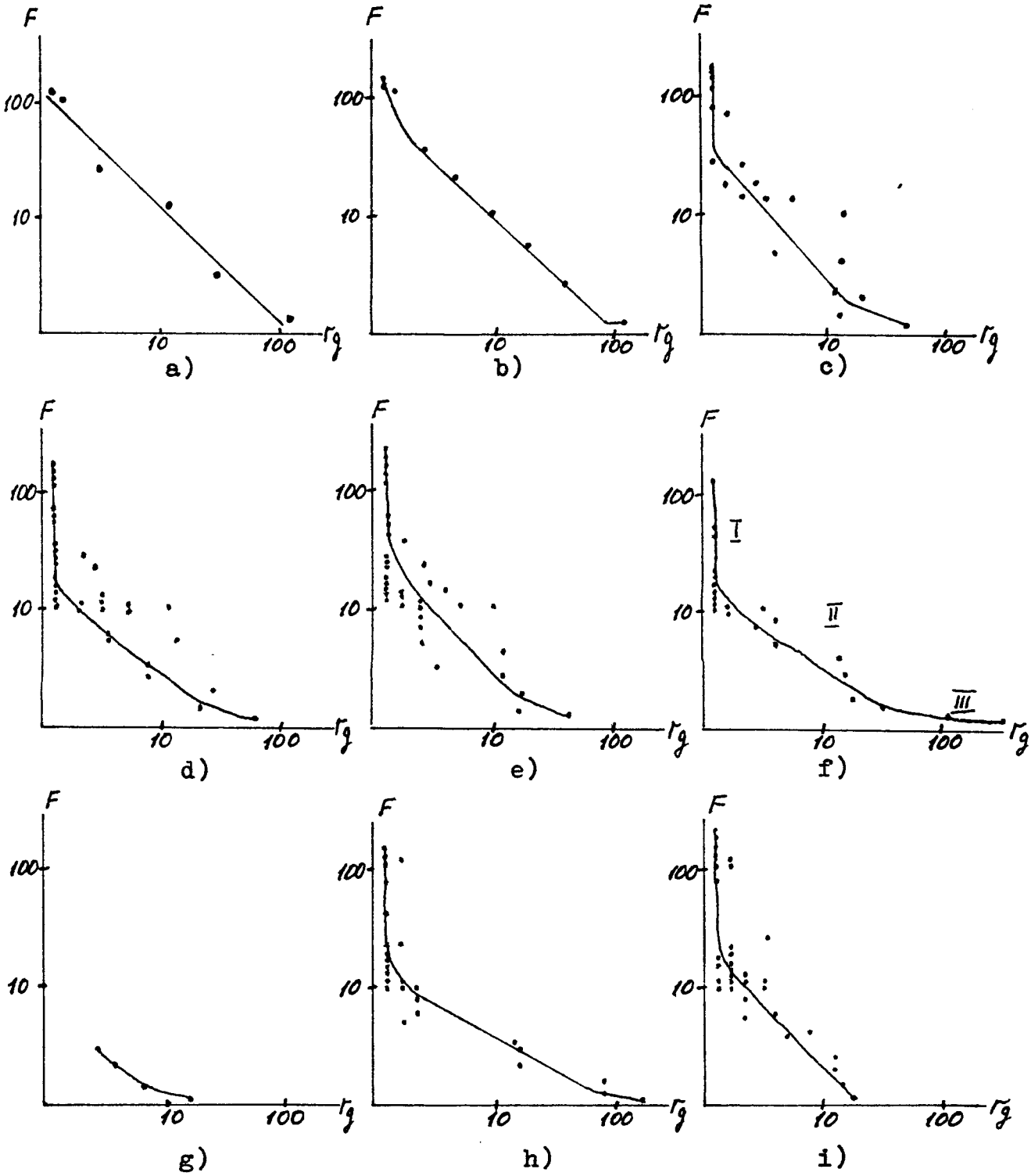


Fig.3. Rate of structure development corresponding to texts a) N° 30, b) N°31, c)N°32, d)N°33, e)N°34, f) N°3, g) N°17, h)N°23, i) N°25.

In fact this segment corresponds to ranks representing auxiliary words (propositions, pronouns, conjunctions, etc.).

The second segment is characterized by a slow development of the structure and corresponding SRS rate is consequently lower than that of a binary [illegible].

The third segment has a high rate of structural development.

If we compare functions in the figure 3 with their approximations (4) we shall see, that the rate of structural development of natural texts (α_n) occupies an intermediate position between those of artificial texts (α_a) and positional systems of calculus (α_p):

$$\alpha_p > \alpha_n > \alpha_a \geqslant \alpha_u$$

where α_u symbolizes primitive (unary) system where union of elements does not result necessarily in qualitative changes; α_p is a system possessing the highest rate of development (qualitative changes). Let us remind that a fundamental discovery for such systems was the demonstration of sufficiency of Z signs to identify N objects, where $Z = \log N$.

Now we may interprete the role assigned to special languages: their function is to facilitate generating of definitely comprehended texts out of signs (notions, definitions). The role of their grammar is to form rules of transforming α_u to α_p.

The main difference between natural and artificial texts is that natural texts are characterized by specific inner construction and the way the concepts and axioms are defined and developed (a posteriori) while artificial texts inevitably have a priori postulated axioms which predetermines text constructions. In this sense SRS may be treated as analogues of natural texts. They contain in their notation (1)-(4) both the technique of etalon construction (genotype) and the regularity of structural relations which differ from the traditional mathematical methods of function representation.

Conclusions.

In place of answering directly the basic question put forth in the subtitle of the present study "Is Dialogue a Science or Art?" let us paraphrase the quotation of Flaubert which we used in the way of an epigraph to introduce our theme.

Strange as it may seem the natural language texts appear to be more scientific than their artificial counterparts on account of the capability of natural languages to be more easily comprehended and handled by arbitrary human end-user.

If we really wish the computer to be utilized effectively by large community of various users we should organize communication with it on the basis of natural language in such a way that both the structure and principle of organization of man-machine interaction correspond to the structure of natural language texts. Consequently, their dialogue will turn into a discourse of "equal" partners mutually striving in a friendly way to achieve common understanding of the concepts being imparted in the form of text.

In the structure of natural language texts the following three segments can be distinctly distinguished:

I - constant part comprising auxiliary or service words;

II - collection of definitions and generalizations (extension and elaboration of the meaning of a word). Here, the rate of structural development is rather low;

III - rapid accumulation of the essential vocabulary representing the basis for text construction. Here, the rate of structural development is the highest.

All three segments together make the foundation of the linguistic approach to construction of information processes. On the basis of exploratory investigation we carried out and by means of the parameter α which we conceived to specify the rate of structural development we are able to formulate now the main features which distinguish linguistic from logical approach to construction of information processes.

At the bottom of the logical approach lies an axiomatic system defined a priori. The texts generated on the basis of the logical approach proved themselves to be very effective in construction of reproducible messages and are characterized by the high rate of structural development. However, when using this approach one has to constantly hold in mind all the rules of the generic axiomatics, and in case of their being modified or extended both conceptual meaning and reproducibility of the text are destroyed.

On the other hand, the linguistic approach is founded upon an a posteriori formed axiomatic system which consists of the invariant characteristics such as winter, summer ..., attributes like good, bad..., and concepts of the form: person, thing ..., all of which could be comprehended either directly or associatively by means of an information model created in the process of interaction with surrounding world. The inambiguity of reproduction may be achieved through gradual narrowing of the indeterminacy region and use of appropriate axioms in description of information processes. Accordingly, the prime objective of culture and language in society seems to be to preserve and synchronize the invariants of perception. At the same time the texts which generate these invariants may take variety of forms and differ in content. In contrast to logical approach the linguistic approach necessitates large vocabulary of generic items to be used for construction of texts (messages).

That is the reason why logical approach appears to be superior to the linguistic approach if measured by the criterion of traditional information theory, i.e. by the number of bits necessary to transmit a message (but certainly not in respect to comprehension capability). If we take into consideration the efforts spent on coding and decoding of semantically significant texts then the advantage is obviously on the side of the linguistic approach.

In case their axiomatic bases coincide the linguistic and logical approach seem to attain equal position. The attempts

to achieve such concurrence resulted in development of problem-oriented systems and languages akin to the natural ones. However, to exploit to the full extent the advantages that presents the use of logical type systems (they perform at the rate of the logarithm of the vocabulary size) one should specify in the rigid form the texts comprising perception invariants which are intended to perform the role of the sign, symbol, operator. The information interaction of the specialists should be based on the logical approach and the linguistic constituent should assume the role of a link in introducing new definitions and generalizations and of an intermediary in interacting with neighbouring subject areas. Information interaction of educational character should be based on the linguistic approach particularly in respect to construction of text perception invariants. Inner conceptual content of such texts is being revealed, however, through the use of external structural forms of knowledge representation which have been developed independently.

Accordingly, the word "good" may be referred to as a perception invariant only in case its use results in provoking corresponding run of activities.

Evidently, this brief report could not cover all the issues we engaged in and all the dilemmas we were confronted with in our exploratory investigation of dialogue process.

It might be said, that it is not carrying out concrete tasks, describing an instruction or creating a program that generates the need for linguistic approach but rather a striving to explore and comprehend the unknown, a wish to establish common understanding, a possibility to learn and evolve.

Thus, to perform an effective dialogue the elements of both linguistic and logical approach should be embedded in its organization and structure. For, in logical approach the knowledge develops in depth (specialization) and in linguistic approach in breadth (universalization). Investigation of texts embodying either of these approaches indicates that they differ significantly in respect to the rate of their structural development. This is due to the logical approach being oriented

to action (they i.e. the texts based on logical approach, are always specialized) and the linguistic approach to self evolution (teaching is always of a universal character). In other words, the transition line separating α_p , α_n and α_a seems to represent division of qualitative properties.

In accordance with the α-characteristics which we introduced to specify the rate of system structural development the program texts may be associated with the unary systems. Although the unary systems represent the most primitive systems in terms of the rules of interaction of vocabulary items they require large computer memory for their storage. Hence, it was not by accident that early programming languages were created in such a way that the individual operators were given the names associated with their actual meaning in natural language (e.g. go to). Further evolution of artificial programming languages led to inclusion of the linguistic constituent (e.g. LISP, APL, etc.). This aspect is clearly exhibited in the case of the texts 26,27,29 which on account of their commentary constituents were classified by our methodology as natural language texts.

It was intriguing to confirm the reputation of text N°3 "Bronze Horseman" by Pushkin as a mysterious poem /22/. Results of classification carried out on the basis of different methods pointed unequivocally to this text as belonging to the class of artificial texts. As can be seen from fig.5 which presents the ranking distribution of word occurrence frequency by means of self-similar recursive structures the Pushkin text exhibits possession of clearcut, stern and formal structural pattern. The method of ranking distribution of word occurrence frequency on the basis of self-similar recursuve structures which we used to interprete the text collection classification allowed us also to explain the phenomenon of the Verlaine texts. It is due to his manner of using service words to create sense and rhytm in the same way the basic words and notions are used to define conceptual content

(total absence of segment I)that Verlaine texts stand out as a separate group in the class of natural language texts.

By using self-similar recursive structures to represent ranking distribution we were able to acquire evaluation criteria of degree of correspondance between the translation and the original (certainly not in terms of the general quality of translations). For example, the translations of Verlaine verses by Ėrenburg and Pasternak exhibit specific structural patterns differing from those of the originals (texts N°19 and 20). It was the poetic genius of these literary giants that transformed the structure of the original into the structures of their own and left the sense intact.

Acknowledgement

The authors wish to thank Dr. A.Reicher for making available the paper in English.

Thanks are also due to Dr.A.P.Mogilyansky who consulted us in linguistic aspects of the poem "Bronze Horseman" by A.S.Pushkin.

References

1. T.Winograd. Understanding Natural Language.- N.Y.: Academic Press, 1972.
2. M.Minsky. A framework for Representing Knowledge. Mass.Inst. of Technology, Cambridge, 1974.
3. J.A.Levin, J.A.Moore. Dialogue-Games, Meta-Communication Structures for Natural Language Interaction.- Cognitive Science, 1977, v.1, N°4.
4. J.Iivary, P.Kerola. A sociocybernetic framework for the feature analysis of information systems design methodologies. North Holland, IFIP, 1983.
5. G.Leibnitz. Sämtliche Schriften und Briefe. Hrsg. von der Deutchen Akad. der Wissenschaften zu Berlin. Reihe Berlin, Akad.-Verl., 1926, Bd.4.
6. Мельчук И.А. Опыт теории лингвистических моделей "Смысл-Текст". М.: Наука, I974.
7. N.Chomsky. Aspects of the theory of syntax. Cambridge, Mass., 1965.
8. Попов Э.В. Общение с ЭВМ на естественном языке. М.: Наука, I982.
9. Пиотровский Р.Г. Инженерная лингвистика и теория языка. Л.: Наука, I979, II2 с.
10. B.Fox. Design-based studies: an action-based "form of knowledge" for thinking, reasoning and operating. IPC Business Press. Ltd, vol.2, N°1, 1981.
11. G.K.Zipf. Human Behavoir and the principle of least effort. Addison Wesley Publ.Co., Inc., Cambridge, Mass., 1949.
12. B.Hill. Zipf's law and prior distribution for the Composition of a Population. Journal of American statistical Association. 65(331): 1220-1232; 1970.
13. Александров В.В., Горский Н.Д. Алгоритмы и программы структурного метода обработки данных. Л.: Наука, I983, с.208.
14. Александров В.В. Самоподобные рекурсивные структуры как способ представлений знаний в ЭВМ. - В кн.: Информационно-вычислительные проблемы автоматизации научных исследований. М.: Наука, I983, с.65-74.

15. Арсентьева А.В., Горский Н.Д. Конструктивный способ определения самоподобных рекурсивных структур. - Там же, с.75-79.
16. G.K.Zipf. The psycho-biology of language. The M.I.T. Press, Cambridge, 1965.
17. Ponomaryov V.M., Alexandrov V.V. Constructive approach to knowledge representation. "Comput.Ling. & Comput. Lang.", 1980, 14, 245-259.
18. A.Moles. Sociodynamique de la culture. Paris - La Hage, Movton, 1967.
19. Орлов Ю.К. Обобщенный закон Ципфа-Мандельброта и частотные структуры информационных единиц различных уровней. - В кн.: Вычислительная лингвистика. М.: Наука, 1976, с.179-202.
20. Александров В.В., Арсентьева А.В., Семенков А.И. Структурный анализ диалога. - Препринт ЛНИВЦ АН СССР,№ 80, Л., 1983.
21. Благой Д.Д. Миф Пушкина о декабристах. Социологическая интерпретация "Медного Всадника". М., 1927.
22. Белый А. Ритм как диалектика и "Медный Всадник". М., 1929.
23. Бонди С.М. История заполнения "Альбома 1833-1835 годов". - В кн.:Рукописи А.С.Пушкина.Фототипическое издание. Альбом 1833-1835 гг. М., 1939.
24. Гроссман Л.П. Пушкин. Биография. М., 1937, с.773-788.
25. N.Chomsky. Language and Mind, N.Y., 1968.
26. Моисеев Н.Н. Математические задачи системного анализа. М.: Наука, 1981.
27. J.Casti. Connectivity, complexity, and catastrophe in large-scale systems. N.Y., 1979.
28. H.Simon. Artificial Intelligence Systems that understand. Boston: MIT, 1977.
29. D.R.Hofstadter. Gödel, Escher, Bach: An eternally golden braid. N.Y.: Harvester press, 1979.
30. Александров В.В., Арсентьева А.В.; Горский Н.Д. Некоторые вопросы построения рекурсивных структур данных. - Управляющие системы и машины, 1981, № 4.
31. D.E.Knuth. The art of computer programming. V.1. Fundamental algorithms. Addison-Wesley Publ.Co., Mass., 1968.

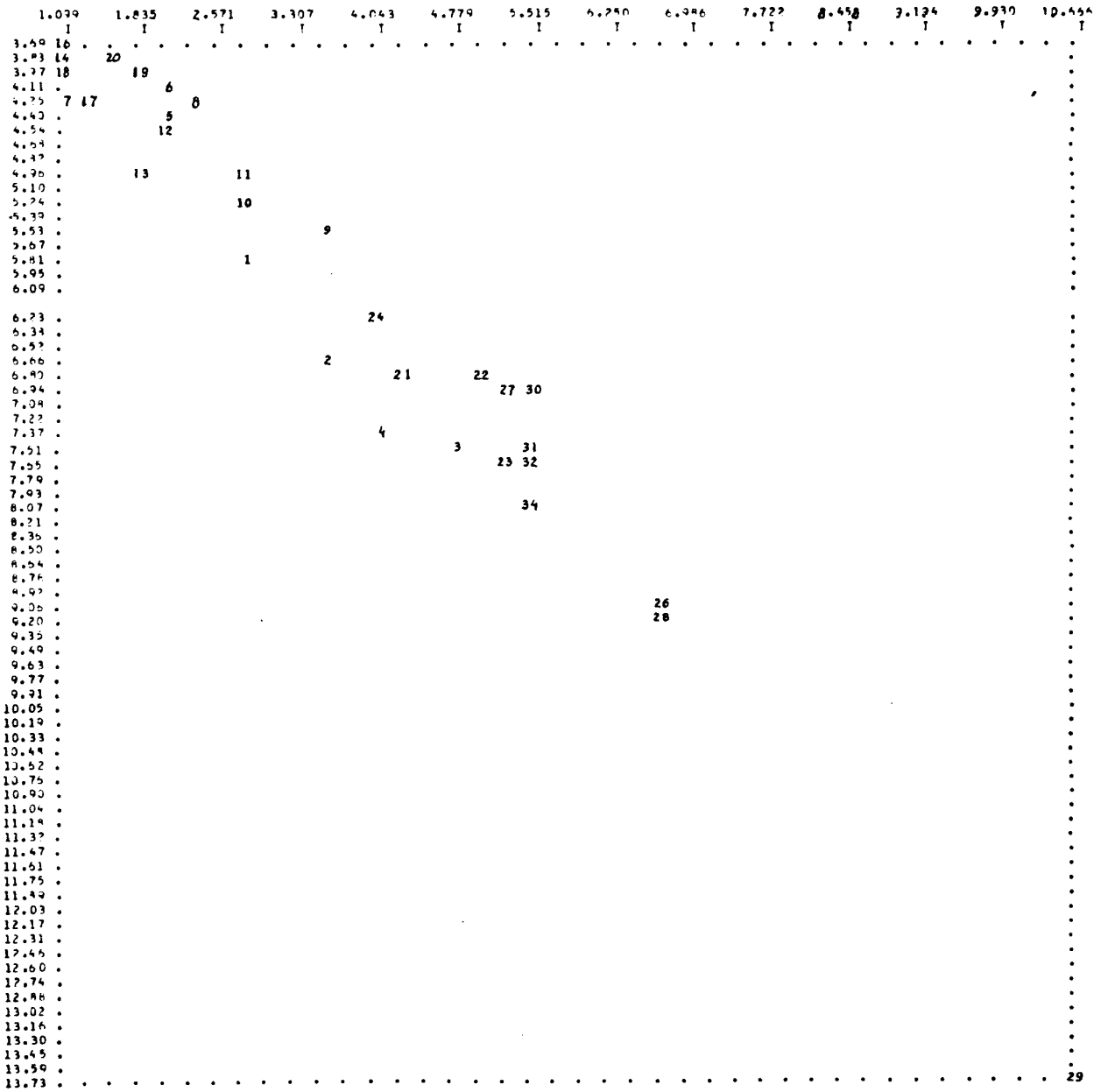
32. Колмогоров А.Н. Три подхода к определению понятия "количество информации". Проблемы передачи информации, т.I, вып.I, 1965.

Table of input data

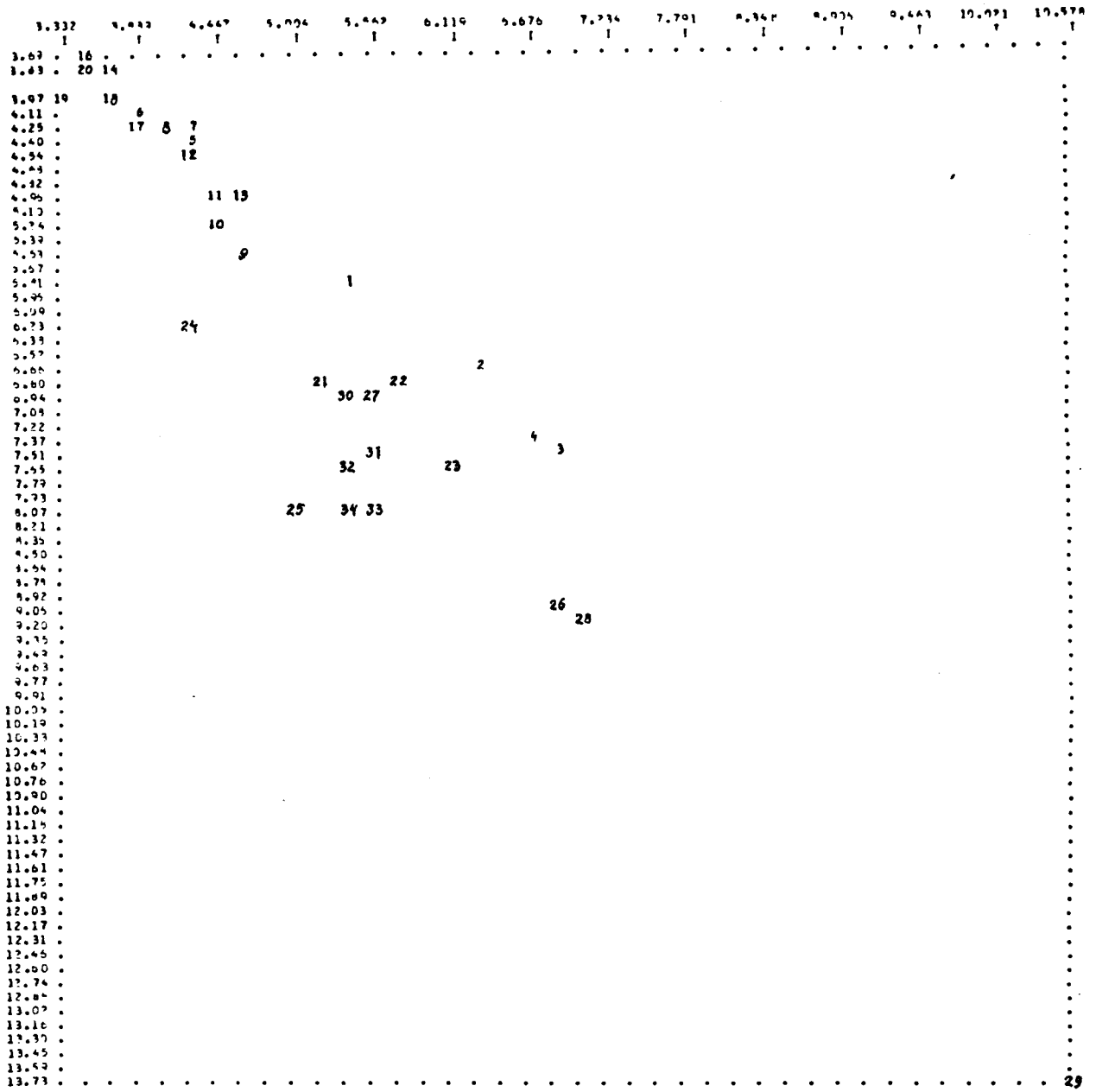
object number	max word appearance $\ln f_1$	text volume $\ln z$	coefficients formula (2) C	coefficients formula (2) β	coefficients formula (3) B	coefficients formula (3) K	coefficients formula (4) a_1	coefficients formula (4) b_1	coefficients formula (5) a_2	coefficients formula (5) b_2	vocabulary volume $\ln V$
	1	2	3	4	5	6	7	8	9	10	11
1	5.900	2.996	.05500	.554	.334	5.092	12.228	1.185	-7.057	-12.985	5.403
2	5.824	3.664	.04200	.582	.273	5.439	12.624	.800	-6.558	-14.449	6.292
3	7.525	4.727	.05500	.688	.212	2.836	12.155	1.216	-5.656	-16.466	6.872
4	7.475	4.007	.03100	.598	.250	6.999	12.672	.478	-6.287	-15.428	6.698
5	4.511	2.303	.110	.538	.434	2.952	11.780	2.036	-7.637	-11.058	4.277
6	4.248	2.079	.114	.521	.481	3.208	11.737	2.091	-7.887	-10.588	3.989
7	4.344	1.099	.03900	.258	.910	22.363	16.095	-.139	-11.890	-8.549	4.263
8	4.369	2.485	.152	.604	.402	1.649	11.173	2.591	-7.184	-11.329	4.111
9	5.595	3.526	.126	.758	.284	1.244	10.353	2.761	-5.984	-13.950	4.654
10	5.298	2.996	.100	.652	.334	2.338	10.786	2.333	-6.552	-12.930	4.595
11	5.106	2.944	.115	.656	.340	1.949	10.463	2.617	-6.474	-12.829	4.489
12	4.682	2.079	.07400	.488	.481	5.492	11.698	1.671	-7.870	-11.036	4.263
13	4.990	1.946	.04800	.422	.514	9.792	12.734	.882	-8.514	-10.863	4.615
14	3.912	1.099	.06000	.296	.910	14.171	14.622	.606	-11.066	-8.430	3.714
15	3.689	1.099	.07500	.304	.910	11.137	14.622	.829	-11.066	-8.207	3.611
16	3.689	1.099	.07500	.307	.910	11.137	14.554	.845	-11.027	-8.223	3.584
17	4.382	1.386	.05000	.356	.721	13.427	12.978	.857	-9.544	-9.621	3.892
18	4.007	1.099	.05500	.292	.910	15.688	14.554	.527	-11.027	-8.541	3.761
19	3.989	1.946	.130	.584	.514	2.964	9.139	3.428	-6.742	-11.406	3.332
20	3.951	1.792	.115	.493	.558	3.837	11.656	2.130	-8.222	-10.033	3.638
21	6.899	4.290	.07400	.822	.233	2.164	10.214	2.287	-5.443	-16.084	5.220
22	6.946	5.094	.157	.887	.196	.251	10.305	3.000	-5.027	-16.892	5.740
23	7.818	5.231	.07500	.837	.191	1.540	10.532	2.160	-5.011	-17.796	6.250
24	5.356	4.220	.118	.971	.237	1.007	6.809	5.206	-4.324	-17.919	4.344
25	8.157	5.642	.08000	1.129	.177	1.213	6.332	5.371	-3.693	-21.704	4.997
26	9.235	6.948	.09200	1.003	.146	.588	8.257	3.669	-3.875	-22.138	6.829
27	7.008	5.308	.183	.959	.188	.03100	9.534	3.545	-4.738	-17.560	5.533
28	9.335	6.837	.08200	.956	.146	.778	9.420	2.810	-4.117	-21.480	7.075
29	13.870	10.656	.04100	1.008	.09400	1.311	12.521	.789	-3.411	-28.529	10.578
30	7.103	5.489	.199	1.000	.182	-.08500	9.811	3.482	-4.724	-17.688	5.489
31	7.625	5.545	.125	1.000	.180	.443	9.471	3.199	-4.519	-18.449	5.545
32	7.792	5.451	.09600	1.000	.183	.906	9.833	2.744	-4.746	-18.328	5.451
33	8.157	5.606	.07800	1.000	.178	1.288	8.912	3.059	-4.458	-19.273	5.605
34	8.187	5.521	.07000	1.000	.181	1.604	8.661	3.107	-4.427	-19.481	5.521

Appendix 2

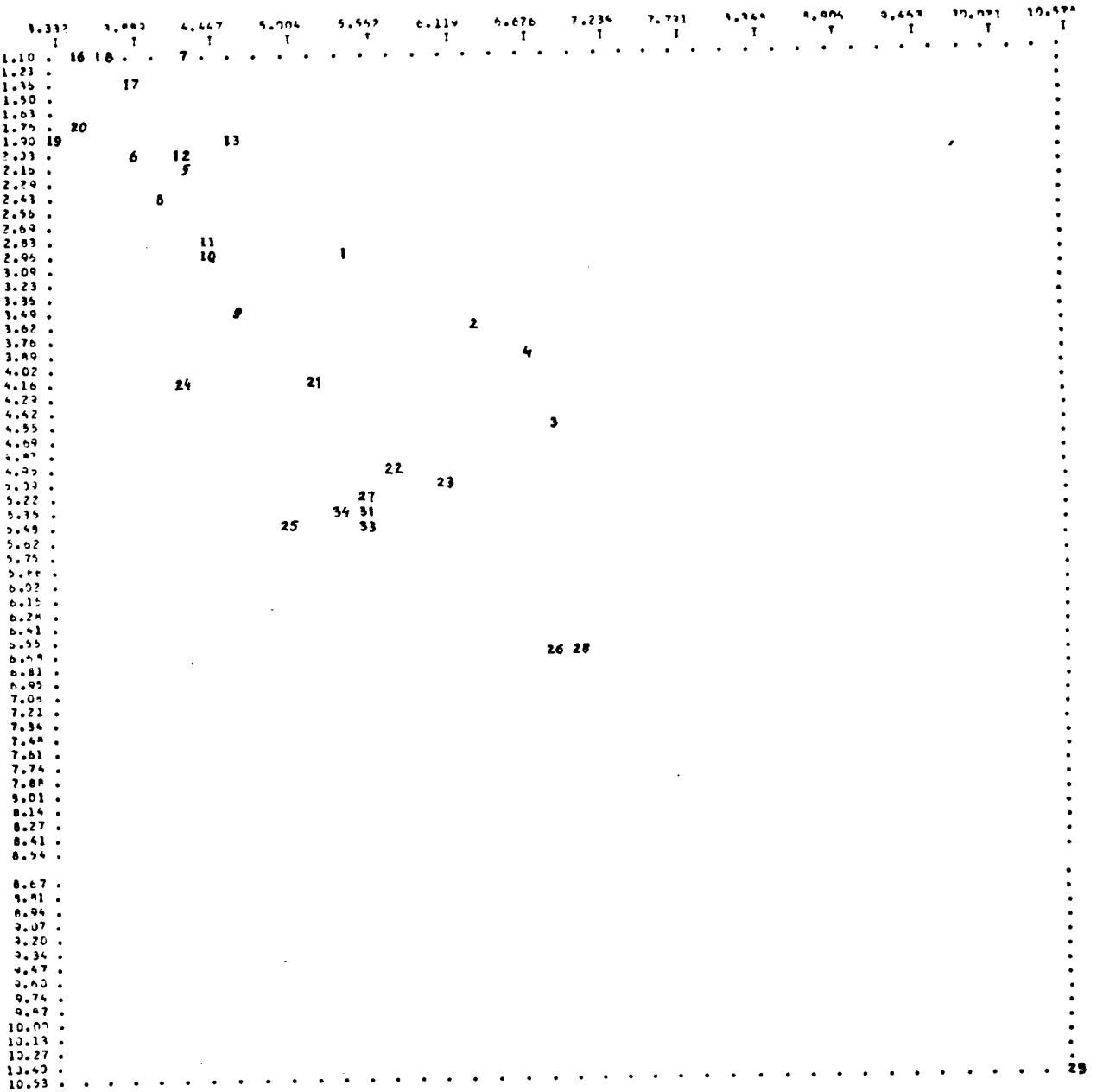
Projection on 1 & 2 features plane



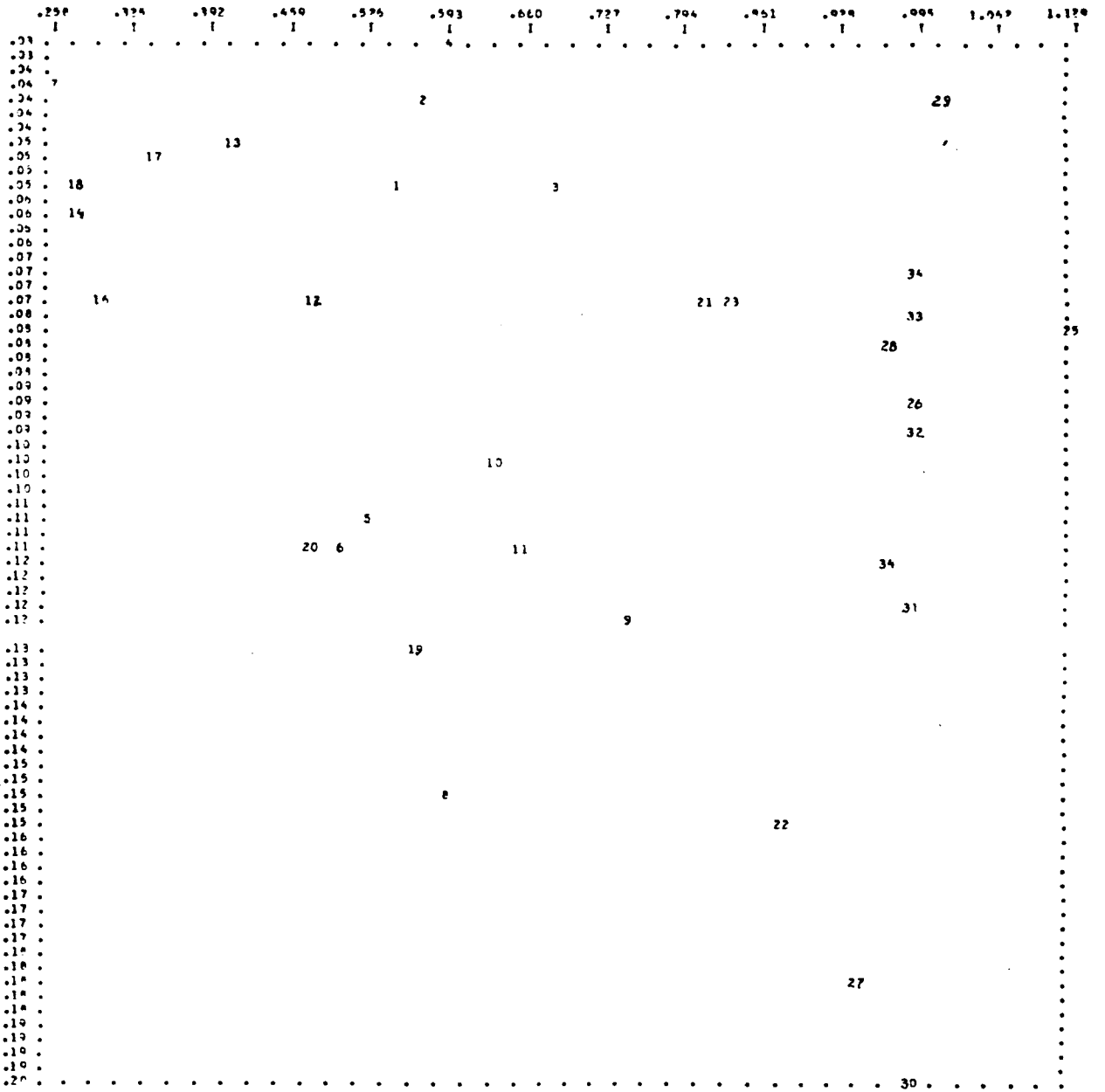
Projection on 1 & 11 features plane



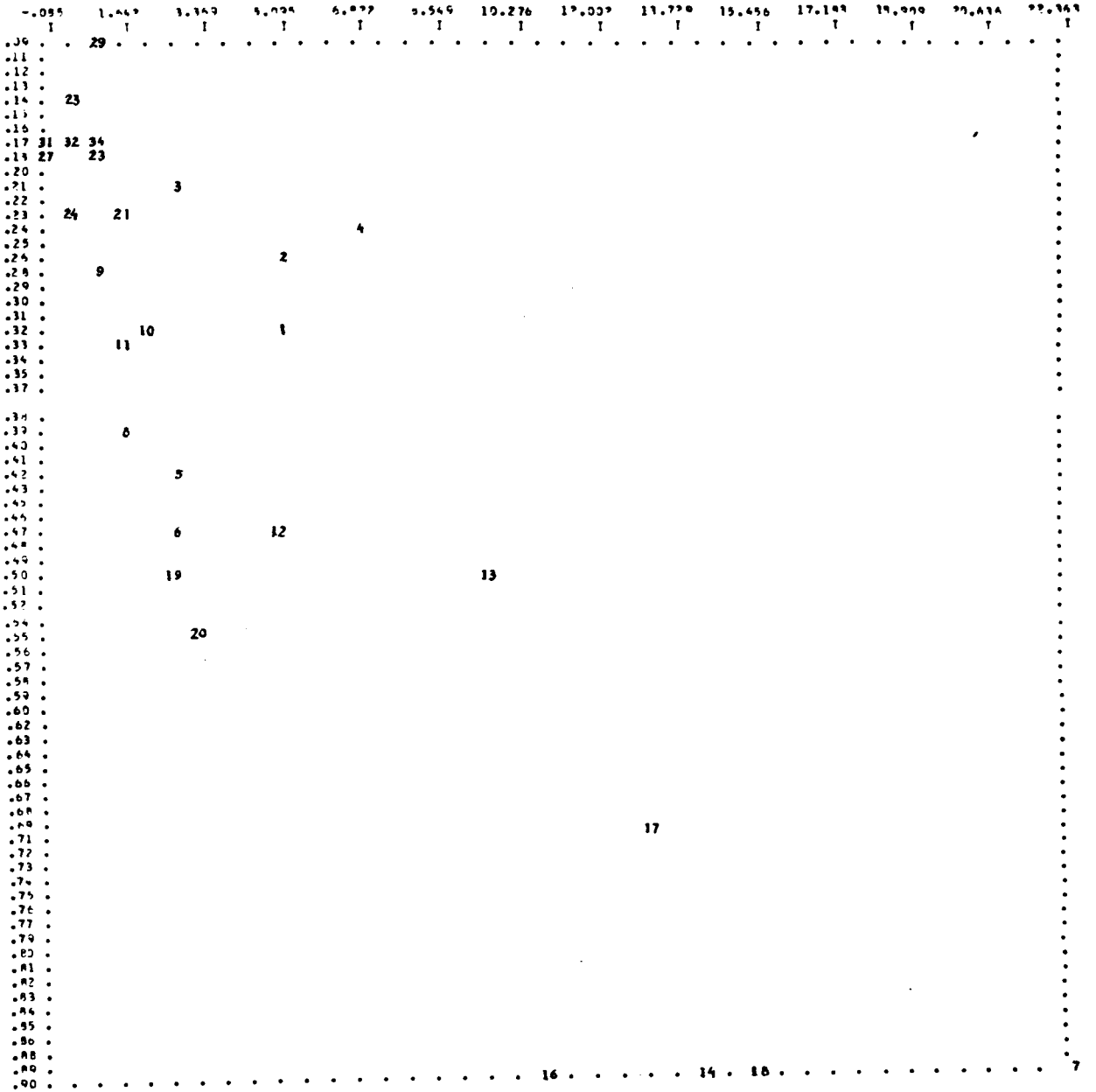
Projection on 2 & 11 features plane



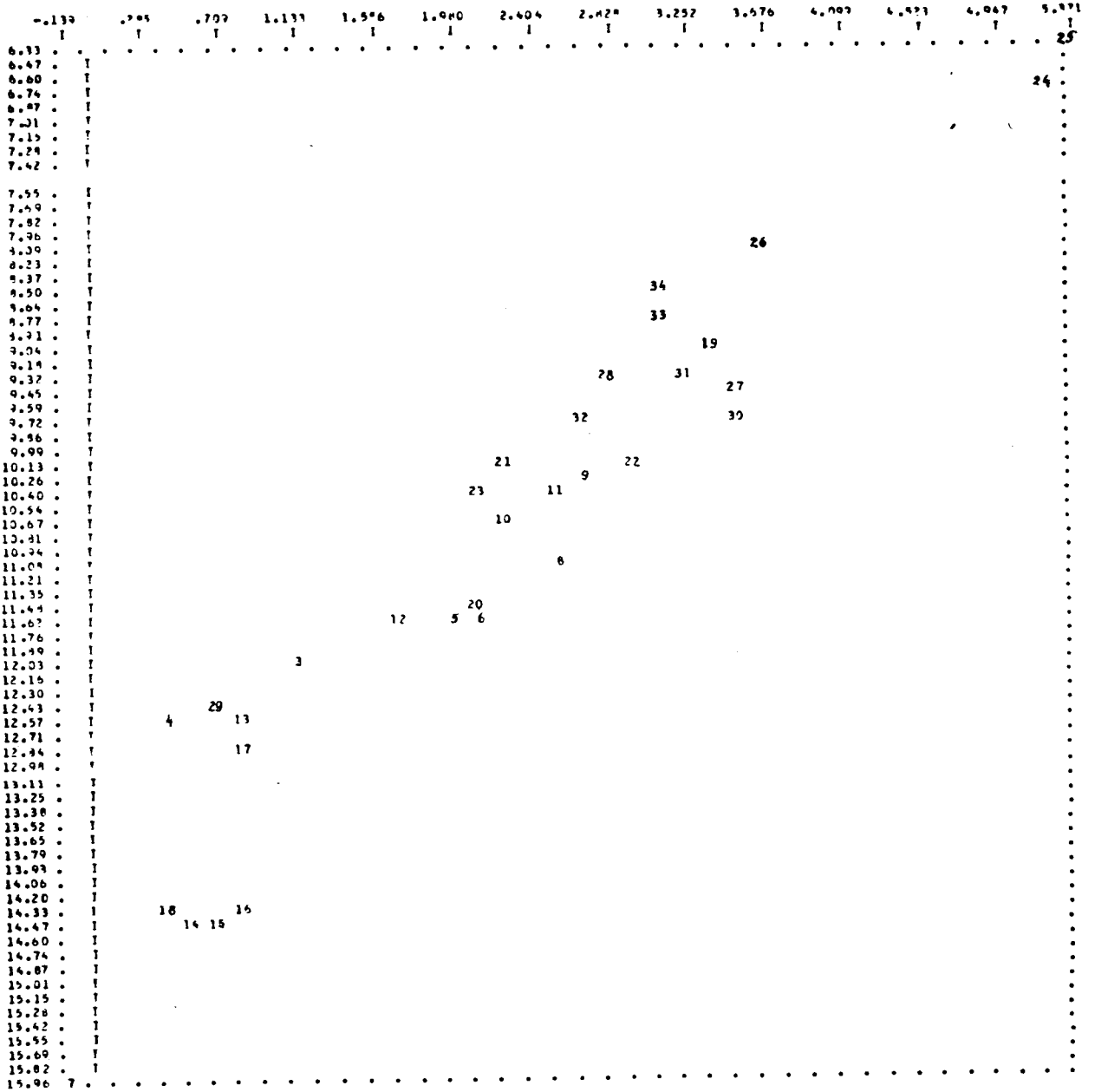
Projection on 3 & 4 features plane



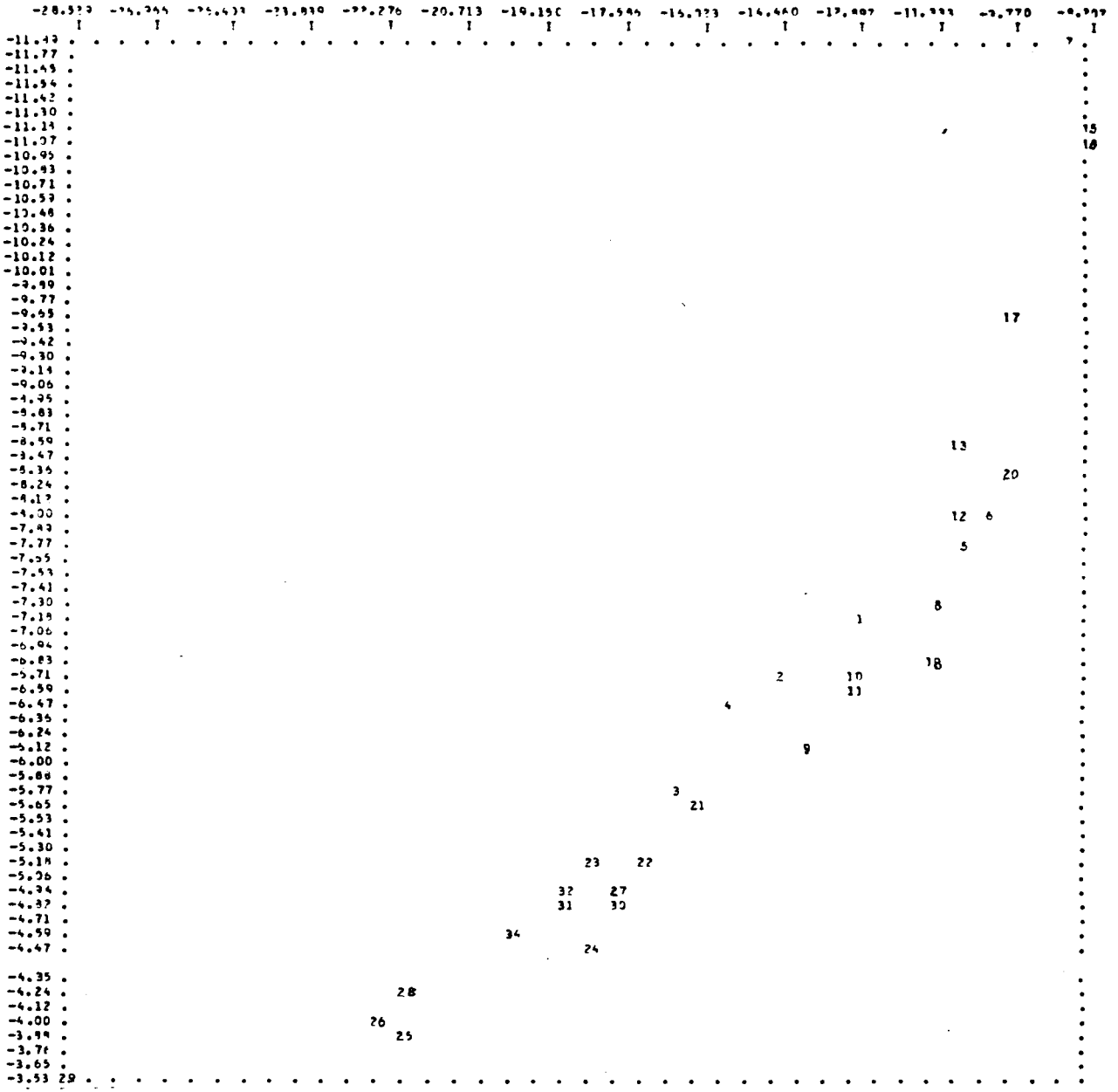
Projection on 5 & 6 features plane



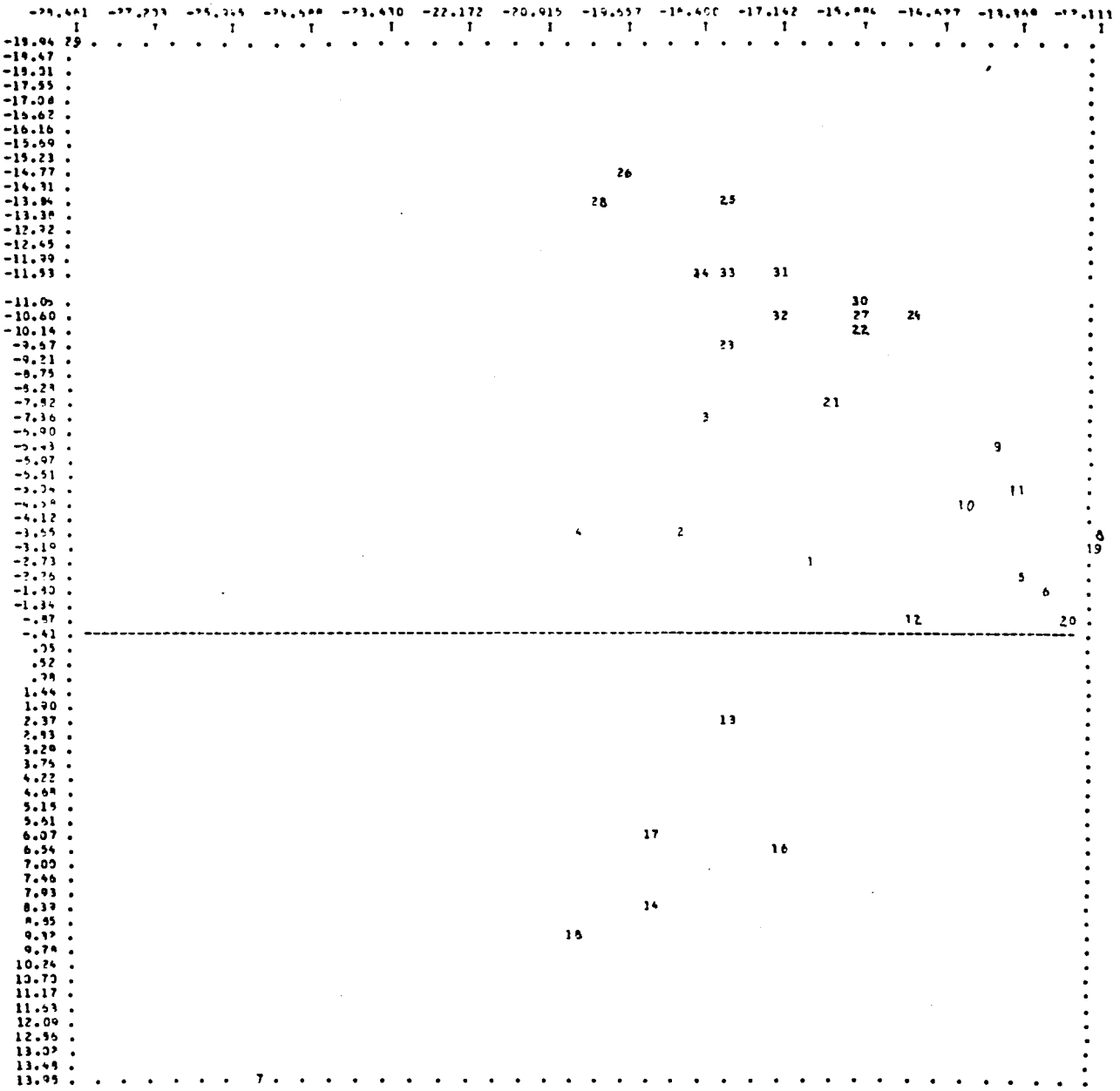
Projection on 7 & 8 features plane



Projection on 9 & 10 features plane

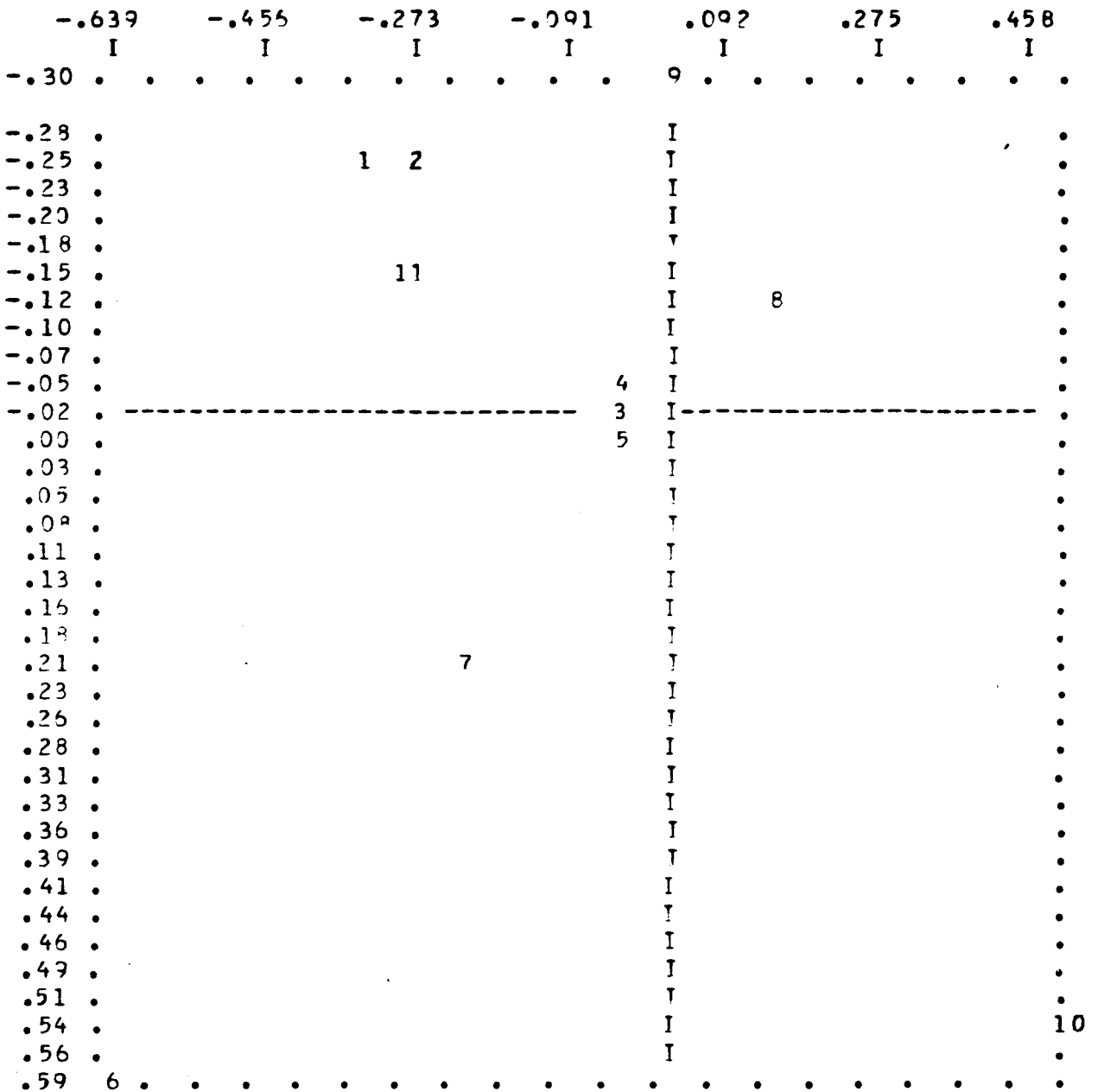


Projection on λ_1 & λ_2 plane



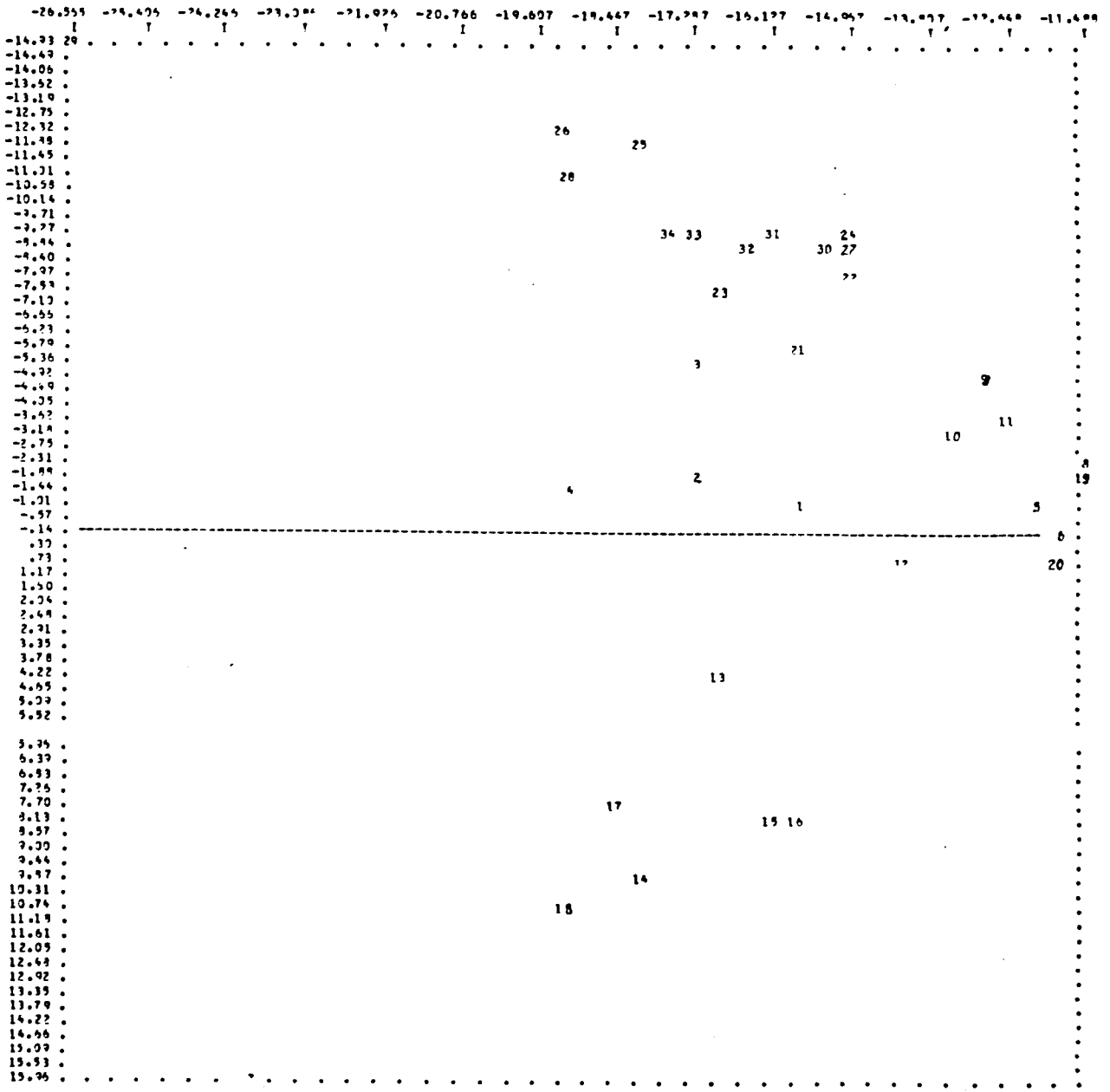
Appendix 3

Features clustering on 1&2 principle components plane

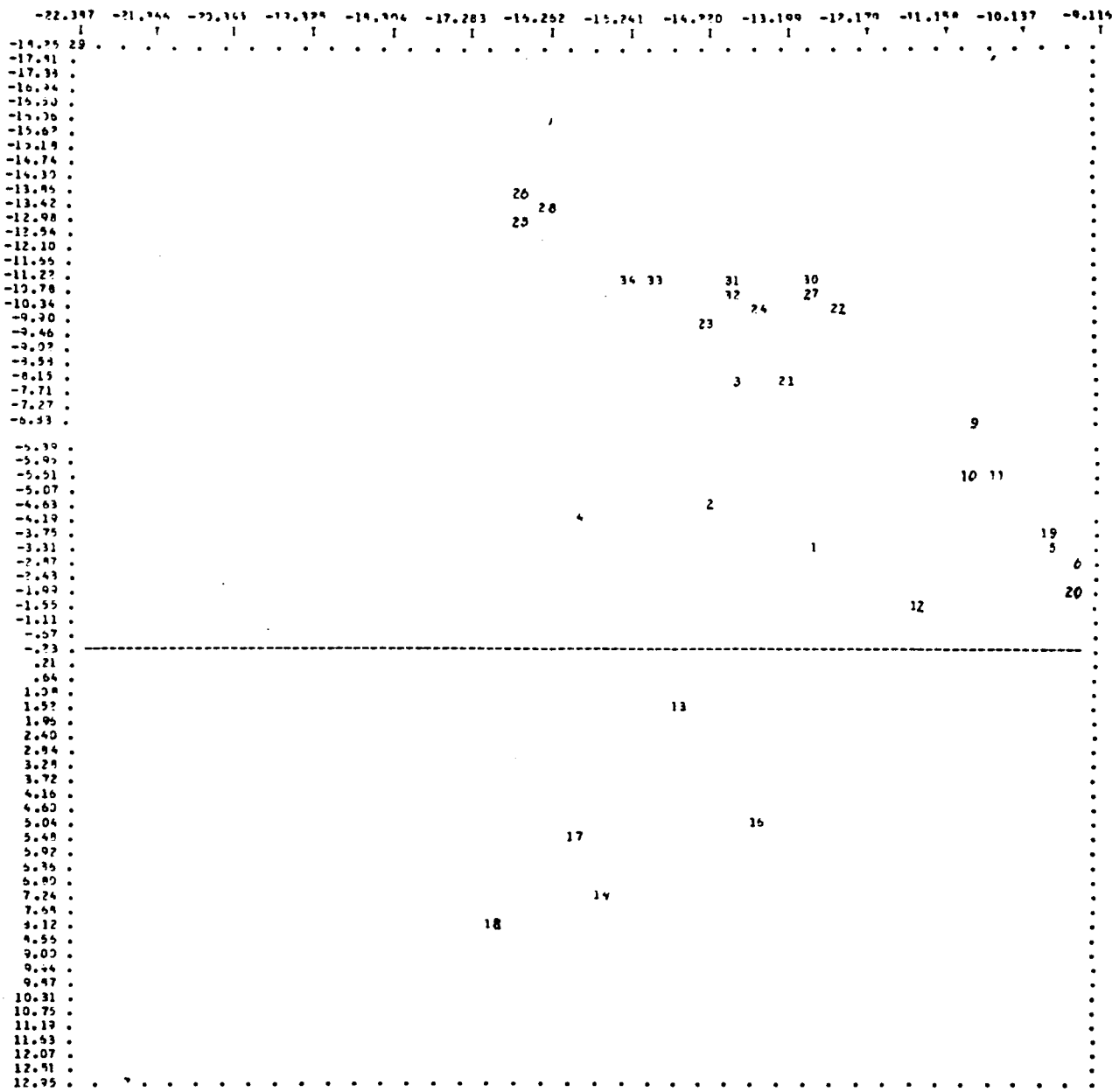


Appendix 4

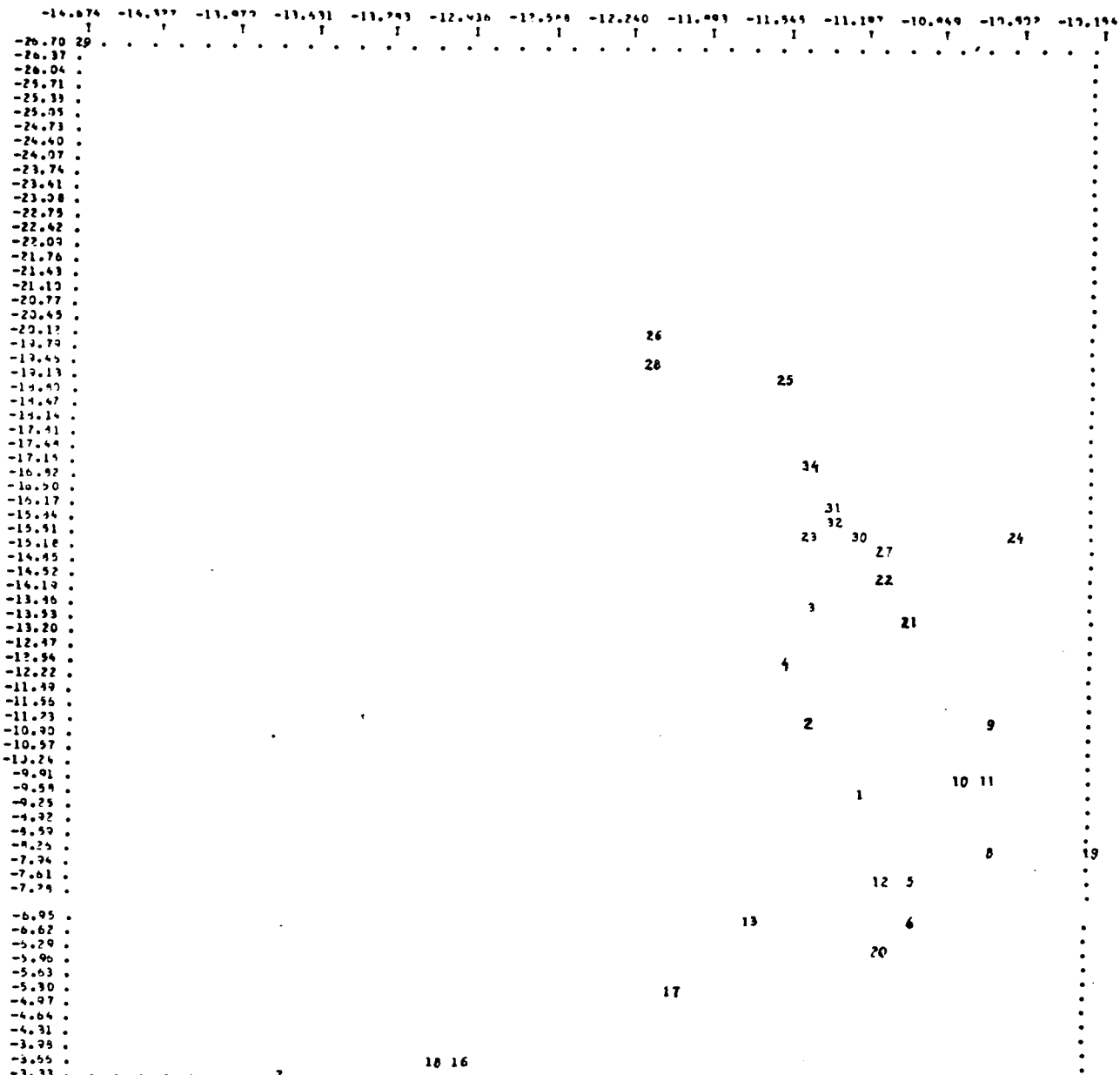
Projection plane $\lambda_1 - \lambda_2$ for feature set: 1,3,6,7,8,9,10



Projection plane $\lambda_1 - \lambda_2$ for feature set: 2,5,6,9,10



Projection plane $\lambda_1 - \lambda_2$ for feature set:
2,4,9,10



Appendix 5

Hierarchical cluster-procedure for the set of 11 features

1) Nearest neighbour variant — Euclidien distance

Step 23

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

2) Mean connection variant — Euclidien distance

Step 27

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

3)Further neighbour variant — Euclidien distance

Step 24

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

Hierarchical cluster-procedure for the set of features: 1,3,6,7,8,9,10

1) Nearest neighbour variant — Euclidien distance

Step 25

class 1: 1 2 4 5 6 8 9 10 11 12 19 20
2: 3 21 22 23 26 27 28 30 31 32 33 34
3: 7
4:13
5:14 17 18

```
6:15 16
7:24
8:25
9:29
```

2) Mean connection variant — Euclidien distance

```
Step 24

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

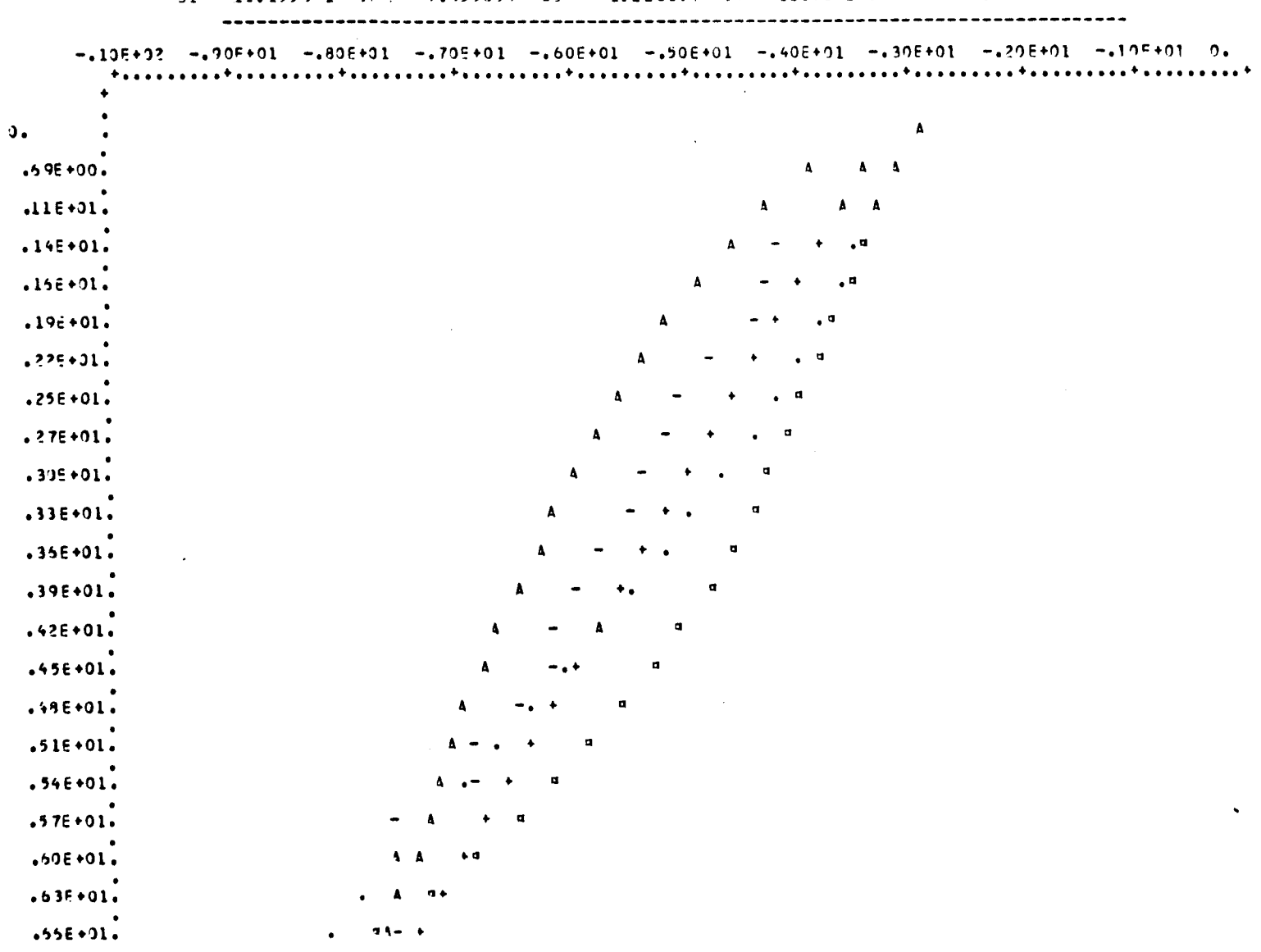
3) Further neighbour variant — Euclidien distance

```
Step 25

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

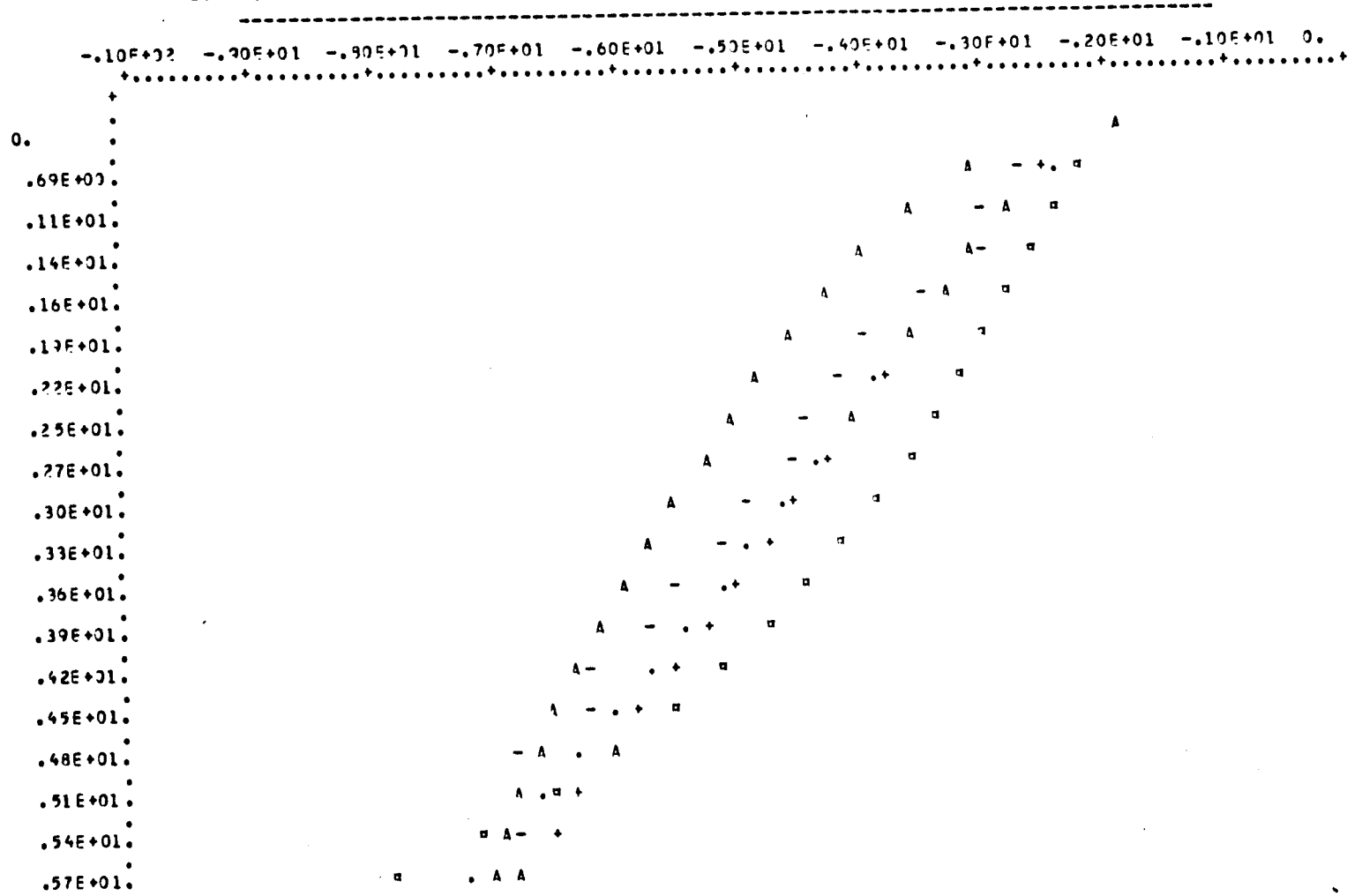
RANK FREQUENCY DISTRIBUTION OF WORDS IN TEXT
A.S.PUSHKIN. THE BRONZE HORSEMAN
TEXT CHARACTERISTICS - TEXT VOLUME Z= 2040 WORDS
VOCABULARY VOLUME V= 962 WORDS

F1= 113.000000 BET= .688219 B= 2.318831 K= .211533
B1= 12.155391 B2= -5.455604 B3= 1.220084 B4= -11.734106 B5= 50.000000



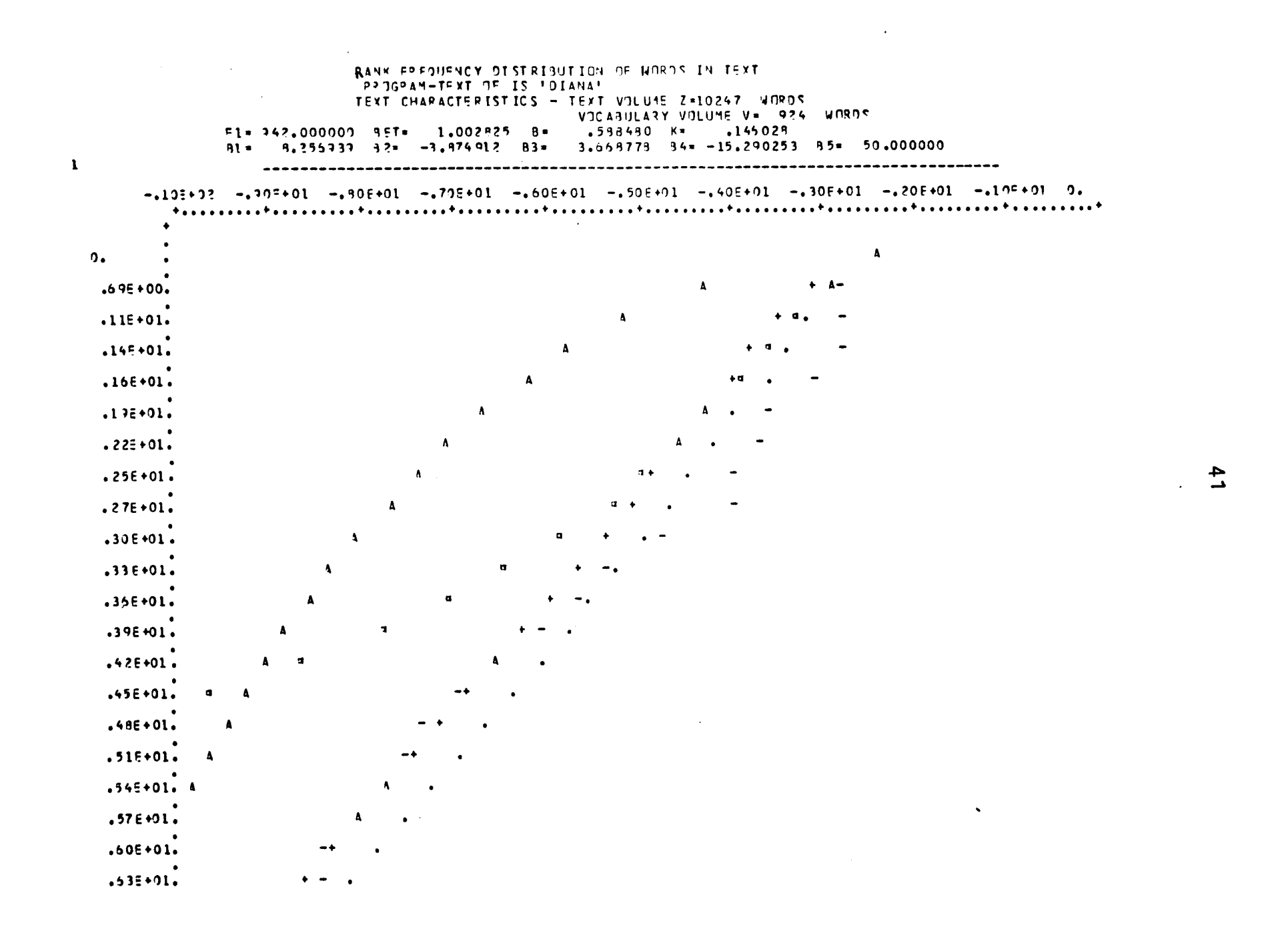
RANK FREQUENCY DISTRIBUTION OF WORDS IN TEXT
DIALOGUE WITH IS 'DIANA'
TEXT CHARACTERISTICS - TEXT VOLUME Z= 1037 WORDS
VOCABULARY VOLUME V= 311 WORDS

F1= 153.000000 BET= .887445 B= .251383 K= .196319
B1= 10.305500 B2= -5.027385 B3= 2.999515 B4= -11.797792 B5= 50.000000



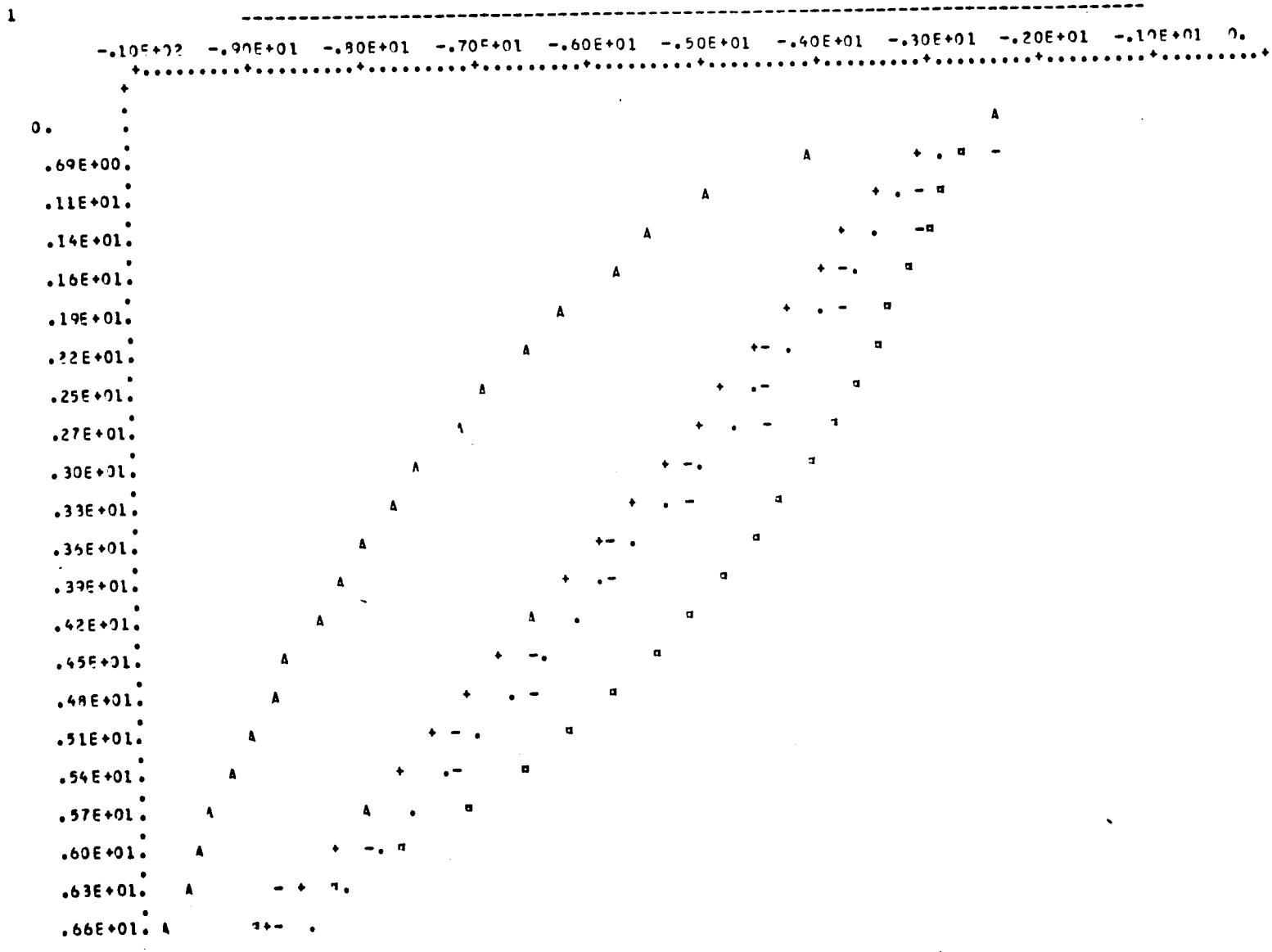
RANK FREQUENCY DISTRIBUTION OF WORDS IN TEXT
PROGRAM-TEXT OF IS 'DIANA'
TEXT CHARACTERISTICS - TEXT VOLUME Z=10247 WORDS
VOCABULARY VOLUME V= 924 WORDS

B1= 342.000000 BET= 1.002825 B= .588480 K= .145028
B1= 8.255939 B2= -3.974912 B3= 3.668778 B4= -15.290253 B5= 50.000000



RANK FREQUENCY DISTRIBUTION OF WORDS IN TEXT
BINARY TREE WITH 256 NODES
TEXT CHARACTERISTICS - TEXT VOLUME Z=22529 WORDS
VOCABULARY VOLUME V= 2048 WORDS

F1=2048.000000 BET= 1.000000 B= .442759 K= .131154
B1= 11.046361 B2= -4.114899 B3= 2.128439 B4= -14.548936 B5= 50.000000



Contents