Now, when the application of mathematical methods and the use of computers are greatly spreading in linguistic analysis, it will not be useless perhaps to give a short summary of the examined fields and the different aspects of investigation.

What is to be examined and by what methods ? – we will try to find an answer to this question.

The examined field will be clearly outlined if we take the possibility of a twofold meaning of the language system into consideration. This 'term' may stand not only for the system of sentences, but of texts/'messages' that are larger than sentences, and have a full value in themselves/, too.

Graphically :

Explanation	Basic unit	Language system
1.	sentence	system of sentences
2.	text	system of texts

Between the two systems a close connection is meant by the fact that the basic unit of the text is the sentence. /

Thinking of the first explanation we generally speak of the analysis of the language, referring to the second one we speak of the analysis of texts.

Traditional linguistics and most of the structural theories and the generative linguistic theory is also sentence- centered. The text-theory that develops the sentence-centered linguistic theory further and contains it somehow is a more characteristic common research field of general linguistics, documentation / the theory of scientific texts / and physology / the theory of literary texts / than the analysis of sentences. The different text types differ from each other much more than the sentences used in these text types.

For approaching a language or a text we can set up the following classes of methods[1] :

	structural	quantitative
1.	-	-
2.	-	+
3.	+	-
4.	+	+

1. On the analysis of 'language'

Let us see the questions of the analysis of 'language' first.

1.1 On the non-structural non-quantitative level of the analysis of language our knowledge concerning the language structure does not get an explicit drafting, there are no systematic investigations for defining the usage of the language elements and structures. The traditional linguistic research – that are, on the other hand, accumulating a very valuable material – are mostly of this character.

1.2 At first the systematic analysis was focussed on the usage of the language. The definition of values of different statistical characteristics belongs to the quantitative non-structural investigations.

For example we can measure the length of syllables, morphemes, word-forms in phonemes; the length of morphemes or word-forms in syllables; the length of sentences in morphemes word-forms or constituents and we can examine the frequency of the so-defined elements.

Besides these to this level belong the investigations that are focussed on the frequency of certain language elements / e.g. certain words, loan-words or key-words /, or that of certain classes of the elements – not forming a complete system – / e.g. certain verb or noun types / concerning a given language.

This kind of the so-called statistical description must be followed by a statistical analysis / setting up and controlling of hypotheses, discovering possible correlations between certain variables of probability / and statistical prediction if possible / drafting of conclusions of general character – effective at given conditions – concerning the whole examined corpus /.

To this first it is necessary to work out such a type of mathematical statistics that is suitable for the examination of the complete language system,

and on the other hand the language system itself has to be described sufficiently.

1.3 In my opinion the adequate structural description of the language system is provided by the generative linguistic theory.

The generative grammar elaborated by Chomsky gives a sentence-centered description of the language.[2] Using finite number of categories / those of syntax and form-classes; number-, case-, mood-, time- and person- indicators / with help of finite number of rules and a lexicon the syntactic component generates the structures of all the grammaticaly correct sentences. A part of these rules refers to the generation of different sentence structures, others to the transformation of these structures. The semantic and phonological components order semantic and phonological interpretation/s/ to the sentences, according to their syntactic structures.

1.4 The knowledge of the linguistic structures makes it possible for the structural-quantitative investigations to be focussed on well-defined sets and for the attained results to be well-situated in the system of the accumulated knowledge concerning the language elements and structures.

The inquirable elements and classes can be summarized as follows :

	Elements / concrete verbal units /	The set of elements	Classes	The set of classes
1.1	phonemes	finite	phonemes classes classes formed on the basis of some point-of-	finite

			view/e.g. the definitive subsets of the distinctive features /	
1.2	phoneme-chains		classes of phoneme-chains	
	digramms	finite	classes of digramms	finite
	trigramms	finite	trigramms	finite
	syllables	bounded	classes of syllables	finite
2.1	morphemes		morpheme classes	
	roots	bounded	classes of roots	finite
	formants	finite	formants	finite
	suffixes	finite	suffixes	finite
2.2	morpheme-chains		classes of morpheme-chains	
	lexical entries	bounded	classes of lexical entries formed on the basis of some point-of-view / syntactic, semantic, phonological categories; origin, etc. /	finite
	word-forms	bounded	different paradigmatic classes	finite
2.3	morpheme-groups		classes of morpheme-groups and constructions	
	couples	bounded	classes of couples	finite
	triplets	bounded	triplets	finite

occurring in
grammatically
correct
utterences

3. According to the different levels of the generating process we can speak of the following structures and classes of structures :

3.11	simple deep structures	bounded	classes of simple deep structures according to the different depths of categorization	bounded
3.12	semantic interpretations	bounded	classes of semantic interpretations	bounded
3.13	surface structures	bounded	classes of surface structures	bounded
3.21	compound deep structures	bounded	classes of compound deep structures according to the different depths of categorization	bounded
3.22	semantic interpretations	bounded	classes of semantic interpretations	bounded
3.23	surface structures	bounded	classes of surface structures	bounded

The term 'bounded' means that the maximum number of the elements of the given set can be defined combinatorically / i.e. it is not infinite /, but

the number of the really existing elements of the set in a given language and at a given moment is not to be defined exactly.

The term 'simple' in point 3. refers only to the deep structures containing hypotaxis, the term 'compound' refers to the deep structures that came to existence by parataxis.

We can speak of quantitative-structural investigations when we define the frequency of all the elements constituting the structure or all the classes concerning the given language, and we weight the elements of the structure with these values / e.g. we define the complete system of the classes of formant-morphemes and the frequency of the separate classes /.

These are the lowest language levels where we can speak of system of elements. On higher levels only of the system of classes. / E.g. the set of the lexical entries of the language cannot be considered a 'complete system' even as a synchronic system. So when compiling the frequency dictionary of the language we can speak only of an approximative completeness. /

We have to mention that up to a certain level the system forming the basis of the frequency investigations itself may also be analysed statistically. In the case of a listable complete system the results are directly characteristic of the examined language. But the results gained on the basis of the statistical examination of a vocabulary of a given language are characteristic only indirectly and with some restrictions.

The results of the quantitative investigations in connection with the system of a given language can be considered relevant both linguistically and mathematically only if they are established on the basis of the sample taken suitably from the corpus and if the limits of error are displayed everywhere.

1.5 Where and how do mathematics and the computer join this analysis process ?

1.51 Of course not on the level of non-structural and non-quantitative analysis.

1.52 On the level of the non-structural quantitative analysis beside simple counting, a more complete apparatus of mathematical statistics can be employed already. Over a certain number of elements even the more simple counting is worth to be done only with a computer. / In certain cases – when defining the frequency concerning a big corpus, or a factor-analysis considering a lot of factors – it is not even possible ! /

1.53 On the level of non-quantitative structural analysis mathematics appears in a more hidden form. The generative linguistic theory is basically related to the automaton-theory / the abstract algebraic theory of the automatons /.

Here we can employ the computer with a double reason. For generating – for 'controlling' the compiled linguistic rules – and for analysis – for discovering the structure of the concrete language units. The experiments completed so far prove that the right strategy is to separate the program of the computer operations from the linguistic rules. Thus we work out an universal algorythm that can operate any grammar even if it is built up the exact way.[3] / Operating such an algorhythm naturally claims not only the suitable compiling of the linguistic rules but a suitably compiled lexicon, too. /

1.54 On the level of quantitative structural analysis both the abstract algebra and the whole apparatus of the mathematical statistics play a role.

With making use of the experience and results of this letter one it is necessary to work out the statistical theory of linguistics.[4]

It is impossible to imagine a successful quantitative structural analysis with our electronic computers. The programs – and the linguistic rules and lexicon as a basis of the work of the programs – has to be built up the way that when employing them, the texts to be processed should not have to be prepared / providing insertions of special codes /. It it is possible to reach this goal, the text-analysis and the automatic text-production can interlock organically, as the bands of the automatic type-setting can be directly used for the analyses, too.

To a complete description of the system of a language not only the compilation of its rules and lexicon belong but also the statistical descriptions reflecting the use of its elements and structures. / We call the latter ones 'etalon-statistics' /.

We can sum up the expected results of the different analyses in the following way :

The analysis		The results of the analyses	
str.	qu.	in reference to sentences	in reference to the system of sentences
-	-	qualitative establishments	
-	+	-	the definition of the frequency of certain elements, of the value of certain characteristics

+	-	the syntactic semantic phonological description of the sentence structures	compilation of different /explanatory, synonimic, conceptual / dictionaries; compilation of a system of rules that is capable to generate all the grammatically correct sentences of a language; elaboration of an exact method for discovering the structure of sentences
+	+	-	the definition of the system of different language elements / not larger than sentences /, complete, if possible, weighted by the values of frequency; frequency dictionaries; typologies;

2. On the analysis of 'texts'

2.1 On the non-structural non-quantitative level of the analysis of texts

our knowledge of the construction of certain texts and the system of texts of different types will not have an explicit drafting. On this level we speak only generally of a literary or scientific work, or of the literary or the scientific language, mentioning mostly the more remarkable features.

2.2 On the level of non-structural quantitative analysis the systematic investigation is focussed on the usage of certain language units, too. But while at the statistical investigation of the language we have always spoken of some characteristic of the 'language', or the frequency concerning the whole 'language' of some system of the elements, here we inquire the characteristics of the individual texts or those of the poetic or scientific language.

2.3 The next is the non-quantitative structural level of texts. The text-structure is similar to the sentence structure but it also shows features different from that. Its analysis is just being developed.

In my opinion the creation of a model which would be able to generate correct 'texts' impossible. Thus the investigation can be focussed only on the elaboration of such analysing processes that are able to 'discover' the structure of given texts.

Till now the analysis was mainly statistical. After the analyses directed to the establishment of the characteristics of certain language types we meet more statistical investigations the purpose of which is the definition of the structural character of one single work-of-art.[5] / This point-of-view could not have come up concerning one single sentence. The length of a sentence measured in different kinds of units says very little about the structure of a sentence ! /

Far fewer experiments are known focussed on the discovering of the so-called hierarchical construction of the text structure; on the discovering of how the separate parts of the text are connected with each other and with each other and with the whole text itself; through what hierarchical levels and what functions are filled in the creation of the whole text structure.[6]

Untill further progress is made in science it is impossible to analyse the different kinds of text-structure-systems./ We may consider such systems the complete works of a writer, the system of the works belonging to a literary form or to an age of the history of literature, etc. /

2.4 The quantitative structural analysis of the language system as a system of text structures expects us to build the language system on larger text units than sentences. For this we have to know what directly higher composition units can come to existence from the sentences and how and through what composition levels can these composition units widen further.

It is this level on which the various kinds of text structure typologies can be formed, that, referring to the verbal works of art, will be an organic part of the theory of literary forms.

2.5 As for the employment of mathematics and the computers, we can say the following.

2.51 Their employment on the level of non-quantitative and non-structural analysis is out of the question.

2.52 At the quantitative non-structural analysis the whole apparatus of the linguistic statistics is used. To what was said in connection with the

quantitative structural analysis of the 'language' we have to add that it is necessary to pay a greater attention to the problems of the statistical analysis of the finite corpuses when analysing a text strucutre. Beside the definition of the values of frequency, the correlation-counting and the so-called type-token investigations play a more important role here.

The employment of computers does not mean a new problem on this level.

2.53 On the level of the structural investigation let us see first the question of the analysis of the individual text structures. We have to work out the strategy of both the primary structural and the primary statistical discovery. What special mathematical and computer problems will arise here we cannot tell yet.

The primary structural approach probably has to precede the statistical one. This discovers the units of structure, referring to which we shall define the frequency of the various language elements. Beside the possible grammatical connections between sentences or sentence-like elements the primary structural approach has to make use of the synonymous and thesaurus-like connections between the lexical units of the separate sentences, too. The experiments accumulated till now at the compilation of the thesauri of documentation will also mean a significant help to the building of a language thesaurus that satisfies this purpose, too.

The primary statistical approach makes some more machine experiments necessary.

We can imagine the structural analysis of the system of text structures only from the material that has been accumulated by the traditional philology in the analyses of a writer's life-works.

			the development of a system to characterize texts, on the basis of which we can create different text typologies; the elaboration of an exact method for discovering the characteristics of different texts;
+	+	the structural-quantitative description of text-structures; compilation of indices concordances 'other dictionaries' referring to different language units of a given text;	the definition of the system of different language elements / larger than sentences /, complete, if possible, weighted by the values of frequency; frequency dictionaries of certain 'text-types'; text-typologies;
		+ □ +	

As we have seen, linguistics, mathematics, documentation-theory, phylology, automatic text-producing and the use of computers are the contributors of a strategy focussed on one purpose for our analysis of the text structures and their systems. To the demonstration of these connections see fig. 1.

2.54 For the quantitative structural analysis first we have to work out the system of the intensional characterization of the separate text structures. This has to be a natural expansion of the characterization of elements of lower levels. Just as we characterize the phonemes by distinctive features, the words by syntactic and semantic markers, so we have to define the methods for characterizing the individual sentences, individual composition units and individual works-of-art. On this base can the various text-structure classes be defined.

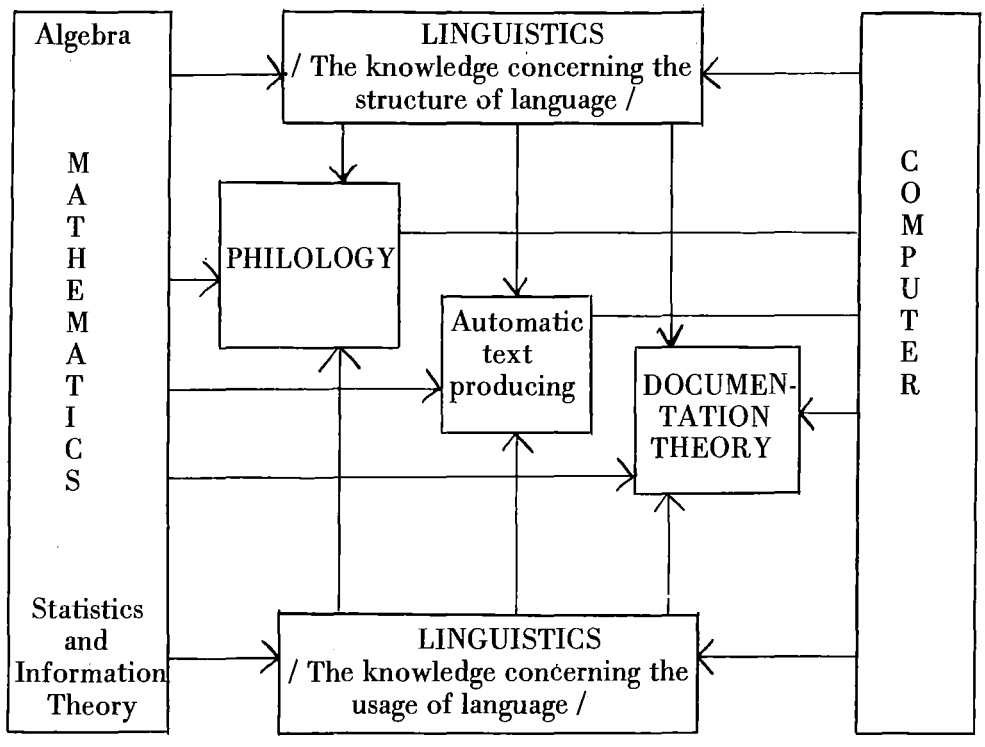
Here the computer will play a role similar to the sorting and classifying of the 'word-cards'.

The expected results of the different analysis we can sum up in the following way :

The analysis str.	qu.	The results of the analyses in reference to the single texts	in reference to the different systems of texts
-	-	qualitative establishments	
-	+	the definition of certain characteristics referring to the given text;	the definition of certain characteristics / referring also to language units larger than sentences / in reference generally to the language or to some 'language type' / scientific, literal, etc. /;
+	-	structural description of a text-structures;	compilation of different dictionaries;

			the development of a system to characterize texts, on the basis of which we can create different text typologies; the elaboration of an exact method for discovering the characteristics of different texts;
+	+	the structural-quantitative description of text-structures; compilation of indices concordances 'other dictionaries' referring to different language units of a given text;	the definition of the system of different language elements / larger than sentences /, complete, if possible, weighted by the values of frequency; frequency dictionaries of certain 'text-types'; text-typologies;
		+ □ +	

As we have seen, linguistics, mathematics, documentation-theory, phylology, automatic text-producing and the use of computers are the contributors of a strategy focussed on one purpose for our analysis of the text structures and their systems. To the demonstration of these connections see fig. 1.



This figure demonstrates the relation between the theory of grammatics and some other basic disciplines as well as one of the fields of applied linguistics.

In my opinion within this framework given in this paper the trends of the further linguistics investigations show themselves, too.

NOTES

1. Cf. H. Spang-Hanssen, Mathematical Linguistics - A Trend in Name or in Fact ? in : Proceedings of the Ninth International Congress of Linguists. Cambridge Mass. 1962. The Hague, 1964. 61-71.

2. N. Chomsky, The formal nature of language in : E. H. Lenneberg, Biological Foundations of Language, New York, 1967.

3. Such an independent computer program was wrought out by Dénes Varga in our Computing Centre.
   D. Varga, Problems of Machine Analysis, Linguistica Antverpiensia 1968. 2. 415-427.
   D. Varga, Postroenie novoj analizirujushchej sistemy predlozhenija, Nauchno-technicheskaja inforçacija, ser 2. 1968. 4.

4. The works of Yule, Guiraud, Herdan came to existence with this demand.

5. Cf. : K. Kroeber, Computers and Research in Literary Analysis, in : Computers in Humanistic Research / Ed. by E.A. Bowles /, New Jersey, 1967. 135-143.

B. O'Donnell, Stephen Crane's The O'Ruddy : A Problem in Authorship Discrimination, in : The Computer and Literary Style / Ed. by J. Leed / Kent, Ohio 1966. 107-116.
Chr. et Cl. Allais, A method of structural analysis with an application to 'Les liaisons dangereuses', in : Revue, International Organization for Ancient Languages analysis by Computer L.A.S.L.A. 1968. 2. 13-32.

6. In the following essays, my purpose was the investigation of the special patterning of the verbal work of art :
On the Structural Linguistic Analysis of Poetic Works of Art, in : Computational Linguistics VI. / Budapest / 1967. 53-82.
Notes on the Semantic Interpretation of Verbal Works of Art, in : Computational Linguistics VII. / Budapest / 1968. 79-105.

I exerted myself to an explanation of the structure of the verbal work of art that considers the complexity of literary texts and at the same time serves as a base for the analysis by computer, too. The application of the model for computers is being wrought out.

János S. Petőfi

Computing Centre of the
Hungarian Academy of
Sciences.