COMPUTATIONAL LINGUISTICS

VIII

COMPUTING CENTRE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST, 1969

COMPUTATIONAL LINGUISTICS
VIII.

COMPUTING CENTRE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST, 1969

Editorial Board

Bálint Dömölki, Ferenc Kiefer (editor), György Szépe, János Szelezsán,
Éva B. Szöllősy (technical editor), Dénes Varga

Felelős kiadó: Balázs János

703507 MTA KESZ Sokszorosító. F. v.: Szabó Gyula

Contents

Reviews and Miscellanous

MULTI - INDEX SYNTACTICAL CALCULUS

Hans Karlgren*

Introduction

In our work on analyzing Swedish nominal phrases as they appear as document titles - particularly titles of articles in periodicals - we have primarily utilized context-free rules. In an endeavour to reduce the cumbersomeness of such rules, we have used the notation:

(1) $a_{xy}\ b_{xy} \rightarrow c_{xy}$ for x = p, q, r and y = u, v

as a shorthand for six substantially similar rules. The gain is not merely that of avoiding scrivener's palsy - and puncher's impatience, since the analysis program also accepts this shorthand - but also that of clarifying the parallelism between the rules. The rule schema reads "a syntagm of type a combines with one of type b to form one of type c, each being respectively of subclass p, q or r and u or v". If the subscripts are interpretable as linguistic categories,

---

*KVAL, Fack, Stockholm 40.
The work reported in this paper has been sponsored by
The Bank of Sweden Tercentenary Fund and
The Swedish Humanistic Research Council

this notation seems quite natural. We might write a fundamental rule of Latin grammar, by way of illustration, thus

$$adj_{ngc}\ nom_{ngc} \rightarrow nom_{ngc}$$

which would mean that to a nominal group may be joined an adjective of the respective number, gender, and case without changing the syntactical category of the group.

This notational little device actually often reduces the intuitive need for context-sensitive rules, since it performs what these rules are required to do in the domain where we have a choice, namely to bring out the common pattern and leave aside for later consideration the minor adjustments.

Now, in practice, we have for each word or syntagm not one subscript but a set of alternative subscripts. On the initiative of Gunnar Ehrling,* who wrote the analyzer, we further reduce the notation by giving a name to all such sets of alternatives and by specifying in a "multiplication table" the name of the set of alternatives forming the intersection between any pair of such sets. Thus, in place of (1) our rules actually read

(2) $$a_{ik}\ b_{jl} \rightarrow / c_{i\cap j}\ .\ k\cap l$$

where the values of $i\cap j$ and $k\cap l$ are taken from the "multiplication table."

We now ask what will happen if we generalize this index "multiplication" so that it will represent not intersection of index sets but an arbitrary binary operation on the set of index symbols. Particularly, we are interested in the case where this multiplication is non-associative and the set of index symbols is not closed under multiplication. This would mean that the restrictions imposed by the indexes on the sentence or part thereof could, in their turn, be written as a context-free - not a finite-state - grammar over the index symbols.

---

*KVAL, Interim Report No 13,
Program för grammatisk analys av texter

When the subscript multiplication rules are generalized so far, they are of the same kind as the "multiplication" on the main level, and we prefer to write a' i' k for a_{ik} and we define multiplication of such index vectors as "inner" multiplication, that is, the corresponding elements are multiplied:

$$a' i' k\ b' j' 1 \rightarrow ab' ij' k1$$

We note that, in general, these rules cannot be reduced to a finite list of common context-free rules, as could rules like (1) and (2). For if we can replace ab by c, we may well be unable to replace ij by anything shorter than ij, the multiplication table being blank for ij or even having no row i or column j, since i and j may, in turn, be strings and not elements in the index set. And if the well-formed sequences of indexes are defined by a general context-free grammar and not by a finite-state one, we cannot remedy this by adding more symbols to the index set: the set of triples i, j, ij may then be infinite.

This paper is an attempt to investigate this problem, elaborating such a multi-index calculus a little. First, however, we may be excused for making a summary of the background of the recognition grammar problems for which such a calculus may be useful. The reader who expects to be bored by such a survey should turn directly to page 14 below.

Reduction systems

We introduce some definitions. The terms employed largely coincide with those of current generative linguistics, but some minor adaptions have been made to make the terms adequate for describing the kind of recognition grammars with which we are concerned.

We consider strings over an alphabet $S = \{a, b, c, ..\}$. We write ab for the string formed by concatenation of two letters a and b, and $\alpha\beta$ for the concatenation of two strings α and β. Concatenation is considered a reflexive, associative but not commutative relation.

We write M for the set of all concatenations of strings in a set M:

$$M^* = M \cup \{ \mu\ \mu^* \mid \mu \in M,\ \mu^* \in M^* \}.$$

A rewriting rule, $\alpha \to \beta$ is a rule which permits us to replace the string α in any string where it may occur by the string β. A reduction rule is a rewriting rule which does not increase the number of words in the string. A reduction system is a set of reduction rules:

$$R = \left\{ \alpha \to \beta \mid \alpha \in a_1\ a_2 \ldots a_n,\ \beta \in b_1\ b_2 \ldots b_n,\ \ a_i \in S,\ b_i \in S,\ m \leq n \right\}$$

By means of R we can define a derivability relation over S^*. We say that α is reducible to β, $\alpha \to \beta$, according to R, if there is a succession of applications of rules in R by which α can be rewritten as β. We include the case where no rule is applied so $\alpha \mapsto \alpha$ for all α. Thus, "$\mapsto$" is a reflexive and transitive relation.

We now define a reduction grammar $G = < S, R, I, T >$ as a specification of a set of strings, a language, over an input alphabet $I \subset S$:

$$L = L\ (<S, R, I, T>) = \left\{ \sigma \mid \sigma \in I^*, \sigma \mapsto \tau \in T \subset S^* \right\}$$

where T is a set of - terminal or, to avoid diametrically opposite associations - target symbols. We say τ is an R-reduction of σ.

Finite Rewriting Systems

Constituent structure grammars and grammar components

We first consider grammars where S is a finite set. We call these grammars constituent structure grammars.

If T contains one single element, say s for sentence, the grammar is a decision grammar, which specifies for each input string whether or not it is grammatical.

Trivially, T can be extended to include a few elements, say s for statement, q for question, and so on. Naturally, we can reformulate a grammar with $T = \{t_1, \ldots, t_n\}$, where n is finite, into a grammar with a unique target element, merely by adding one element, say s, to S and incorporating a few rules $\{t_i \rightarrow s \mid i = 1, \ldots, n\}$ to R.

However, allowing T to be an infinite set is not necessarily a trivial extension.

Trivial but occasionally practical is to define a language $L(S, R, I, A^*)$ where the targets are all the strings over an output alphabet $A \subset S$.

If T is some non-trivially defined subset set, L' of strings over a subset A of S, we have

$$L = L(S, R, L\ L')$$

where L' must be defined by some grammar $G^1 = \langle S', R', A, T \rangle$ We say that $G'' = \langle S, R, I, A \rangle$ is a grammar component and note that G'' and G' together completely specify L. We shall come back to this concept later when we describe more complex grammars as combinations of simple ones.

With the restriction imposed on the rules of R that the right hand side should never be longer than the left hand side, it is obviously always possible in a finite number of steps to decide whether or not a given finite string is reducible to some element in T, i.e., whether or not it is an element in the set L. For if the given string σ contains m symbols and S contains n different symbols, σ can be shortened at most (m - 1) times and after the i:th time it has been shortened, (i = 0, 1, ..., m - 1), it can be rewritten without shortening at most $(n^{m-i} - 1)$ times without being rewritten as σ, which can always be avoided by keeping a finite record of historical information.

Disjoint constituent grammars

1. A reduction rule where the right hand side contains exactly one symbol is called a context-free rule. If all the rules are context-free we say the grammar and the language is context-free.

If the grammar is context-free we may give it the following interpretation. Let the letters of I be sets, "categories", of strings of linguistic signs. Let ab mean the set of strings consisting of one string contained in category a followed by one contained in b. Let the reduction rules mean inclusion so that, e.g., $ab \subset c$ means that the set ab is included in the set c.

A string σ over I then represents a grammatical sentence of type t, if and only if, $R \Rightarrow \sigma \subset t \in T$.

2. A context-free constituent grammar, then, can be adequately described as a classificational system with finer and broader terms where all classes can be written as concatenations - interpreted as the set of concatenations of the cartesian products - of a finite set S of categories. The process of analyzing sentences of such a language can be performed as a classificational procedure and the result is adequately and exhaustively statable as the class adherence of sets of successive substrings, representable, e.g., by a tree with no crossing branches.

One may note that the character of a context-free language well conforms with what used to be defined as agglutinative languages, that is with the agglutinative languages as they were commonly defined, not as any existing natural language of any particular group.

The assumptions behind an attempt to describe a real language by a context-free grammar, therefore, are very strong. It is not astonishing that these attempts partially fail; it is astonishing that they have carried as far as they have. For instance, there is no convincing empirical evidence that a decision grammar for a natural language cannot be written as a context-free grammar, though there are ample theoretical reasons not to stake too much on the prediction that no practical counter-examples will turn up in the future.

3. If we add to our context-free grammar rules of the type

$$ab \rightarrow bc$$

or, generally, permutation rules where the same elements recur on the right, though in different order, we broaden, of course, the family of languages under considerations and the interpretation above under 2. no more holds true. But all what was said about the highly specialized character of the languages remains true, except that class adherence is now not confined to sets of successive substrings; the language is characterized by the existence of discontinuous constituents, and except that the tree drawn will have crossing branches here and there. But it is still possible to assign each substring to exactly one immediately higher order constituent and it is still possible to draw a tree.

We may summarize the constituent
so far mentioned under the name disjoint-constituent grammars, i.e., grammars where each constituent is either disjoint from or included in another and where, accordingly, the constituents can be defined as a hierarchial set of equivalence classes over the substrings of the given input string.

Such a classification of substrings is called a p-marker. The hope of expressing the essence of the syntactical structure of a sentence by one p-marker therefore implies strong assumptions about the language.

Overlapping constituent grammar

If the rules of R do not obey the restrictions mentioned for disjoint-constituent structure grammars, that is, if rules occur of the type

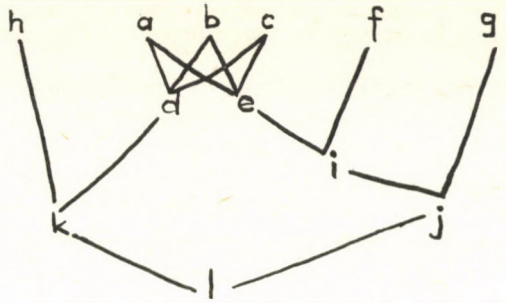
$$abc \rightarrow de$$

or

$$abc \rightarrow dc$$

no equivalence classification of substrings is obvious and no tree can be drawn without further assumptions.

The most natural would be to draw a graph of the following kind:



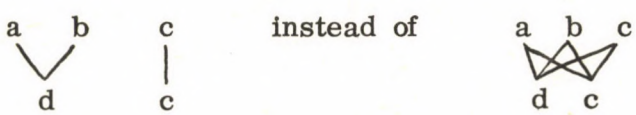
Unlike p-markers, this graph attributes one and the same substring of the input string to more than one higher constituent also when these higher constituents are disjoint. Here abc belongs to d and to e, to k and to i.

It is by no means an unnatural description of a sentence to let one segment have more than one function, nor is it impractical to represent such structures as graphs. On the contrary, that is what graphs are for, and in the special case where no two branches ever coalesce, the graph seems to be so utterly simple that it is, at any rate, rather a waste of paper to print drawings of it.

For a subset of the grammars now under discussion we can, with some good will, construct p-markers, although the same rules contain more than a single right handed element. If the rules are of the type

$$abc \rightarrow dc$$

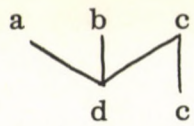
or, generally, only one symbol on the right is different from the corresponding symbol to the left, we may, by convention*, consider ab to be a constituent of type d, whereas c only functions as a context. For these context-sensitive cases we therefore can agree to represent our reduction as follows:



It might seem as natural to draw

---

* Chomsky (1963) p. 294, Handbook of Mathematical Psychology, edited by Luce, Bush, and Galanter.



saying that d is a representation of c as well as of ab, since d could not have been rendered as ab unless c had been present. One would then have overlapping constituents in cases such as Swedish gott, reducible to godt:

Nobody seems to be over-happy with this attempt to "add conditions to guarantee that a p-marker for a terminal string can be recovered uniquely from its derivation" and for this and more serious reasons linguists turn away from these types of constituent grammars altogether. But it is characteristic that one attempts to find "unique" equivalence classifications, i.e., tree graphs of the simple kind described. "We assume that such a tree graph must be a part of the structural description of any sentence; we refer to it as a phrase-marker p-marker. A grammar must for adequacy provide a p-marker for each sentence".* In other words, rather than modify the kind of graph employed, one replaces it, in transformational grammar, by an ordered set of such simple graphs.

The multi-index notation permits an alternative mode of presentation, as will appear in the next few paragraphs.

---

*Chomsky, op. cit. p. 288.

Infinite Rewriting Systems

We now consider the case where a grammar $G = \langle S, R, I, T \rangle$ contains an infinite alphabet S.

In particular, we consider the set S of vectors over a finite set S' of indexes:

$$S = S' \cup \{ s_i' s_2' \ldots' s_n \mid s_i \in S' \}$$

For S we introduce the general multi-index multiplication schema:

(1) $$(s_1' s_2' \ldots' s_n)\ (t_1' t_2' \ldots' t_m) \rightarrow$$

$$(s_1 t_1)'\ (s_2 t_2)' \ldots' (s_n t_n)' t_{n+1}' \ldots' t_m \qquad \text{if } n < m$$

$$(s_1 t_1)'\ (s_2 t_2)' \ldots' (s_n t_n) \qquad \text{if } n = m$$

$$(s_1 t_1)'\ (s_2 t_2)' \ldots' (s_m t_m)' s_{m+1}' \ldots' s_n \qquad \text{if } n > m$$

that is, for $i > n$ and $j > m$ we consider $s_i = t_j = e$, where e is a unit element such that $ae = ea = e$ for all a.

R' contains, except the general multi-index schema (1), a finite set R' of rules or rule schemata over S

(2) $$R' = \{ \alpha \rightarrow \beta \mid \alpha = a_1 a_2 \ldots a_n,\ \beta = b_1 b_2 \ldots b_m,\quad n \leq m \}$$

where a_i and b_j are elements in S or variables over S or over specified subsets thereof.

T is given either explicitly or as an infinite subset of S

$$T = \{ t' x \mid t \in A \subset S,\ x \in S \}$$

i.e., as those elements in S which consist of an element in a finite set A, arbitrarily subscripted.

We note that every element s in S defines an infinite class of elements beginning with the vector s, just as a decimal number defines a class of number with the same or a greater number of digits.

The rules of R are such as

1 $ab \rightarrow c$

2 $a'x\ b'y \rightarrow c'z$

3 $a'x \rightarrow b$

4 $a \rightarrow b'x$

and so on. To make a language decidable it is obviously sufficient - by way of analogy with the reasoning above - to require that the right-hand side should never contain more letters out of the alphabet S' than the left-hand side, thus excluding rules like rule 4 above. The fact that the letters are here distributed over different levels, so constituting one or more symbols of S, cannot invalidate that argument.

The conclusion obviously also remains intact if we accept rules with a longer right-hand side for rewriting symbols which never occur on the right-hand side of any rule, that is, if we make allowance for assignment rules.

In the following we shall restrict ourselves to <u>context-free</u> multi-index rules, that is, the rules shall

a) contain one element of S on the right-hand side and wherever practical* the rules shall also

b) contain at most as many elements of S' on the right-hand side as on the left-hand side, except where the left-hand side consists exclusively of elements which occur on the right-hand side of no rule.

---

*The second restriction is unnecessarily severe. One may well include, e.g., rules which are not reductive with reference to S'' but which are <u>strictly</u> reductive on the highest level they refer to and which do not increase the number of levels referred to by any rule.

Though each rule is a context-free rule, such a multiindex grammar is not a disjoint-constituent grammar; constituents do overlap:

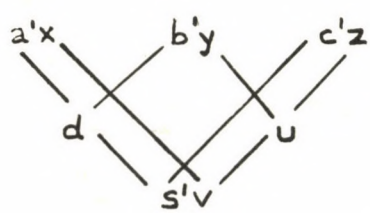
Let us consider a grammar where

$$ad \rightarrow d$$
$$dc \rightarrow s$$
$$xy \rightarrow u$$
$$uz \rightarrow v$$

and where $s'v \in T$. Let us consider the analysis of the string $a'x \;\; b'y \;\; c'z$:

$$a'x\ b'y\ c'z$$
$$d'xy\ \ c'z$$
$$s'xyz$$
$$s'xu$$
$$s'v$$

or graphically:



We see that segmentation is overlapping but that each level of indexes represents one equivalence classification and one tree-shape graph.

In many cases, context-free multi-index rules are weakly equivalent to context-sensitive rules, as will appear from the following few examples of languages which notoriously cannot be described with ordinary context-free rules. Crudely, we may say that taking an index on another level into account is an implicit way of regarding context.

Example 1. The language "$a^n b^n c^n$".

R: $a \rightarrow x'p$

$b \rightarrow y'p$

$c \rightarrow z'q$

$xy \rightarrow s$

$xsy \rightarrow s$

$sz \rightarrow s$

$ppq \rightarrow e$ where e is the unity element.

Illustration:

aabbcc

x'p x'p y'p y'p z'q z'q

x'p s'pp y'p z'q z'q

s'pppp z'q z'q

s'pp z'q

s'e = s

T = s

Example 2. The "reduplication" language, consisting of an arbitrary string of a's and b's followed by the same string repeated.

R : $xy \rightarrow x'y$ for x = a,b and y = a,b

$xx \rightarrow s$ for x = a,b

$s's \rightarrow s$

Illustration:

abbababbab

a'(b'(b'(a'b))) a'(b'(b'(a'b)))

s'(s'(s'(s's)))

s

Example 3. The language $(a^n b^n)^m$

R': $x\ x'y \rightarrow x'(x'y)$ for $x = a,b$ and for all $y \in S$

$ab \rightarrow t$

$t'x\ t'x \rightarrow t'x$ for all $x \in S$

$t \rightarrow s$

$s's \rightarrow s$

$T = \{s\}$

Illustration:

aaabbbaaabbb

$a'(a'b)\ b'(b'b)\ a'(a'a)\ b'(b'b)$

$t'(t't')\ t'(t't)$

$t'(t't)$

$s'(s's)$

s

Example 4. The language $a^m b^n c^{mn}$

R': $x\ \ x'y \rightarrow x'(b'y)$ for $x = b,c$ and all $y \in S$

$a\ b'x\ c'x \rightarrow b'x$ for all $x \in S$

$b \rightarrow s$

$s's \rightarrow s$

$T = \{s\}$

Illustration:

aaabbbbccccccccccc

$aaa\ b'(b'(b'b))\ c'(b'(b'b))\ c'(b'(b'b))\ c'(b'(b'b))$

$aa\ b'(b'(b'b))\ c'(b'(b'b))\ c'(b'(b'b))$

$b'(b'(b'b))$

$s'(s'(s's))$

s

Thus, the possibility to add further index levels at option provides a means of performing arithmetical operations. The context-free multi-index rules are powerful and cover many languages of what is known as the context-sensitive type.

We shall now turn to linguistic interpretations of such a calculus.

Multi-index Calculus in Linguistics

The multi-index calculus can be applied in linguistics above all for two purposes: to replace context-sensitive rules and to provide a means of representing p-markers.

Context-free multi-index rules derived from context-sensitive rules

It is possible to replace many - all? - context-sensitive rules by an equivalent set of context-free multi-index rules.

Thus, the rule

$$a \rightarrow b / \_\_ c$$

can be replaced by

$$a \rightarrow b'p,\ c \rightarrow c'q \text{ and } pq \rightarrow e \quad \text{or, more cautiously}$$

by the assignment rules

$$a \rightarrow A'p$$
$$c \rightarrow C'q$$

and the reduction rules

$A'p \rightarrow A'r$

$A'r \rightarrow B'r$

$rq \rightarrow e$

$p \rightarrow e$

$q \rightarrow e$ where e is the unity element.

Let us consider the following little grammar:

$j \rightarrow i/g\ -$

$h\ g \rightarrow g\ h$

$i \rightarrow d/h\ -$

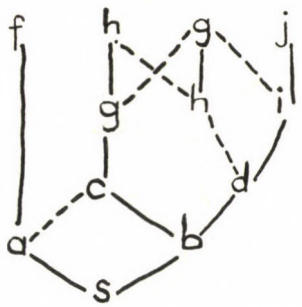
$g\ h \rightarrow c$

$f \rightarrow a/\ -\ c$

$c\ d \rightarrow b$

$a\ b \rightarrow s$

With this grammar, the sentence <u>fhgj</u> will be analyzed thus



We have here adopted a "mixed" tree representation for context-sensitive structures, with obvious significance.

We can reduce the same sentence to <u>s</u> by the following set of rules:

$j \rightarrow i'k$

$g \rightarrow g'1$

$1k \rightarrow e$

$h \rightarrow g'm$

$g \rightarrow g't$

$g \rightarrow h'n$

$mt \rightarrow m$

$mn \rightarrow e$

$i \rightarrow d'p$

$h \rightarrow h'q$

$qp \rightarrow e$

$gh \rightarrow c$

$f \rightarrow a'r$

$c \rightarrow c't$

$rt \rightarrow e$

$cd \rightarrow d$

$ab \rightarrow s$

Thus,

f h g j

f h g'l i'k

f h gi' (lk)

f g'm h'n i

f gh' (mn) i

f g h'q d'p

f g hd' (qp)

f c d

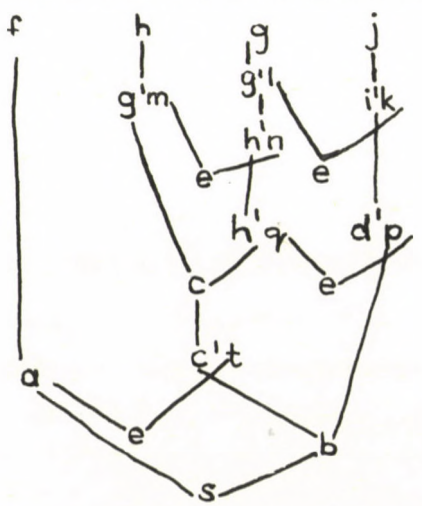
a'r c't d

ac' (rt) d

a b

s

Graphically, this means that we have a cet of interconnected treegraphs:



In a transformational grammar, we interpret G'' as a grammar component, adding to our grammar a component $G' = \langle S', R', I', T' \rangle$ where I' is the set T'' of p-markers, T' is a subset thereof and R' is a set of multi-index rewriting rules such as

$$a'x \rightarrow a'y$$

$$a'x\ b'y \rightarrow c'x$$

$$a'x\ a'x\ b'y \rightarrow a'x\ b'y\ a'x\ b'y$$

$$a'x\ .\ b'y \rightarrow b'y\ .\ a'x$$

for specified sets of values for x, y, etc., that is, substitution, reduction, expansion and permutation rules for which the conditions are not confined to one index level at a time.

Regarding the analysis as a syntactic tree, we may characterize transformational rules as such where the conditions for some symbol(s) to be rewritten in a specified way refer to the "vertical" neighbours (not to the "horizontal" neighbours as in context-sensitive rules). We might speak about pretext and posttext sensitive rules, or generally about "kintext sensitive" rules. Obviously and notoriously, "kintext" must play a different role in generative and in recognition procedures, since pretext in one case is posttext in another.

Thus, one component may map the input strings on $T'' = \left\{ t_i'x \mid t_i \in T;\ x \in S'' \right\}$ and a transformation component may map $I' = T''$ on $T' = \left\{ t'y \mid t \in A \right\}$ and $y = \left\{ a_1' a_2' a_3' \ldots \mid a_i \in B \right\}$ where B is a subset of S'' and $A \subseteq T$. Or we may define the target set for each component in other ways.

Multi-index calculus in a transformational grammar

Given a constituent structure grammar $G = \langle S, R, I, T \rangle$ we obtain an infinite grammar G'' by replacing S by

$S'' = SU\left\{s_1's_2's_3'\ldots \mid s_i \in S''\right\}$ and R by

$$R'' = \left\{a_1a_2\ldots a_n \rightarrow b'(a_1a_2\ldots a_n) \mid (a_1a_2\ldots a_n \rightarrow b) \in R\right\}$$

if R is context-free and otherwise

$$R'' = \left\{a_1a_2\ldots a_n \rightarrow b_1'(a_1a_2\ldots a_n) \cdot b_2\,(a_1a_2\ldots a_n) \cdot \ldots \cdot b_m'\right.$$

$$\left.(a_1a_2\ldots a_n) \mid (a_1a_2\ldots a_n \rightarrow b_1b_2\ldots b_m) \in R\right\}$$

and replacing $T = \left\{t_1, t_2, \ldots, t_k\right\}$ by

$$T'' = \left\{t_i' \; x \mid t_i \in T \; x \in S''\right\} .$$

That is, we obtain a grammar* which maps given strings on an infinite set which may be considered as a set of p-markers** G'' is then an interpretation grammar, corresponding to G.

Thus, one-level reduction rules suffice for a decision grammar for a constituent-structure language and multi-index reduction rules suffice for an interpretation grammar for such languages. Multi-index rules also suffice for a decision grammar for a transformationally defined language.*** The question remains if they suffice for an interpretation grammar for the latter.

---

* a decidable one, see hints above, footnote. The number of levels does increase, but all rules refer exclusively to the uppermost level.

** These multi-index expressions naturally contain all information that transformations operate upon. Indeed, they will often contain too much, but superfluous indexes can easily be eliminated by multi-index rules; the point is that no side conditions for permissible transformational rewritings need be observed. Everything needed for the calculus is in the string.

*** if this is decidable. They may also, incidentally, provide simple decidability criteria for a transformational grammar. Cf. the hints above.

A structural description of the sentence may be given as the sequence of p-markers obtained during the analysis. Now, since the relative order of operations is not inherently fixed, we would like to find a representation of such sequences such that equivalence can easily be defined. That is, we want to find an adequate interpretative grammar corresponding to G'. Can multi-index rules serve those purposes?

The unified formalization, provided by the multi-index representation, might prove an aid to finding an effective interpretative calculus for transformationally defined languages.

Conclusion

The multi-index calculus seems promising for several linguistic purposes, especially where restrictions can be assigned to several, weakly interacting levels.

THE RECONSTRUCTION OF THE NOTION OF "AGGLUTINATION" IN THE FRAMEWORK OF GENERATIVE GRAMMAR

(Preliminary version)

Ferenc Kiefer

The aim of this paper is to examine certain aspects of agglutination in the light of generative grammar. The pertinent data will be drawn from Hungarian. For simplicity's sake I shall restrict myself to Hungarian noun inflection. On the basis of some observations I shall stipulate a new definition of the notion of "agglutination". It will be left open, however, whether the thus reconstructed notion of "agglutination" will be applicable to agglutinating languages other than Hungarian as well. It should also be made clear that I do not want to account for all aspects of noun inflection in Hungarian.

With the above restrictions and qualifications in mind we may divide the paradigms of the Hungarian noun into two major groups. One will contain forms as

(1)	hajó	ship
	hajót	ship (acc.)
	hajónak	to the ship (dat.)
	hajóban	in the ship (loc.)

(1) are usually considered to be the case forms of the Hungarian noun. Without going into details one may safely say that these forms as well as others

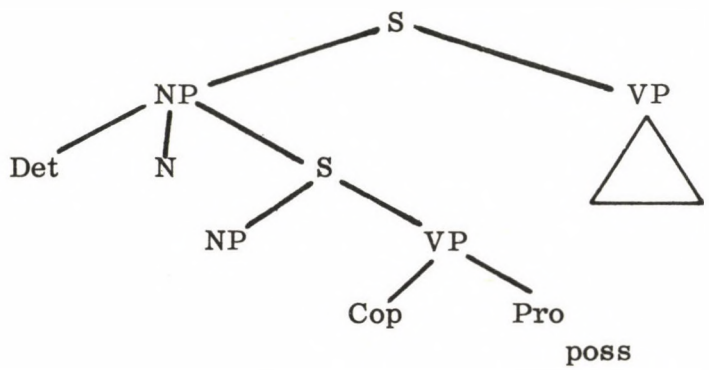
of similar (case) function can be accounted for by means of morphological rules in one way or another. It is clear that agglutination does not enter into the picture here.[1/]

The second group of noun forms in Hungarian can be exemplified by the following forms:

(2)	(i)	hajóm	my ship
		hajói	his/her ships
	(ii)	hajóé	belonging to the ship
		hajóké	belonging to the ships

These forms, it will be claimed, cannot be accounted for by what is usually understood by morphological rules. Notice first that there seems to be an essential difference between (2) (i) and (ii). The endings in (2) (i) are generally referred to as "possessive personal endings" (birkokos személyrag). They can apparently be derived from underlying structures of the following type (details ommitted):

(3)



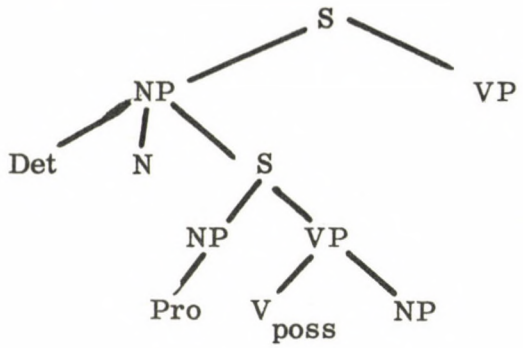
In other words, hajóm, for example, seems to go back to

(4) a hajó, amely az enyém
the ship that (is) mine

or, more precisely, a structure like (3) can be realized on the surface by either (4) or by the "agglutinated" form hajóm.

If we were to consider the possessive pronouns in Hungarian as derived categories we would have the structure (5) instead of (3):

(5)



which is the structure of sentences like

a hajó, amelyet én birtokolok
the ship which I possess

(This sentence in undoubtedly a grammatical sentence though somewhat odd.) I will, however, leave the question open for the moment which one of the two alternative treatments should be given preference. I will take up this question later on. The morpheme é which appears in the forms in (2) (ii) is generally referred to as "possession sign" or "possession ending" (birtokjel). In order to understand the function of this morpheme let us consider the following sentence:

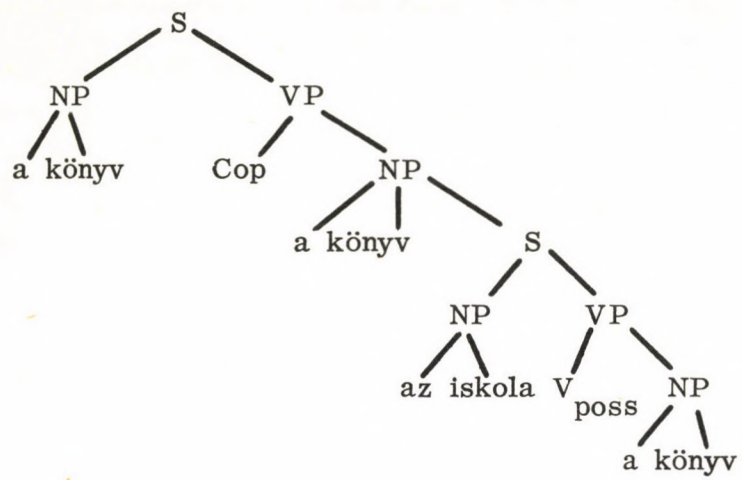
(6) A könyv az iskoláé.
the book the school-belongs-to
The book belongs to the school.

(6) can be paraphrased as (7):

(7) A könyv az iskolának a könyve.
the book the school-of the book-its
The book is the book of the school.

Most genitive constructions can evidently be derived from an underlying sentence containing some verb of possession. Thus, the underlying structure of (7) can be rendered by (8) (details omitted):

(8)



Notice, incidentally, that the position of the two NP's in the embedded S is immaterial because it depends on the verb of possession which we happen to choose. Transformational rules can take quite easily account of this difference. Important for our considerations is the fact that in the case of (3) as well as in that of (8) an NP together with an embedded S can be condensed to a noun plus a morpheme. The latter must then be attached to the noun in some way.

But this is not yet the whole story. The possessive forms and the "case forms" can combine. Thus, we get, among others, forms like

	iskoláméit	those belonging to my schools (acc.)
(9)	iskoláméinak	to those belonging to my schools (dat.)
	iskoláméból	from that belonging to my school (elat.)

(9) have an underlying structure which is, roughly speaking, a combination of (3) and (8).

Forms like (2) (ii) and (9) are generally used elliptically. E.g.

(10) (i)

Kinek az érdekeit képviseled?
whose the interests represent-you
Whose interests do you represent?

Az iskoláméit.
the school-my belonging-to
The interests of my school.

(ii)

Kinek a pénzéből élsz?
whose the money-from live-you
From whose money do you make your livelihood?

Apáméból.
the father-my-belonging-to-from
From that of my father.

Thus, while in cases like (2) (i), if the alternative (3) is adopted, the transformational rule generally referred to as Equi-Noun-Phrase-Deletion will operate before the rules of agglutination are put to work, in cases like (2) (ii) or (9) both instances of the noun phrase underlying for the possession ending must be deleted in order to get the agglutinated forms. Of course, the generation of them would require some additional rules. The condition for the aforementioned deletion must be sought in the linguistic context of the sentence under consideration. The deletion can be carried out only under the presupposition that the noun to be deleted is known from the previous context. That is all that can be said about this topic for the moment. It is clear, however, that agglutination hinges on deletion of this sort in an essential way.

Let us now turn first to the generation of the possession ending. For the possession ending <u>é</u> the structure to be reduced is the following one:

(11) $$\left[N_1 \left[N_1 \quad V_{poss} \quad N_2 \right] S \right] NP$$

and the outcome of the reduction should be

(12) $$N_2 \quad é$$

In (11) the second N_1 can be deleted by means of the Equi-Noun-Phrase-Deletion rule. This would give us the surface form exemplified in (7). In order to get (6) we need a rule like (13):

(13) $$\left[\left[\begin{matrix} N_2 \\ +Gen \end{matrix}\right] \quad N_1\right]_{NP} \longrightarrow N_2\ \underline{é}$$

This amounts to saying that é stands for a noun which is part of a gentitive construction. We still have one N_1 left in (8), this can be deleted under the presupposition that it is known from the previous context.

Now we may take up the problem of possessive pronouns again. We have seen that for the é - forms we have two surface realizations. Both are derived from an underlying structure similar to that depicted in (8). Now it might be expedient to look for two surface realizations for (2) (i) as well. One might consider

(14)
a könyv az enyém
the book the mine
the book is mine

and

(15)
a könyvem
the book-my
my book

as derivable from the common source

(16)
x a könyv, amelyet az én birtokol
the book which the I possess

through the intermediate stage of (17):

(17)
x a könyv az énnek a könyve
the book the I_{gen} book
x the book is I's book

'x' indicates that the forms are not well-formed surface forms (they can of course figure as deep structure sources). (16) has the underlying structure (5).

This treatment would thus mean that we have a common source for the é-forms and for the possessive pronouns and endings. Notice that the slight motivation which we gave above may be a little bit strengthened by pointing out that the non-elliptical form of (15) is rather (18):

(18)

az én könyvem
the I book-my

This is explained in a natural way if we have I in the underlying structure. Now we may set up two rules that account for the forms (14) and (18):

(19) $$\begin{bmatrix} \text{Pro} \\ \\ +\text{Gen} \end{bmatrix} \quad N \xrightarrow[NP]{} \text{Pro} \quad \acute{e}$$

and

(20) $$\begin{bmatrix} \text{Pro} \\ +\text{Gen} \end{bmatrix} \quad N \xrightarrow[NP]{} \text{Pro} \quad \text{Suff}$$

For the case characterized by (19) we would have a further rule that introduces possessive pronouns

(21) $$\text{Pro } \acute{e} \longrightarrow \left\{ \begin{matrix} \text{enyém} \\ \text{tied} \\ \text{övé} \\ \text{mienk} \\ \text{tietek} \\ \text{övék} \end{matrix} \right\}$$

The parallelism between (19) and (13) would speak for the alternative (5) rather than for that given by (3). Furthermore, (21) would imply that there is no such category as possessive pronoun in Hungarian syntax, or, to put it differently, that possessive pronouns are agglutinated forms of deep-seated syntactic relations.

The analysis of possessive pronouns put forward above might get further evidence from the fact that the sentence

whose book is this?

is rendered in Hungarian by

kié ez a könyv?
who/é this the book

and the answer can be either one of the pronouns in (21) or anything else generated by the rule (13). Consequently, the rules (13) and (19) should be considered to be independent of each other. I shall, however, not pursue this issue any further.[2/]

We can now conclude with some safety that the most important aspect of agglutination, in the case of nouns, is the reduction of a noun to a "bound" morpheme. This is brought out clearly by rules (13), (19) and (20). Let me refer to these rules as lexeme reduction rules.

The notion of "lexeme" should be understood somewhat informally here (corresponding to that of lexical formative). The "bound" morphemes which are the output of the lexeme reduction rules resemble to inflectional endings. I would not consider either of them as being part of the lexicon of a given language.

The "bound" morphemes must undergo further rules which, roughly speaking, will attach them in case of (13) to the noun N_2, in case of (20) to the first N in (17). Finally, in case of (19) the rule (21) would obligatorily apply.

Examples in (10) show clearly that there must be a strict ordering of the "attachment" rules. First, the output of (20) must be taken care of, then the output of (13) and finally the rules would apply that spell out inflectional endings. All these rules apply only once and obligatorily. They are, therefore, clearly morphological in nature.[3/] What kind of rules are, however, the lexeme reduction rules?

They seem to be optional. This is clearly the case with respect to rule (13). The counterpart to (13) would be a rule that developes an unagglutinated, "explicit" structure (genitive construction) out of the underlying structure. Rules

(19) and (20) constitute two alternatives. They differ from (13) in that that one of them must obligatorily apply. This would suggest that they are syntactic rules rather than morphological ones.

If they are syntactic rules they are certainly not cyclical. In this respect there is one important point to make. Notice that structures like (5) and (8) are recursive. Consider examples (22) (i) and (ii):

(22)

(i) az iskola könyvtárának a könyve
the school library-of the book-its
the book of the school's library

(ii) a kerület iskolája könyvtárának a könyve
the district school-its library-of book-its
the book of the library of the district's school
etc.

(22) (i) and (ii) do not reduce, however, to (23) (i) and (ii), respectively.

(23)

(i) +az iskoláéé

(ii) +a kerületéé

This means that lexeme reduction rules do not apply cyclically.

The claim that lexeme reduction rules are syntactic rules is rather clear in case of (13). It cannot be chosen after the other option, the genitive construction, has been fully developed. As to (19) and (20) it seems to me a particular feature of modern Hungarian that it has no real verb of possession ("lexical gap"). The verb "birtokol" is rarely used. From this comes the oddity of a sentence like

a könyv, amelyet birtokolok
the book which possess-I

which might be a sentence derivable from (16). I am not quite sure, however, whether there is really no other possibility of deriving a nonagglutinating surface form out of (16). Anyhow, since rules (19) and (20) are quite parallel to (13), we may consider them with good reason to be syntactic rules. It should be made clear, however, that further research is needed in order to clarify this issue.

After the above rather inconclusive discussion of the nature of lexeme reduction rules we may take up again our main theme, the notion of agglutination.

Notice first that the attachment of bound morphemes to a stem can also happen in languages generally not considered to be agglutinating. So, for example, the Swedish noun form

(24)
flickornas
girls-the-'s
the girls'

comes from an underlying structure something like

N Aff Aff Aff

as a result of the operation of so-called affixation rules. The affixes are, then, spelled out and attached to N in some fixed order by inflection-rules. The important thing about agglutination, therefore, seems not to be the attachment of one or more bound morphemes to a stem morpheme (lexeme) but rather the fact that these morphemes come from quite different sources in the case of languages like Hungarian than in the case like those as Swedish (though, of course, some of them may have similar sources, e.g. the inflectional endings proper).[4/] As a result we may claim that the most important aspect of agglutination is given by the rules (13), (19 and (20). This entails that agglutination is more a syntactic phenomenon than a morphological one contrary to the view oftan expressed in traditional linguistics. Furthermore, agglutination is a superficial aspect of language brought about by rather late rules in syntax.

Now we may stipulate, though tentatively, the following definition of "agglutinating language".

A language L is called to be agglutinating with respect to a category K in L iff the grammar G of L contains lexeme reduction rules such that the lexemes involved are dominated by K.

The qualification "with respect to K" seems to be necessary in view of the fact that most agglutinating languages are not entirely agglutinating, i.e. they may not be so with respect to some category K' different from K.

A language L is called to be agglutinating iff the grammar G of L does not contain any lexical category K^+ such that lexemes dominated by K^+ do not enter regularly agglutination.

Hungarian, for example, seems not to be agglutinating with respect to adverbs but it is not quite clear whether the category "Adverb" should be considered a lexical category at all. (It might be, and, in fact, in the majority of cases, it <u>is</u> derived from other categories.) If this is so, then Hungarian could be considered to be one example of an agglutinating language.

My last remark concerns the contextual restrictions of the lexeme reduction rules. In (13), (19) and (20) nothing suggests that they can operate under certain conditions only. But the fact that a noun is reduced in a non-recoverable way indicates that there must be another "copy" of this noun present at some point of the discourse. As already pointed out, for the time being, there seems to be no way to state formally any such contextual condition.

To summarize, we have seen that agglutination raises a great number of interesting questions which are worth to be explored in more detail. From these I would put emphasis on two problems. One is the contextual determination of agglutination just referred to. The other one is the fact that agglutination seems to be much more syntactically conditioned than anticipated thus far.

<u>Footnotes</u>

1. It would be wrong to claim that particles known as prepositions in languages like English are agglutinated in Hungarian. Hungarian is simply synthetic where English is analytic. In other words, some English prepositional phrases would roughly correspond to various case forms in Hungarian.

2. I do not want to give here a detailed and full-fledged account of the possessive pronouns in Hungarian, of course. The structures (3) and (5) are deliberately oversimplified and serve the purposes of my argument only. For a more detailed discussion of this topic see, for example, Lotz.

3. For a detailed discussion of morphological processes in generative grammar see Bierwisch and Kiefer.

4. The descriptive definition of agglutinating language runs as follows. In an agglutinating language (i) words can uniquely and easily be segmented into morphs ("determinacy with respect to segmentation"), and (ii) there holds a one-to-one correspondence between morph and morpheme. (See, Lyons, pp. 188-189.) This definition does not take into account the possibility of different deep structure sources and would therefore bring under the same heading (10) and (24), which is clearly an undesired result.

References

M. Bierwisch: Syntactic Features in Morphology: General Problems of So-Called Pronominal Inflection in German, in: To Honor Roman Jakobson, Mouton and Co., The Hague (1967), pp. 239-270.

F. Kiefer: Morphological Processes in Generative Grammar, An Outline of Swedish Morphology, forthcoming in Acta Linguistica Hung.

J. Lyons: Introduction to Theoretical Linguistics, Cambridge University Press, 1968.

ON THE LINEAR PATTERNING OF VERBAL WORKS OF ART

János S. Petőfi

0. Introduction

Computers have opened new ways in the analysis of works of art. This is especially significant in the investigation of verbal works of art. As in this type the text is not only a vehicle but also a dynamic creator of the message, its analysis is of basic importance. (Of course, one must add that the analysis of the text-structure is only one component of the complete analysis of the verbal work of art.)

In the text-structure we distinguish the linguistic and the sound-textural component. In the first one the syntactic-semantic, in the second one the rhythmic-euphonetic structure of the text is manifested. Both can be hierarchically and linearly patterned.

The 'hierarchical patterning' means the way the text as a whole is built up from the basic units of the structure through the levels of the composition units of different complexity.

The 'linear patterning' is a network of the repetitive returns of elements of the most different character and level that interweave the whole text.[1]

In my paper I wish to deal with the linear patterning of the linguistic component. The analysts have generally paid great attention to the investigation of this aspect. Besides that this kind of patterning is really significant (Jakobson sees the essence of the poetic language in a phenomenon closely connected with this one), its analysis is simpler on every level than that of the hierarchical patterning. In the following I shall try to examine the conditions and methods of the computer analysis of the linear patterning, in accord with the generative linguistic theory.

1. The conditions and methods of the analysis by computer

For the computer to discover all the manifestations of the repetitive returns we need an exact grammar and a suitable analysing algorithm and program.

We see it necessary to create the possibilities of an analysis where the text to be analysed has to be preprocessed to a minimum extent only.

1.1 On the grammar

The two main components of the grammar are the lexicon and the system of the syntactic-semantic and phonological rules. The analysis of verbal works of art raises special demands to both of them, that is taking into consideration that we deal with 'verbal works of art' and 'analysis'.

1.1.1 The lexicon

The lexicon has to contain all the information that can guarantee the micro-grammatical, grammatical and conpositional analysis of texts, that is the discovering of the structure of words, sentences and text-units larger than sentences.

For a complex lexicon that is suitable for this purpose I shall use the term thesaurus. A linguistic thesaurus has to consist of two main sectors: the sector of definitions of the individual lexical units and grammatical formatives, and the sector of classifications made according to various aspects.

A/ The sector of definitions contains the 'dictionary entries'.

The dictionary entry of a lexical unit (LU) has to contain the following classes of information:

i/ phonological matrix (PHM)

the matrix of the phonological feature-bundles of the phonemes that form the LU

ii/ syntactic information (SYI)

a/ the undifferentiated syntactic category (USC) of the LU

b/ the undifferentiated syntactic category-chain (USC-CH) marking the microsyntactic structure of the LU

c/ the ordered sets of syntactic markers (SMS) of the LU: from the undifferentiated ones towards the more and more differentiated ones

d/ morphological information (MI)

iii/ semantic information (SEI)

the ordered sets of semantic markers of the LU

iv/ condition information (COI)

information referring to the grammatical use of the LU

v/ other information (OTH)

the 'origin', 'stylistic value', etc. of the LU

vi/ indices

referring to the sector of classifications (syntactic-semantic and different thematic indices)

I do not wish to deal with the detailed analysis of the above listed information-classes[2], I just want to add some observations to them. Of course the nature of the Hungarian language is reflected in these observations, too.

i/ The phonological matrix represents a definite word-form with every kind of word (noun: singular, nominative; verb: active, indicative mood, present tense, third person singular; adjective: positive degree, singular nominative; etc.).

ii/a. The undifferentiated syntactic category represents the form class of the given word (N: noun; V: verb; A: adjective, etc.).

b. When giving category-chains to the mycrosyntactic structure we may represent the affixes by small letters, e.g.:

N: V + n kirándulás (kirándul+ás) excursion
N/A/Adv + n távolság (távol+ság) distance

V: prefix + V elalszik (el+alszik) falls asleep
V + v szunnyadozik (szunnyad+ozik) slumbers etc.

c. If we accept Chomsky's system of the syntactic markers of the noun, we get the lines of the following character:

-Common	+Animate	+Human	Balázs	Blaise
+Common	-Count	-Abstract	robogás	rattling
+Common	+Count	-Animate	cukor	sugar

etc.

The following categories serve for the characterization of the verbs:

+transitive	megkap	gets
-transitive	szundit	dozes
+progressive	alszik	sleeps
-progressive +beginning	elalszik	falls asleep

d. The flection of Hungarian nouns is defined by the root-type, the accusative singular, the nominative plural and the personal possessive suffix of third person singular, e.g.:

1. type:	-at,	-ak,	-ja	vad	beast
	- t,	- k,	-ja	üveggolyó	chrystal ball
	-et,	-ek,	- e	fej, szem	head, eye
2. type:	-ot,	-ok,	- a	álom	dream

(The first type shows no root-changing, but the fifth does; for example certain suffixes join not to álom, but to álm.)

With the verbs we also have to mark the changing of the root. E.g. at the verb alszik we have to make certain forms from the root alsz, others from the root alud.

iii/ In connection with semantic information we have to define a system of 'elementary meanings' with the help of which the meanings of words are expressable. Here I give the definitions of the Hungarian Explanatory Dictionary. (A Magyar Nyelv Értelmező Szótára) (These are far from being ideal definitions.)

V: szendereg (dozes) : rests in light sleep, half awake

N: üveggolyó (chrystal ball): a small ball,
used as toy,
usually made of solid glass,
with colored veins

iv/ Syntactic condition for example

V: aluszik (slumbers) (the form 'aluszík' can be used only in indicative mood, present tense)

Semantic conditions:

With nouns we have to give the conditions of the compatibility of the possessor-possession relations, and of their employment as predicates. E.g.:

szem eye (a part of body of '+Animate')

With verbs we have to give the conditions referring to the arguments, e.g:

alszik	(sleeps)	+Animate
lehunyja	(closes)	only in this constant word-group: +Animate szemét (his, her, its eyes)

v/ As I have mentioned above 'style valuing' belongs to the other informations. E.g:

aluszik is a variety of 'alszik'.

vi/ (See later.)

Hence the dictionary entry of a LU has the following structure:

üveggolyó (chrystal ball) elalszik (falls asleep)

PHM /ü//v//e//g/+/g//o//ly//ó/ PHM /e//l//a//l//sz//i//k/

SYI	USC	N	SYI	USC	V
	USC-CH	N + N		USC-CH	prefix + V
	SYMS	+Common +Count -Animate		SYMS	+Active -Transitive -Progressive +Beginning
	MOI	1. root-type -t, -k, -ja		MOI	elalsz elalud
SEI		used as a toy a small ball usually made of solid glass with colored veins	SEI		gets into the state of sleeping
COI		-	COI		+Animate
OTH		-	OTH		-
indices		.....	indices		.....

The structure of the dictionary entries of the grammatical morphemes /GM/ is simpler, though it shows analogous features to the structure of the dictionary entries of lexical units.

i/ phonological matrix;

ii/ the undifferentiated syntactic category of the word, that the given GM can join;

iii/ the grammatic meaning of the GM;

iv/ information referring to the immediate context of the GM;

v/ information referring to the rules that realize the morphonological changings that are connected with the employment of the GM;

vi/ indices referring to the sector of classifications.

The system of Hungarian suffixes is rather complicated. I just put some notes to this sketch of dictionary entry.

i/ A part of the suffixes occurs in two or more forms containing different vowels. E.g:

-ban, -ben,	in
-hoz, -hez, -höz	to

The phonological matrix represents the so called archiphoneme:

-bAn,

-hOz.

ii/ As the homonymous suffixes are already differentiated to a certain extent by the syntactic category of the basic word that the given formative can join this has to be given anyway.

iii/ The grammatical meaning is made specific by the semantic markers of the basic word. E.g:

-bAn	adverbial suffix
/... +Place.../ -bAn	adverbial suffix of place
/... +Time.../ -bAn	adverbial suffix of time

iv/ The order of formatives is bound in Hungarian. The possible immediate context of every formative has to be given.

v/ Certain suffixes expect the lengthening of the last vowel of the previous morpheme, others expect a linking-sound to be inserted, etc. E.g:

a/	szem	-e	-t	szemét
	eye	possessive personal suffix third person	acc.	its eye
b/	szem	-O	-t	szemet
			acc.	eye

vi/ (See later.)

Hence the dictionary entry of grammatical morphemes is the following:

-e		-j	
PHM	/e/	PHM	/j/
SYI USC	N	SYI USC	V
SEI	possessive personal suffix third person singular	SEI	the sign of imperative mood
COI	root or formative suffix suffix	COI	root or formative suffix suffix
OTH	.....	OTH	.....

Of course all this is only the representation of a possible form and not a sketchy account of some definitive solution.

B. The sector of classifications firstly has to contain a basic list that is a guide of finding words in the sector of definitions. This contains all the possible forms, always referring to the dictionary entry.

For example

álom		ban	bAn
álm	álom	ben	bAn
...		...	
alsz	alszik	hoz	hOz
alszik		höz	hOz
alud	alszik	...	

The underlined forms (as those standing on the right of the arrows) represent the dictionary entries. If a form represents homonymous entries, we have to give the number of the homonyms.

The syntactic classifications list the words in alphabetic order under different category-sets of one or more syntactic markers as 'headwords'. (Hence e.g.: nouns, verbs, adjectives, ...; active - non active, transitive - not transitive verbs, ...; etc.)

When generating we choose the terminal elements of the phrase-markers from the sets labelled by them.

The groups of words derived from similar roots, the synonyms, the mirror-words (e.g: give - get), the antonyms (e.g: falls asleep - wakes) belong to the semantic classifications.

Such words as head, eye, foot on the one hand, and meadow, wood, excursion, or beetle, wasp, buzzing on the other get beside each other in a thematic classification. Among these are different connections:

head, eye, foot	are part of a living creature's body,
meadow, wood	are parts of nature,
beetle, wasp	are living creatures of nature,
	(besides beetle is a genus-term, while wasp is a species-term),

buzzing — is the sound given by the wasp,

excursion — links up associatively with nature in general, that is to the meadow and to the wood, too.

(To define the types of these connections is a very significant task from the point of view of text-analysis.)

With the classifications sketched above we can create analogous groups from the grammatical morphemes, too. Besides we can create miscellanous thematic groups, as for example the group of nouns, pronouns, adjectives, postpositions, suffixes, etc. that serve for the expression of time or place.

These classifications show the competence of the native speakers referring to the lexicon on one hand, and contribute the development of the analysis of performance in a great extent on the other.

In the individual dictionary entries the indices refer to the lists in which the given entry occurs.

1.1.2 The system of rules

The other component of the grammar is a system that contains rules of different character. This problem will be discussed only very shortly here. The classes of rules are as follows:

i/ phonological, morphonological rules,

ii/ syntactic rules working within the frame of sentences,

iii/ semantic rules working within the frame of sentences,

iv/ co-textual rules

a. to text-units built up immediately from sentences,

b. to text-units built up from other text-units.

So far the regularities of the first two classes have been investigated mainly. (Here I mean investigations of a generative character.)

In the field of semantic rules -- as in the semantic characterization of dictionary entries -- there are a lot of unsolved problems. The analysis until

now was focussed on the investigation of simple sentences. (We know for example very little about the problems of semantic interpretation of sentences containing different kinds of adverbials.)

The investigation of the types of co-textual rules has just recently started. This requires the processing of a big corpus.[3]

The compilation of rules in generating and analysing has to be in permanent interaction.

1.2. On the algorithm

When sketching the algorithm I supposed we had a thesaurus that satisfies the conditions mentioned above, and a suitable rule-system. (Naturally, both the thesaurus and the rule-system reflect a given state of language. The occasional differences of poetic language are related to this.)

My conception considers not the sentences but what are called communication units as compositional basic units. For the time being we may allow the giving of their boundaries as a former preparation of the text. (I think that later, when me have gained enough experience, these boundaries will be establishable automatically.)

1.2.1 The first step of analysis is of a general character, it is the first phase of all kinds of language analysis: the morphological analysis.

Because of the complexity of Hungarian morphology the use of different special morphological "tool-lexicons" can be required.

In the morphological analysis we 'cut off' the elementary suffixes from the words step by step, and we substitute syntactic categories found in the vocabulary entries for these and the word-roots. E.g:

szemét	szem	+e	+t
(his eyes acc.)	N	possessive	acc.
	+Common	personal	
	+Count	suffix	
	-Animate	third person	

When the morphological analysis of the given text is finished we get the sequence of categories instead of the sequence of concrete word-forms. E.g: Instead of the sentence

"Lehunyja kék szemét az ég." (The sky closes its blue eyes.)

we get the following line of categories:

/V/ + /third person/ + /A/ + /N/ + /poss/ + /T/ + /N/
singular

The categories in brackets stand for the proper differentiated categories.

1.2.2. In the second phase, in the syntactic analysis on the level of communication units we discover the deep structure of the given unit from these category-lines. This happens in several steps.

In the analysis the surface structure is immediately given. First we have to discover the relations between the elements of this linear chain, then to establish the deep structure by performing the inverses of nominalizations, embedded and wordordering transformations.

With the help of the information accumulated during the analysis, we order two syntactic characteristics (characteristic sets) to every immediate constituent and communication unit. The first one shows what they are, referring to the character of the -- not word-ordering -- transformations that form the surface-structure, the second one shows in what succession the constituents take place in the surface structure.

1.2.3 The third phase is the semantic analysis on the level of communication units. As the syntactic analysis operates with syntactic categories, here, as a first step we take the semantic categories and the conditions of context.

This semantic analysis means first of all the examination of compatibility of constituents. As a result of the analysis we order a semantic characteristic to the examined communication unit.

1.2.4 This does not mean yet that we really analysed and characterized every sentence and every sentence-like unit in the three phases. Not all questions can be answered on their own levels. The answering of a morphological question may require syntactic information, syntactic questions may call for semantic and/or co-textual ones, etc.

Hence the next phase is that of the minimal co-textual analysis of communication units that cannot be analysed between their own boundaries.

With this the individual analysis of communication units comes to an end. Now follow the levels of the establishment of syntactic- semantic connections among communication units and among composition units that arise from other composition units -- but here I do not wish to deal with these questions.[4]

2. The analysis of linear patterning

As an illustration of the theoretical questions examined above let us see a short analysis.

The analysed poem is József Attila's Lullaby.

Altató

Lehunyja kék szemét az ég,
Lehunyja sok szemét a ház,
dunna alatt alszik a rét --
aludj el szépen, kis Balázs.

Lábára lehajtja fejét,
alszik a bogár, a darázs,
vele alszik a zümmögés --
aludj el szépen, kis Balázs.

A villamos is aluszik
s mig szendereg a robogás
álmában csönget egy picit --
aludj el szépen, kis Balázs.

Alszik a széken a kabát,
szunnyadozik a szakadás,
máma már nem hasad tovább --
aludj el szépen, kis Balázs.

Szundit a lapda, meg a sip,
az erdő, a kirándulás,
a jó cukor is aluszik --
aludj el szépen, kis Balázs.

A távolságot, mint üveg-
golyót, megkapod, óriás
leszel, csak hunyd le kis szemed
aludj el szépen, kis Balázs.

Lullaby

The sky closes its blue eyes,
The house closes its many eyes,
under a wadded quilt the meadow is sleeping
fall asleep nicely, little Blaise.

The beetle bows its head
on its feet, the wasp is sleeping,
the buzzing is sleeping with it --
fall asleep nicely, little Blaise.

The streetcar is slumbering, too,
and while the rattling is dozing,
it tings a bit in its dream --
fall asleep nicely, little Blaise.

The jacket is sleeping on the chair,
the tear is snoosing
today it does not stretch longer --
fall asleep nicely, little Blaise.

The ball and the whistle are napping,
the wood, the excursion do so,
the good sugar is slumbering too --
fall asleep nicely, little Blaise.

You will get the distance
as a chrystal ball, you will be
a giant, just close your little eyes
fall asleep nicely, little Blaise.

Tüzoltó leszel s katona!	You will be fireman and soldier!
Vadakat terelő juhász!	Shepherd who drives beasts!
Látod, elalszik anyuka --	You see, mummy falls asleep --
aludj el szépen, kis Balázs.	fall asleep nicely, little Blaise.

The poem to be analysed consists of the following linguistic communication units:

1.	Lehunyja kék szemét az ég	The sky closes its blue eyes
2.	lehunyja sok szemét a ház	The house closes its many eyes
3.	dunna alatt alszik a rét	under a wadded quilt the meadow is sleeping
4.	aludj el szépen kis Balázs	fall asleep nicely, little Blaise
5.	lábára lehajtja fejét alszik a bogár	The beetle bows its head on its feet, is sleeping
6.	a darázs	the wasp
7.	vele alszik a zümmögés	the buzzing is sleeping with it
8.	aludj el szépen kis Balázs	fall asleep nicely, little Blaise
9.	a villamos is aluszik	the streetcar is slumbering, too
10.	s mig szendereg a robogás álmában csenget egy picit	and while the rattling is dozing it tings a bit in its dream
11.	aludj el szépen kis Balázs	fall asleep nicely little Blaise
12.	alszik a széken a kabát	the jacket is sleeping on the chair
13.	szunnyadozik a szakadás	the tear is snoozing,
14.	máma már nem hasad tovább	today it does not stretch longer
15.	aludj el szépen kis Balázs	fall asleep nicely, little Blaise
16.	szundit a lapda	the ball is napping
17.	meg a sip	and the whistle
18.	az erdő	the wood
19.	a kirándulás	the excursion
20.	a jó cukor is aluszik	the good sugar is slumbering too
21.	aludj el szépen kis Balázs	fall asleep nicely, little Blaise
22.	A távolságot, mint üveggolyót megkapod	you will get the distance as a chrystal ball
23.	óriás leszel	you will be a giant

24. csak hunyd le kis szemed	just close your little eyes
25. aludj el szépen kis Balázs	fall asleep nicely, little Blaise
26. tüzoltó leszel	you will be fireman
27. s katona	and soldier
28. vadakat terelő juhász	shepherd who drives beasts
29. látod, elalszik anyuka	you see, mummy falls asleep
30. aludj el szépen, kis Balázs	fall asleep nicely, little Blaise

(In the subsequent lists I shall refer to these unit numbers.)

The analysis of the linear patterning consists of

- the individual analysis of the units of the structure,
- the compilement of different indices and
- the analysis of these indices.

2.1 The individual analysis of structure units is performed in the possession of the described conditions, according to the described algorithm. In neglect here the presentation of how the algorithm works with all the communication units. The lists -- as the results of the analyses -- serve informations in connection with the analysis, too.

2.2 Here I shall present some lists referring to the lexical units of communication units

2.2.1 Let us see first the alphabetic list of the word-forms of the poem

a	(the)	2,3,6,7,9, 10, 12, 12, 13, 16, 17, 19, 20, 22;
alatt	(under)	3;
álmában	(in its dream)	10;
alszik	(sleeps)	3, 5, 7, 12;
aludj el	(fall asleep)	4, 8, 11, 15, 21, 25, 30;

aluszik	(slumbers)	9, 20;
anyuka	(mummy)	29;
az	(the)	1, 18;
Balázs	(Blaise)	4, 8, 11, 15, 21, 25, 30;
bogár	(beetle)	5;
cukor	(sugar)	20;
csak	(just)	24;
csönget	(tings)	10;
darázs	(wasp)	6;
dunna	(wadded quilt)	3;
ég	(sky)	1;
egy	(a)	10;
elalszik	(falls asleep)	29;
erdő	(wood)	18;
fejét	(its head)	5;
hasad	(stretches)	14;
ház	(house)	2;
hunyd le	(close)	24;
is	(too)	9, 20;
jó	(good)	20;
juhász	(shepherd)	28;
kabát	(jacket)	12;
katona	(soldier)	27;
kék	(blue)	1;
kirándulás	(excursion)	19;
kis	(little)	4, 8, 11, 15, 21, 24, 25, 30;
lábára	(on its feet)	5;
lapda	(ball)	17;
látod	(you see)	29;
lehajtja	(bows)	5;
lehunyja	(closes)	1, 2;
leszel	(you will be)	23, 26;
máma	(today)	14;

már	(already)	14;
meg	(and)	17;
megkapod	(you will get)	22;
mig	(while)	10;
mint	(as)	22;
nem	(not)	14;
óriás	(giant)	23;
picit	(bit)	10;
rét	(meadow)	3;
robogás	(rattling)	10;
s	(and)	10, 27;
sip	(whistle)	17;
sok	(many)	2;
szakadás	(tear)	13;
széken	(on the chair)	12;
szemed	(your eyes)	24;
szemét	(his eyes)	1, 2;
szendereg	(dozes)	10;
szépen	(nicely)	4, 8, 11, 15, 21, 24, 30;
szundit	(naps)	17;
szunnyadozik	(snoozes)	13;
távolságot	(distance)	22;
terelő	(who drives)	28;
tovább	(longer)	14;
tüzoltó	(fireman)	26;
üveggolyót	(chrystal ball)	22;
vadakat	(beasts)	28;
vele	(with it)	7;
villamos	(streetcar)	9;
zümmögés	(buzzing)	7;

2.2.2 The next list consists of the lists of words arranged according to the syntactic categories. Let us see the nouns and verbs from these in a more detailed way.

Nouns:

-Common			Blaise
+Common	-Count	-Abstract	rattling, buzzing,
	-Count	+Abstract	-
	+Count	-Animate	dream, sugar, wadded quilt, wood, sky, head, house, coat, excursion, ball, foot, meadow, whistle, tear, eye, chair, distance, chrystal ball, beast, streetcar;
	+Animate	-Human	beetle, wasp;
		+Human	mummy, shepherd, soldier, fireman;

Verbs:

action:	sleeps, slumbers, rings, falls asleep, sees, bows, closes, gets, dozes, naps snoozes,
event:	stretches;
being:	be;
+transitive:	sees, bows, closes, gets;
-transitive:	sleeps, slumbers, rings, falls asleep, dozes, naps, snoozes, stretches, be;

2.2.3 From the semantic indices let us see a list of synonyms: the synonyms of the verb alszik (sleeps).[5] I wish to demonstrate the semantic connections among the members of the synonym-set with the representation of the meanings of these verbs found in the Hungarian Explanatory Dictionary:

alszik	(sleeps)	:	he (she) is in the state of sleeping
aluszik	(slumbers)	:	(only in indicative mood, present tense) sleeps

elalszik	(falls asleep)	:	gets into the state of sleeping
lehunyja szemét	(closes his-her-eyes)	:	(only in this word connection) sleeps for a time
szendereg	(dozes)	:	rests in easy sleep or halfsleep
szundit	(naps)	:	sleeps for a shorter time, not deeply
szunnyadozik	(snoozes)	:	sleeps silently, calmly, not deeply, with breaks

2.2.4 In morphological forms the poem is very poor. (I shall return to this later.)

2.2.5 To the demonstration of syntactic lists let us see first the list of the deep and surface structures of the communication units of the poem.

The first column shows the deep structures. In these deep structure representations the constituents signed by the sign '—' stand for the elements absent from the surface structures. The second column contains the indices of the surface structures. As a matter of fact we give here not the indices of the complete deep and surface structure, but only the categories of the sentence-base in a normalized form (first column) and in a succession according to the surface structure (second column).

1. NP + VP/V + NP/A+N// ;VP/V NP/A N// NP
2. NP + VP/V + NP/Num+N// ;VP/V NP/Num N//NP
3. NP + VP + Adv_{loc}/N+postp/ ;$AdvP_{loc}$/N postp/ VP NP
4. NP/A + N/ VP $AdvP_{mod}$;VP $AdvP_{mod}$ NP/A N/
5. NP + VP/V + NP/ +$AdvP_{loc}$;$AdvP_{loc}$ VP/V NP/ VP NP
6. NP + /-VP/ :NP
7. NP + VP + $AdvP_{soc}$/Pr/ ;$AdvP_{soc}$ /Pr/ VP NP
8. NP/A + N/ VP $AdvP_{mod}$;VP $AdvP_{mod}$NP/A N/
9. conj + NP + VP ;NP conj VP
10. /-NP/ + VP + $AdvP_{mod}$+$AdvP_{loc}$ + $AdvP_{temp}$/conj + NP + VP/; AdvP/conj VP NP/ $AdvP_{loc}$ VP $AdvP_{mod}$

11. NP/A + N/ + VP + $AdvP_{mod}$; VP $AdvP_{mod}$ NP/A N/
12. NP + VP + $AdvP_{loc}$; VP $AdvP_{loc}$ NP
13. NP + VP ; VP NP
14. /-NP/ + negVP + $AdvP_{temp}$ + $AdvP_{temp}$; $AdvP_{temp}$ $AdvP_{temp}$ negVP
15. NP/A + N/ + VP + $AdvP_{mod}$; VP $AdvP_{mod}$ NP/A N/
16. NP + VP ; VP NP
17. conj + NP + /-VP/ ; conj NP
18. NP + /-VP/ ; NP
19. NP + /-VP/ ; NP
20. conj + NP/A + N/ + VP ; NP/A N/ conj VP
21. NP/A + N/ + VP + $AdvP_{mod}$; VP $AdvP_{mod}$ NP/A N/
22. /-NP/ + VP/V + NP as NP ; VP/NP as NP V/
23. /-NP/ + VP/V + NP/ ; VP/NP V/
24. conj +/-NP/ + VP/V + NP/A + N//; conj VP/V NP/A N//
25. NP/A + N/ + VP + $AdvP_{mod}$; VP $AdvP_{mod}$ NP/A N/
26. /-NP/+VP/V + NP/ ; VP/NP V/
27. /-NP/ + VP//-V/ + NP/ ; VP//-V/ NP/
28. /-NP/ + VP//-V/ + NP/N + S_{adj}//-NP/+ VP/V + NP////; VP/ A/NP A_{verb}/ NP/
29. /-NP/ + VP/V+NP/S/NP + VPP///; VP/V NP/S/VP NP//
30. NP/A + N/ + VP + $AdvP_{mod}$; VP $AdvP_{mod}$ NP /A N/

The structures contained by this list can be put among the following lists of structure-classes:

2.2.6 The complete list of deep structure types:

NP + /-VP/ 6, 17, 18, 19;
NP + /VP/Y 9, 13, 16;
NP/A + N/ + VP + AdvP 4, 8, 11, 15, 21, 25, 30;
.....

2.2.7 The list of deep structures that contain some special immediate constituent. E.g:

containing $AdvP_{loc}$: 3, 5, 10, 12;

containing VP//-V/+NP/ : 27, 28;

.....

2.2.8 The complete list of surface structure types

VP/NP V/ 23, 26;

VP NP 13, 16;

VP $AdvP_{mod}$ NP/A N/ 4, 8, 11, 15, 21, 25, 30;

.....

2.2.9 The list of surface structures that contain some constituent at a given place. E.g:

containing NP in the last place 1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 15, 16, 17, 18, 19, 21, 25, 30;

beginning with AdvP 3, 5, 7, 10, 14;

.....

2.2.10 From the semantic lists the lists of deep structures containing incompatibilities (semantically incompatible constituents) are important.

In the NP + VP structures, for example, the incompatibility is between the syntactic-semantic markers of NP and the context conditions of VP.

NP/-Animate/ +VP/+Animate__/ 1, 2, 9, 12, 13, 16, 17, 18, 19, 20;

NP/+Abstract/ +VP/+Animate__/ 10;

.....

In order to discover the incompatibilities where the adverbial part plays a role, we generally have to take into account the thematical connections, too. E.g:

$$NP/-Animate/+VP/+Animate/+AdvP_{loc}/\begin{matrix}\text{referring}\\ \text{to +Human}\end{matrix}/\quad 3;$$

2.2.11 Besides the mentioned lists -- if the lexical units contain style-valuing as well -- we can make so called style lists, too.

The compilation of the different style classes of words and the incompatibilities influencing style markers are probably of primary interest.

2.3. It is likely that we have to make programs that are able to construct indices referring to aspect-combinations given optionally, but it is only some basic lists that have to be made at every analysis. The compilation of other lists may be necessary on the base of the analysis of these basic lists.

Over a certain quantity of text it is expendient to discover the fact of complete or partial concordance among the different structure units with a program. On the layer of communication units this can happen for example in a way that we order to every communication unit the signs of those classes (or lists) to which the examined communication units and/or their elements belong, and then we establish the 'relations of affinity' of these sign-sequences.

3. The description of linear patterning

It cannot be our purpose to collect all the structure elements in the description of the linear patterning. The restrictions of the scoope of description are always prescribed by a specific net of aims.

The most characteristic features of the linear patterning of the linguistic component of the analysed poem are as follows:

7 of the 30 communication units are the repeated returnings of the communication unit "fall asleep nicely, little Blaise" corresponding the title of the poem.

Not considering these, the first 20 communication units -- explicitly or implicitly -- contain verbal predicates of active voice, present tense, third person singular.

With the exception of the communication unit 14, in all of these communication units the synonyms of the verb "sleep" can be found as a predicate. In one case -- in the communication unit 10 -- it is as the predicate of an adverbial clause, but here there is an adverb "in its dream" besides. (Though it does not belong to the theme of a work of art centered analysis the following thing can be mentioned for the sake of interest: when the poem was first drafted, it did not have this varied array of synonyms. Originally in the communication units 10, 13, 16, 17, 18, 19, there was the verb "sleeps" everywhere.)

These predicates that require living agents, refer to inanimate subjects in all cases. The "subjects" are the last elements of the communication units nearly everywhere.

From the communication unit 22 on there are explicit or implicit predicates of second person, indicative mood, present tense, future tense and imperative mood.

Besides the verb-form of imperative mood in the repeatedly returning communication unit "fall asleep nicely, little Blaise" there is only one verb in the imperative in the whole poem, the one "close" in the communication unit 24. This "close" echoes the predicates "closes" in the first and second communication units. (The verb "closes its eyes" occurs no more times in the poem.)

Besides the imperative the attribute "little" also connects the communication unit 24 to the refrain. The adjective "little" occurs only here besides the refrain -- and the "your little eyes" harks back the constructions "its blue eyes" and "its many eyes" of the first two communication units.

In this group of communication units there is only one verb form of third person signular, in the communication unit 29: "falls asleep". This "falls

asleep" continues morphologically and semantically the predicates of the communication units of the first group on one hand, and it is the only occurence of the verb "fall asleep" besides the refrain on the other. Hence "fall asleep" is the verb only of "little Blaise" and "mummy".

In these communication units the "you will be something" predicates are dominating and no semantic incompatibility occurs.

The linear patterning -- as we can see -- partly suggests the skeleton of the hierarchal patterning, too, dividing the poem into two composition units of different character. But I do not wish to deal with the structure of the hierarchal patterning in this place.

4. Concluding remarks

In my paper I have dealt with the aspects of the analysis of verbal works of art by computer.

In wanted to show that with the application of the methods of the modern linguistics and with the use of computers not only the analysis of language but also a part of the work of art centered analysing activity of man can be modelled.

This possibility spares man the trouble of time-wasting collecting work that precedes the description of the structure, and at the same time carries out this collecting work independently of subjectivity. The application of these methods is useful first of all in the analysis of text structures of great quantity.

Thought I only deal with the questions of the analysis of verbal texts, in my opinion these problems have many common features with the analysis of other types of work of art, first of all the constructions of musical language or of dance language.

I think it very important to make the intensional characterization of the structure units presented here, too. This manner of the characterization makes it possible that the newest results of investigations relating to the tolerance spaces and the theory of classification should be able to be applied in this field,

too. As I see the problem, from the point of view of the computer analysis of both the language and the works of art these territories are first of all to be investigated.[6]

Notes

1. My paper entitled "On the Structural Linguistic Analysis of Poetic Works of Art" (Computational Linguistics VI. Budapest, 1967. 53-82.) deals with the questions of the building up of the text structure.

2. In connenction with the establishing of the structure of the dictionary entries Charles J. Fillmore's investigations are very interesting.

   In connection with the informations to be added to the dictionary entries of the Hungarian words you can have a general information in: "Reverse-Alphabetized Dictionary of the Hungarian Language" Comp. by F. Papp. Budapest, 1969.

3. In this paper I do not want to deal with the questions of the repetitive returns of text units larger than sentences.

4. In our Centre it is Dénes Varga and Éva Szöllősy who deal with the problems of morphological and syntactical analysis by computer.

   See: Dénes Varga, Problems of Machine Analysis, Linguistica Antverpiensia 1968. 2. 415-427; Postroenie novoj analizirujushchej sistemy predlozhenija, Nauchno-technicheskaja informacija, ser 2. 1968. 4.

   I examined the questions of the theorethical questions of the 'reokacenebt' of the deep structure in "Notes on the Semantic Interpretation of Verbal Works of Art" Computational Linguistics, VII. Budapest, 1968. 79-105.

5. Here it must be emphasized, that the theory of semantics should develop so far to be able to discover even the synonymic relations of the sentences, too. In case of success we shall be able to construct synonym lists on sentence layer, too. In connection with this question we think the investigations of Melchuk and Zholkovskij important.

6. See: E. C. Zeeman and O. P. Buneman, Tolerance spaces and the Brain in: Toward a Theoretical Biology -- Ju. A. Srejder, Matematicheskaja modelj teorii klassifikacii in: Nauchno-technicheskaja informacija, ser 2. 1968. 10.

COMPUTERS IN FOLKLORE RESEARCH

János S. Petőfi, Éva Szöllősy

Introduction

The application of computers in different fields of sciences and humanities usually has a twofold result. On the one hand, research methods are to be modified and that at the least means a much more precise formulation of the questions to be asked. On the other hand the processing of data becomes much quicker and much more objective than it could ever be without computerized methods.

As to the humanities, linguistics was the first to have created the possibility of subjecting the object of its research to computerized methods. It was not long before computers came to be applied in the analysis of verbal works of art and in the the analysis of verbal folklore works of art as well.

We do not intend to give here even the shortest summary or analysis of the methods that have been applied so far. What we would like to show is how the method outlined above could be applied in other fields of research - in ethnography, including non-verbal works of folklore art as well - and also what requirements should be met by the description and characterization of the 'objects of research'.

The principal idea of the method is the intensional characterization of the objects of our research and the automatic recognition of the relationships among the objects thus characterized.

As to the linguistic analysis, the intensional characterization of words can be carried out by human work only, that of any other structural unit of the language could be done automatically, i.e. by computers.

The variety of objects of ethnographic research will of course entail a variety of tasks to be solved.

1. On the objects of ethnographic research

To begin with, we must confess we are no experts in the field of ethnography. It seemed to be interesting, however, to see how methods applied in computational linguistics could be applied in the analysis of non-language-like systems or to put it more precisely in that of systems consisting of heterogeneous elements. The system of the objects of ethnographic research is a par excellence example of the latter.

Without striving at completeness we may state that ethnography deals with objects, verbal and musical texts and various actions ('distinguished day' customs, popular cures, etc.).

So the objects investigated are either primary texts (folk tales, ballads) or they were 'collected' and 'stored' in the form of verbal descriptions (incantations, faith cures, etc.) or they could be described verbally also in a more or less satisfactory way (houses, embroidery, shepherd's staffs etc.). Music and dance can also be described in their 'own' language.

Thus the analysis of the objects to be examined always means the processing of texts. (Let us now pass over the otherwise interesting point, namely that decoration elements could be stored and 'processed' in 'analogous form' - if we had a suitable computer.)

The processing of verbal texts is of course not quite the same as the linguistic analysis of verbal works of art. Now we are going to point out several specific problems connected with the examination of the different groups of objects of research - excluding questions concerning the analysis of music and dance language texts.

2. On the characterization of research objects of different types

2.1. The analysis and characterization of primary verbal texts, i.e. of the first group of research objects, is near to the analysis of verbal works of art. Up to the sentence level they are the same. In the analysis of the 'thematic' structure of texts the difference lies in the fact that themes and groups of themes in folklore texts (tales, ballads) are realizations of a finite number of patterns as it has been pointed out several times. The characters are representatives of 'types', what they do are 'typical actions'. All of these should be denoted when characterizing a text. (This way of characterization could best be compared to abstracting articles on chemistry whose relevant words are supplied together with their role indicators).

Thus the most important sets that we assign to the text as its characteristics are

- the thematic orderings of its words;
- the classes of the deep structures, surface structures and semantic interpretations of its syntagms and sentences;
- 'descriptive lists' of the characters, actions, those of the objects, places, etc. that have a role in the action (these lists could be compiled by computers);
- the thematic structure of the plot's development.

So this way of analysis requires that all the methods yielded by linguistics for the analysis of verbal works of art should be made use of by the analysis of folklore products.

2.2. When the research objects are stored in the form of verbal descriptions (descriptions of popular cures, customs, etc.) the analysis of the textual framework is only of mediating importance - except for those parts of the text that are of primary character (e.g. the text of an incantation, etc.).

The characteristics to be determined here are

- the characteristics of the primary parts included;
- 'descriptive lists' of the characters, actions and those of the objects, places, etc. that play a role in the action;
- the thematic structure of the action's build-up.

When developing an algorithm and program for the above purposes the given ways of description should of course be kept in mind, and automatic processing should be considered when new data are collected.

2.3. In connection with the analysis or rather with the preparatory tasks of analyzing objects (such as various utensils, needlework, etc.) a problem arises that - to a smaller or larger extent - concerns the above analyses as well.

Linguistic analysis - as we have seen - presupposes the existence of a proper vocabulary (thesaurus) and a proper set of rules. As for the rules, we could arrive at completeness but this is hardly true for the thesaurus. The only solution we could imagine is the compilation of a 'kernel' of the thesaurus. The kernel would contain the basic vocabulary and it could from time to time be completed by the specific vocabulary of the field the text to be analyzed belongs to.

Parts of the vocabulary of folklore texts also belong to the specific category and will not be included in the basic vocabulary. In order to determine their vocabulary entries - for their semantic characterization first of all - we have to have their KWIC index lists made on the basis of the texts given. This happens just as in the case of analyzing verbal works of art, together with the morphological analysis.

When characterizing objects (buildings, devices, decoration motives, etc.) the general part of their characterization is to write the appropriate vocabulary entry, whilst the specific part of the characterization consists of listing the specific characteristics of the object. The classes of the specific characteristics and the order of these classes should be made constant for every 'object-type'.

The characterization of the objects could not be carried out except by human work (it is analogous to the compilation of the vocabulary of the sector of definitions necessary to linguistic analysis).

3. On the analysis of the different types of objects

As far as verbal works of art are concerned the analysis of a single work of art is important in itself. With folklore products, however, the primary task is the analysis of the various sets of objects.

The object-centered intensional characterization of objects makes the following types of analysis possible:

3.1. The analysis of a type of objects -

Objects of the same type (tales, folksongs, objects of the same character, etc.) are characterized in the same way. Consequently, the ways they might be related to each other can be determined exactly as it has been shown for sentences of verbal works of art.

3.2. The determination of the following type of 'object systems' -

A complex object type- i.e. one consisting of heterogeneous elements may be related to several homogeneous object types. E.g. a folk tale may be related to popular beliefs by its plot, a distinguished day custom may be related to songs, dances, objects, etc.

On the basis of the characteristics that have automatically or non-automatically been assigned to an object, the automatic determination of the so called 'object systems' may be done by examining the object types that can be partially related.

3.3. The determination of different groups of objects and their relationships -

On the basis of the identifiers of different folklore objects various lists of object groups can be compiled. The list of folklore objects belonging to the same geographical unit will certainly be a distinguished one. These lists could be thought of as the folklore characteristics of the given geographical unit and the relationships among these folklore characteristics could again be examined by the above methods.

x x x

What we really wanted to show is that by a gradual and consistent abstracting process it is possible to determine such a structure of characteristics for widely varying objects that makes the application of a universal 'relation determining' program possible. Its elaboration and development, however, require a deeper insight into the problems of tolerance spaces.

PROBLEMS OF IMPROVING THE EFFICIENCIES OF PARSING SYSTEMS

Dénes Varga

1. The number of all the possible structures as a function of the sentence length

As soon as practical applications are considered the efficiency of the parsing method is of fundamental importance whether natural or programming languages are to be processed. The problem of efficiency arises because the relationship between the length of a string of symbols and the number of structures that may in theory be assigned to the string is far from being linear, the growing number of symbols entails a much more rapidly growing number of possible structures.

For CF grammars it is comparatively easy to determine how the number of structures depends on the length of the string. Considering binary branchings only and excluding the possibilities that arise from having different labels attached to one node

$$f(n) = \frac{1}{2n-1} \binom{2n-1}{n-1}$$

is the number of different trees that can be assigned to a linear string of n elements [1]. This means that for a string with 10 element the number of different trees is slightly less than 5000, for a string with 20 element this number becomes more than 1.75 milliard.

In order to include non-binary branchings as well I suggest the following recursive formula

$$g(1) = g(2) = 1$$

$$g(n) = 2 \left[g(2) \sum_{j=1}^{n-2} g(j) + g(3) \sum_{j=1}^{n-3} g(j) + \ldots \right.$$

$$\left. \ldots + g(n-2) \sum_{j=1}^{2} g(j) + g(n-1)\, g(1) \right] + 1$$

where n is the number of elements in the string. Accordingly, more than 100 000 different structures can be assigned to a 10 element string, and 1.6×10^{12} different structures to a 20 element string.

Let us stress again that what we have calculated here is the number of the essentially different derivations, i.e. the number of those yielding different results. The number of possible derivational paths for 10 elements is 18 times larger than the number of the different results, for 20 elements the number of paths is 750 times larger than that of the different structures

$$\left(\frac{20!}{2} \quad 1.2 \times 10^{18} \right).$$

2. Syntactic ambiguities

Natural language utilizes but a small fraction of the possibilities. As to the number of possible structures of concrete sentences, the syntactic restrictions are very strong yet far from sufficient to yield information for unambiguous assignement. The number of structures allowed by the formal syntactic rules is in most cases definitely larger than the number of structures a human being becomes aware of in the course of speech.

A well-known point is that unambiguity cannot always be ensured by grammatical means even for artificial languages whose structure is immensely

less complicated [2]. It is worth mentioning that the authors of ALGOL-68 decided to let some ambiguities remain in the language which could have been eliminated but by making the grammar a lot more complicated [3].

Where does the majority of syntactic ambiguities in natural languages come from?

1. There is a number of words with different ranges and vice versa: some words may fall within the scope of several different words and it cannot be determined by formal syntactic means - nor yet by semantic ones at times - whose range they really belong to. These two things often combine, especially in complex genitive constructions.

A fine Russian specimen of which is as follows:

... вследствие других законов сохранения
и особенностей взаимодействия частиц...

The corresponding string of symbols:

$$Pr^{g}\ A_{g}\ N_{g}\ N_{g}\ E\ N_{g}\ N_{g}\ N_{g}$$

The rules of reductions:

(i) $$A_{g} + \left\{ \begin{matrix} N_{g} \\ NP_{g} \end{matrix} \right\} = NP_{g}$$

(ii) $$\left\{ \begin{matrix} N_{g} \\ NP_{g} \end{matrix} \right\} + \left\{ \begin{matrix} N_{g} \\ NP_{g} \end{matrix} \right\} = NP_{g}$$

(iii) $$\left\{ \begin{matrix} N_{g} \\ NP_{g} \end{matrix} \right\} + E + \left\{ \begin{matrix} N_{g} \\ NP_{g} \end{matrix} \right\} = NP_{g}$$

(iv) $$Pr^{g} + \left\{ \begin{matrix} N_{g} \\ NP_{g} \end{matrix} \right\} = C$$

Apparently a number of different structures can be determined by changing the order of rule application.

2. Another source of syntactic ambiguities is that not even the string of symbols (categories) can always be unambiguously assigned to the sentence, i.e. homonymy may often appear on the morphological level. Homonymy arises either because formal differentiation between parts of speech is absent (as in English) or because the correspondence of the functions of words and the morphological means of expressing them is ambiguous, the morphological functions are not unambiguously expressed (it is tipical for many languages e.g. in Russian). Completely independent words with or without affices too can of course agree in form as well.

Still one seldom comes across a sentence that could be assigned several entirely different structures. Sentences of this type are usually puns or grammatical examples (cf. "Time flies like an arrow"). It is the so-called local syntactic ambiguity that normally troubles us, i.e. part of the sentence that can be assigned several different part-structures without influencing the remainder of the sentence-structure. Now if there are several locally ambiguous parts in the sentence and they are independent from each other, the number of ambiguities for the whole sentence will considerably increase: it will be the arithmetic product of the numbers of the independent local ambiguities.

3. Questions of tactics

The above numeric data clearly show how hopeless it is to simple proceed by checking on all the theoretically possible structures. But it is also apparent that syntax-directed parsing systems will fail in a considerable number of cases just because the sentence structure is syntactically undetermined [4]. The development of an effective analyzer is at least as much a mathematical as a linguistic problem.

The most important demands a parsing algorithm should meet are as follows:

(i) It should be able to determine all the conceivable parsings that a given sentence is assigned by a particular grammar.

(ii) It should be consistent in the sense that one parsing could not be arrived at but in one single way. (It should be a 'one-to-one algorithm'.)

(iii) In some way or other it should counterbalance the immense growth of the number of possible structures. Our goal to be approached is the linear relationship between the steps of the algorithm and the length of the sentence.

The efficiency of the algorithm depends considerably on factors that are independent of the particular method one has chosen to apply. These problems arise with any algorithm even if in different forms. The most important 'tactical' questions of this type are as follows:

(i) Assuming a large set of rules how does the algorithm select the rules that are to be (may be) applied?

(ii) How does it check for the conditions of applying them?

(iii) How does it recognize 'blind alleys' i.e. illegal paths (if any)?

(iv) How does it return from the illegal path to the legal one (or to the one that has not proved to be illegal as yet)?

Some of the well-known methods for solving (i) are as follows:

(a) Each rule is explicitly assigned the set of rules by which it could be continued. But choosing this method for a complicated grammar with a large number of possibilities one might face troubles.

(b) The rules are divided into several groups on the basis of different characteristics such as the number or the character of the symbols within the rule etc. Searching is then carried out within a comparatively small set of rules.

(c) Each symbol is assigned a set of all the rules this particular symbol appears in. Assignment can be done according to the position numbers within the rules. The so called initial symbols, i.e. symbols in first position, play then a distinguished role in the rule selection.

Whatever method one applies one may choose one of the several possible ways of practical realization. In case of (c) the choice made will be of immense importance (e.g. rules arranged in matrix form, chainlike representation etc.).

Problems (i) and (ii) are strongly interconnected. How are we to decide whether the conditions of applying a rule are met?

In the case of CF grammars checking could be carried out quite easily. For top-to-bottom analysis all we have to do is the identification of the left-hand side symbol of the rule. For bottom-to-top analysis based on normal form CF rules (i.e. binary branchings) only, once again it is not too difficult to check a twodimensional table for the possibilities of connecting a pair of symbols.

If CS or general form CF grammars are applied, the problem is not trivial at all, it turns out to be that of identifying strings of symbols. It could of course be solved in a trivial way but this would require an awful lot of work to do. B. Dömölki has developed an elegant method that would examine a whole series of rules simultaneously. The checking is performed on Boolean vectors, and the point Dömölki has made an excellent use of is that computers carry out logical operations on all the bits of a machine word at the same time [5].

Two subproblems connected with checking rules should be discussed:

(a) When should it start at all? Suppose that the symbol string is processed in sequential order (left-to-right or right-to-left) and a possibly applicable rule or a given context should be checked for. Then we could either go back to symbols that have already been examined (and check them repeatedly when checking for the applicability of various rules) or have already begun and completed certain examinations so that we have finished checking by the time its result is needed. (The second solution could of course be applied only if an appropriate mechanism automatically provides the checking for the conceivable conditions and the (gradual) cancelling of the non-realizable possibilities.)

(b) Is some kind of an additional examination necessary before the checking is completed? Namely it might turn out that the whole checking was superfluous because its result cannot be used later on or it will not lead to a correct result.

We have come very near to (iii), i.e. to the problem of how the occasional impasses (blind alleys) could be recognized in the course of the analysis? This is a cardinal problem concerning the efficiency of automatic analysis. The growing length of the sentence (symbol string) entails not only a growing number of possible structures but the number of inappropriate part-

structures growing as well. These 'torsoes' correspond to certain parts of the sentence but are incompatible with the remainder of it. What is more, the longer a sentence the more levels it may have i.e. the deeper its structure can be. This holds for the blind alleys as well: the longer the sentence the deeper the blind alley can be, the more branches and the more valid elements it may contain. Sentences that are monosemantic though syntactically ambiguous could be thought of as bottomless blind alleys not yet explored whose exploration needs either a wider context or the use of interrelationships not contained in the text.

The problem once again becomes twofold:

a/ What is the criterion of having got into a blind alley?

b/ How could we prevent getting into a blind alley at least in some cases?

The answer to these questions may be different, of course, for each algorithm and plays a subordinate though extremely important role regarding the "strategy" applied.

Just to give an example I would like to mention an elegant method of defining and "calculating" the criterion of blind alleys using an algorithm based on operations with Boolean vectors. Dömölki [5] -- who condenses the information related to the hypothetically accepted part structure and to the given symbol string under processing into a state vector defined recursively -- applies the following criteria to determine the impossibility of continuing the analysis along the given line

$$(T(Q_t) \vee B) \wedge H\left[x_{t+1}\right] = 0$$

Accordingly, the new symbol x_{t+1} to be processed may neither continue the paths the previous vector of state contained that have proved possible so far, i.e.

$$T(Q_t) \wedge H\left[x_{t+1}\right] = 0, \quad \text{nor begin a new rule, i.e.}$$

$$B \wedge H\left[x_{t+1}\right] = 0 .$$

The only handicap of Dömölki's method is that impasses can be recognized only after the algorithm has got into them -- the algorithm cannot pick out the paths that will lead into an impasse later on. So we have modified the algorithm and instead of using Dömölki's vector B - that would 'activate' the first position of each of the rules - we let only those of the rules become active that provide (direct or indirect) continuation of the paths that have already proved to be legal [6].

Experience so far shows three practical methods of an at least partial avoidance of impasses: (i) the consideration of the context; (ii) the use of the transitive connectivity of the rules; (iii) the checking ahead the number of symbols not yet processed.

Taking into consideration the context means making use - if possible -- of only one direction of the context to avoid the repetition of the tests performed. Today such analyzing grammars play an important role in the analysis of artificial languages [7].

In my opinion the use of the transitive joining of rules has yet many important possibilities to offer. P. Z. Ingerman's analysis is a good example of experiments in this direction [8].

Taking into consideration the number of symbols not yet processed, saves the analysis many unnecessary tests. There have been attempts at doing a preliminary global analysis of the complete symbol string on this basis to assess in advance the possibilities of each path of the analysis [9].

Finally let us mention the question as the last of the questions of tactics:

(iv) How to find the way from an illegal path back to a legal one?

This is the task that must somehow be solved by the parsing algorithm. So it is not enough to give a sign or "flag" at the points where the decision may perhaps be a failure. (i) It must be ensured that the state prior to committing the error is reconstructed. (ii) It would be advantageous to return to the state immediately prior to committing the error thus avoiding unnecessary delays.

(Nevertheless, there exist fine algorithms with no assurance that every error could be corrected. One of them is the well-known 'compiler compiler' that would never reinterpret a part of the symbol string if the part has once been accepted, consequently it is unable to recognize certain structures.)

One of the possible solutions to the problem in question is to have the "current state" of the analysis stored whilst proceeding so that it could be accessed later on. What we have termed "current state" here may include all the half-finished and abandoned rule applications that could be continued only after other rules have been applied. Whenever reaching back for a previous "current state" the possibilities that have ceased to exist in the meantime can always be cancelled. The techniques followed for practical realization may vary depending on the amount of information to be stored, on the memory area available for the working fields, etc. (In most cases some kind of a push down store is applied.)

4. The strategy of analysis I.

The problems mentioned so far are common in varying degrees for all parsing systems, the ways they are solved have no decisive influence on the whole flow of analysis (though they are of decisive importance as far as efficiency is concerned).

T.V. Griffiths and S. R. Petrick's classification [10] of the types of parsing systems is based on two considerations (whilst stressing that 'some procedures are described in these terms only with difficulty' and 'others seem to allow no such classification'):

(i) In what direction does the parsing proceed - is it a top-to-bottom or a bottom-to-top analysis? (The third type mentioned - 'direct substitution algorithms' - is a subclass of the bottom-to-top algorithms.)

(ii) Does the algorithm apply any means of a preventive reduction of the number of blind alleys, i.e. for increasing the 'selectivity' of the algorithm?

Their most important findings concerning the efficiency of the different types of algorithms are as follows:

(a) Algorithms proceeding from top to bottom - especially those of the direct substitution type - are the more efficient ones.

(b) Methods of increasing selectivity are of no special importance in the case of top-to-bottom analyses but they do considerably increase the efficiency in the case of bottom-to-top analyses.

(c) Efficiency is demonstratably influenced by the asymmetry (left-branching or right-branching) of the structure to be analyzed. In the case of analysis proceeding from top to bottom it is influenced in the reverse direction if compared with the analysis proceeding from the bottom upwards. (We assume that the analysis proceeds either from right to left in both cases of from left to right in both cases.)

They considered the parsing time of the following sentence types:

ab^n	a^nb	a^nb^n	ab^ncd
left-branching	right-branching	embedding	compound
('regressive' in Yngve's term)	('progressive' in Yngve's term)		(left-branching with respect to recursivity)+

S → Ab	S → aB	S → aSb	S → AB
A → Ab	B → aB	S → ab	A → Ab
A → a	B → b		A → a
			B → Bd
			B → bc

Parsing time as a function of sentence length increases - according to Griffiths' and Patrick's data - as follows:

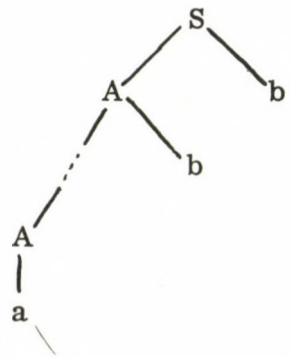
+The grammar given would have allowed right recursivity as well (ab^ncd^m) but in the measurements the above restrictions of grammar are dealt with only.

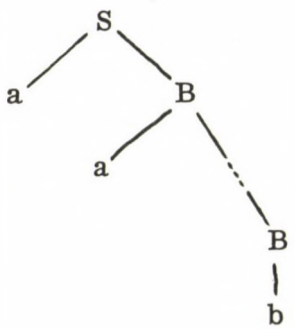
ab^n

	non-select.	selective
top-to-bottom	quadratic	
bottom-to-top	linear	

a^nb

non-select.	selective
linear	
exponential	linear



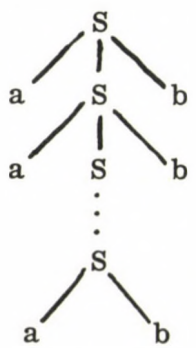


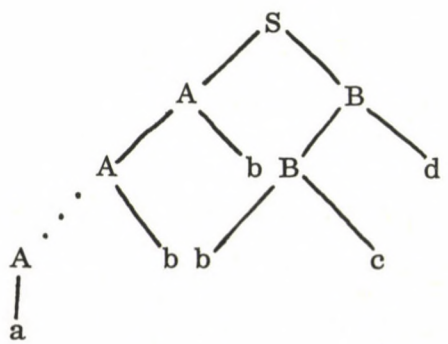
a^nb^n

	non-select.	selective
top-to-bottom	linear	
bottom-to-top	exp.	linear

ab^ncd

non-select.	selective
exponential	
exp.	cubic





According to Griffiths's and Patrick's data it is the bottom-to-top selective parser alone that is able to analyze sentences of the last, comparatively simple type of grammar with a better than exponential efficiency.

What are the underlying reasons for the results obtained by Griffiths and Petrick?

(i) Bottom-to-top algorithms are characterized by the fact that they take their start from what actually exists instead of looking for what "could be" [11].

In the case of exceedingly extensive grammars the top-to-bottom analysis must work with a huge number of potential possibilities and the elements of the symbol string to be analyzed will but slowly filter out the possibilities that may not be realized.

(ii) Selectivity, in the sense Griffiths and Petrick use the term, does not influence all this to any degree as the filtering on the basis of a precedence-matrix extends only to testing the first element. It will be shown later on that selectivity can be considerably increased and, going even further, it could be made the basis of the strategy of the analysis.

(iii) In the case of bottom-to-top analysis the situation is entirely different. Here the seemingly identical apparatus works with a much greater efficiency. But (a) the "look ahead" condition suggested by Bastian [12] (i.e. the possibility of the resultant symbol achieving its aim checks the compatibility) one level higher up and the distance from the top is so much less. (b) Here only such rules are to be realized in which all the components can be found, the others are omitted in the course of the rule controls. It is out of the question therefor to regard this selectivity as analogous to the top-to-bottom selectivity that is based on the first symbol of the lowest level.

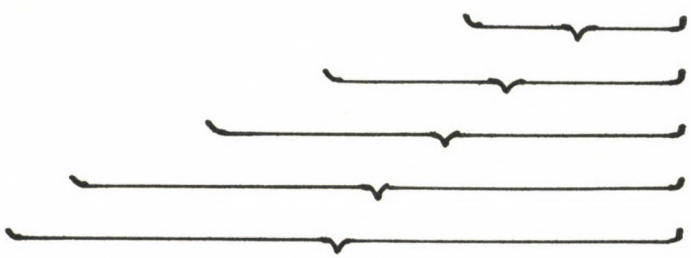
(iv) Griffiths's and Patrick's measurements of the effect of the asymmetry of sentences on the efficiency of the analysis are a practical justification of an observation I made in 1964. In an article about Yngve's hypothesis [13] I developed the idea that for languages that have mostly "progressive" (right-branching) structures it is the right-to-left analysis that is more effective in the case of analysis from bottom to top. (The right-to-left analysis is equivalent of course to a left-to-right analysis in a system that is a mirror image of the original.)

In case of pure structures the explanation of the phenomenon is simple: In a right-branching structure the number of erroneous linkings is started at the end of a sentence. Let us take the example from the above mentioned article of mine:

Вы знаете много теорем о пределах.

Its processing from right to left is very simple:

Вы знаете много теорем о пределах



If, however, the analysis is started from the beginning of the sentence we get erroneous (or incomplete) linkings one after the other

Вы знаете
знаете много
много теорем

In the case of complex structures the situation is more complicated. In this case the effectivity greatly depends on the method used for eliminating the impasses.

(On the disadvantages of vertical analysis see the next paragraph.)

5. The strategy of analysis II.

When determining the type of analysis apart from its starting point it is also very important to know along what paths the analysis proceeds towards its goal, or in other words in what sequence the tests are carried out together with the inseparable question of in what form or structure the part-results are stored.

On the basis of these considerations there are two basic types of parsers. In theory this classification is independent of the fact whether the analysis proceeds from the bottom upwards or from the top downwards.

(i) Those parsers that proceed with "maximum width" from level to level working on the whole symbol string, first produce all the reductions that may be achieved by applying a single rule, then those that may be obtained by applying two rules and so on until the part-structures thus obtained are gradually linked. (In the analysis that proceeds from top downwards, these correspond to the derivations produced by applying two, there, ... rules, followed by the comparison of the terminal symbols thus obtained with the symbol string being analyzed.)

(ii) The parsers that proceed with a "minimum width" and the "steepest slope", while gradually extending the elements of the symbol string take the first opportunity to apply a rule and will not extend the analysis to a new symbol until there are new rules that could be built on the rules applied so far.

We could mention as an example for the first method the Sakai-Nagao algorithm [14] [15] the Cocke algorithm [16] or its application by Kuno to context sensitive languages [17] (the same strategy is applied by Vauquois in his analysis of Russian). The algorithms by Woods [18] by Borscsev [19] and the Dömölki-Varga algorithms [5] [6] are examples of the second method.

Both methods have their advantages and disadvantages. It will perhaps be useful to draw the attention to them.

The great advantage of the analysis that proceeds from level to level is the ease with which in case of appropriate storage the part-analyses that could be continued along the same line, are contracted (see Griffiths-Petrick: "Merging similar sections of different TM [Turing Machine] paths").

Its disadvantage consists in the fact that

a/ relatively large number of independent part structures has to be stored,

b/ it needs relatively lengthy tests to determine whether the individual part structures are compatible.

The strategy of "maximal hierarchization" is more advantageous beyond doubt as far as economy in storage is concerned because in this case a single push down store will suffice to store the results and all the paths that have proved incorrect may be removed once and for all from the push down store together with all the derivations. This principle may be formalized as follows.

Let us denote according to the inverse Polish notation the result of the rule applied to the elements

$a_k\ a_{k+1} \cdots\ a_{k+r}$ with the result B_m as

$a_k\ a_{k+1} \cdots\ a_{k+r}\ B_m^r$. In other words let the elements of the symbol string that we applied the rule remain in the symbol string and let us simply add to the end of the string the symbol obtained as the result of the rule application.

Accordingly the resultant symbol string will be

$$\min_i\ a_1 \ldots a_i\ B_1^{r_1} \qquad r_1 \leq i$$

after applying the first applicable rule.

Let us suppose that there are at most m-1 more applicable rules following the first one while no new symbol is read ($m \geq 0$)

The symbol string will become

$$\max_m\ \min_i\ a_1 \ldots a_i\ B_1^{r_1} \ldots\ B_m^{r_m}. \qquad r_j \leq i$$

While continuing the application of this principle the symbol string will be increased by new terminal and non-terminal symbols:

$$\min_j\ \max_m\ \min_i\ a_1 \ldots\ a_i\ B_1^{r_1} \ \ldots\ B_m^{r_m}\ a_{i+1} \ \ldots\ a_j^{r_j}\ ;$$

$$\max_{n}\ \min_{j}\ \max_{m}\ \min_{i}\ a_1 \ldots a_i B_1^{r_1} \ldots B_m^{r_m}\ a_{i+1} \ldots a_j B_{m+1}^{r_j} B_n^{r_n}$$
$$\ldots\ldots$$

If the analysis gets into an impasse and cannot continue, then we have to return to the symbol $B_s^{r_s}$ last applied, remove it and continue the analysis applying the above principle. (First an attempt is made at applying another permissible rule in the same place and only if this fails shall we take a new a_t symbol and continue the analysis.)

The return from an impasse always means the deletion of the last non-terminal symbol and the reconstruction of the symbol string following it. (We would like to mention that this principle of analysis may be quite easily adopted for the analysis of context-sensitive languages as well).

This undoubtedly elegant principle of application produces the first possible analysis relatively rapidly, in its canonic form.

The increased selectivity of the analysis gives us a procedure that could be very well used in practical applications. Going further after having obtained the first analysis if the analysis is continued on the same principles (just as if the first correct analysis were in an impasse) all the other analysis may be likewise produced.

The disadvantages of the applied strategy of analysis are as follows:

(a) If right at the beginning of the analysis we have taken an incorrect path, then the correction of this error may only be done after all the following and in part independent applications of the rules have been deleted. This means that the correct, or perhaps the only possible partresults are lost: after putting the error right they have to be re-generated.

(b) The position is somewhat similar as far as the erroneous part-results are concerned: the analysis may get into a "local" impasse several times.

(c) A new, different system of storage and searching must be provided if we wish to ensure a new generation of the identical <u>continuations</u> -- supposing that <u>previously</u> some kind of a change took place in the determined structure.

6. A new strategy suggested for analyzing CF languages

The exponential increase in the time of analysis in various systems of analysis is obviously due to the increase in the number and depth of impasses, to their various branches -- in short to their dangerousness increasing with the length of the symbol string.

This is the dangerous point I tried to dodge by elaborating a parsing system that applies selectivity not as an additional device for increasing the efficiency of some method but as an independent method itself.

The linearity of the increase in the process of analysis may be best achieved if the symbol string to be analyzed can be segmented in accordance with the highest level rules applicable and these parts could be analyzed separately. If several parsings can be assigned to any of these segments (cf. what we have said about homonymy on p. 73) the structures corresponding to the whole sentence can be produced from the local part-results by combinatorical means.

The algorithm requires the following apparatus:

(i) The matrix of the direct inter-rule linkages (matrix A (r, i, r'))

(ii) The matrix of the direct rule-symbol linkages (matrix A (r, i, t.))

(iii) The connectivity matrices for the beginning symbol of the i^{th} component of the rules (matrix B_i (t, r))

(iv) The connectivity matrices for the middle part of the i^{th} component of the rules (matrix C_i (t, r))

(v) The connectivity matrices for the end symbol of the i^{th} component of the rules (matrix E_i (t, r))

(vi) The matrix for checking the component number of the rules.

The parser accomplishes the junction of departure from both directions from the top and from the bottom of the tree as follows.

The appropriate row of matrix (i) gives the total set of the applicable rules for the next top-to-bottom step in the parsing process. Matrices (ii)-(vi) serve for the filtration of these possibilities by the help of the sequence of

terminal symbols, i.e. from the very bottom of the tree. A step of the algorithm means a segmentation of the actual terminal string (which is a proper or improper substring of the original string to be analyzed) to several substrings as its components. This process goes on recursively with cyclical application of the segmenting routine until the terminal elements of the segments are approached (matrix (ii)). The input of every application contains two kinds of information:

(a) which substring of the original string is to be segmented,

(b) which component of which rule can the given substring be?

On the basis of this information the first possible segmentation (and the corresponding rules) are choosen where the conditions of the selectivity for the beginning, middle and end symbols of every component are fulfilled. This checking is accomplished by the help of matrices (ii)-(v). Matrix (vi) serves for getting the rule(s) with minimal number of components which meet(s) our conditions and contain(s) all the terminal symbols of the given string in the given sequence of components.

The resulting part-segmentations are stored partionally as well for checking the necessity of the next segmentation.

In case of two or more possible continuations in the analysis the canonic one is applied and the others are stored for self-correction and/or for revealing all the possible structures of the given string.

By cyclical continuation this process either we arrive finally at the terminal ending in case of all components or the given segmentation is found to be incorrect.

In case of incorrect segmentation first the permissible branches of the latest segmentation are tested by the algorithm. In our experience the selectivity of the system is considerable. Therefore even the storage of relatively small quantity of information allows a rapid examination of all the possibilities.

During segmentation we apply a "principle of segmentation" that is analogous to the principle discussed in connection with the "maximum hierarchization": the shortest component that is nearest to the beginning of the segment or to the end of the previous component, is taken and used until it becomes evident that for some reason the given segmentation is not applicable.

In this case an attempt is made at solving the situation by shifting the last border of segmentation to the right: only if this leads to no result, is the previous border of segmentation changed. The high effectivity of the method applied is due to

a/ making best use of the bottle-neck for the reduction in analyzing time;

b/ the fact that the tests for the possibilities of various part-segmentations can be quickly performed;

c/ the possibility of testing each segment in complete separation from all the other segments;

d/ the fact that the two-side approach leads to much fewer unnecessary part results than either Cock's or the well-known top-to-bottom algorithms.

Bibliography

1. Berge, C. Théorie des graphes et ses applications. Dunod, Paris, 1958.
2. Ginsburg, S. The Mathematical Theory of Context-Free Languages, McGraw Hill, New York, 1966.
3. Algol-68. MR 95. (mimeographed)
4. Shrejder, Ju. A. Teorija tolerantnosti, Nauchno-Texnicheskaja Informacija, Ser. 2
5. Dömölki, B. Voprosy sintaksicheskogo analiza dlja formal'nyx jazykov, Computational Linguistics 5, pp. 41-93.
6. Varga, D. Problems of Machine Analysis, Linguistica Antverpiensia II. pp. 415-428.
7. Knuth, D.E. On the Translation of Languages from Left to Right, Information and Control Vol. 8, No 6, pp. 607-639.

   Kaufman, V. Sh. O raspoznavanii nekotoryx svojstv kontekstno-svobodnyx grammatik, I-ya Vsesojuznaja konferencija po programmirovaniju, Kiev, 1968.
8. Ingerman, P. Z. A Syntax-Oriented Translator, Academic Press, New York, 1966.
9. Unger, S. M. A Global Parser for Context-Free Phrase Structure Grammars, Comm. ACM, Vol. 11, No 4, pp. 240-247.

10. Griffiths, T. V., Petrick, S. R. On the Relative Efficiencies of Context-Free Grammar Recognizers, Comm. ACM, Vol. 8. No 5, pp. 289-300.

11. Cf. Vakulovskaja, G. V., Kulagina, O. S. Ob odnom algoritme sintaksicheskogo analiza russkix tekstov, Problemy kibernetiki 18, p.218.

12. Bastian, L. A Phrase-Structure Language Translator, AFCRL Rep. 62-549, AF Cambridge Research Labs., Bedford, Aug. 1962.

13. Varga, D. Yngve's Hypothesis and Some Problems of the Mechanical Analysis, Computational Linguistics 3, pp. 47-72.

14. Sakai, I. Syntax in Universal Translation, Proc. 1961 Internat. Conf. on MT of Languages and Applied Language Analysis, London, 1962, pp. 593-608.

15. Nagao, M. Studies on Language Analysis Procedure and Character Recognition, Kyoto University, 1965.

16. Cf. Hays, D. G. Automatic Language-Data Processing, Computer Applications in the Behavioral Sciences, Prentice-Hall, Englewood Cliffs, N. J., 1962, pp. 394-421.

17. Kuno, S. A Context-Sensitive Recognition Procedure, NSF-18, Aug. 1967. VII-1-28.

18. Woods, W. A. Context-Sensitive Recognition, NSF-18. Aug. 1967. VIII-1-23.

19. Borscsev, V. B., Efimova, E. N., O sokrashchenii perebora pri sintaksicheskom analize, Nauchno-Texnicheskaja Informacija, 1967, No 10, pp. 26-33.

REVIEWS AND MISCELLANOUS

NOTES ON BOOKS

1. Erich Mater: Deutsche Verben, Band 1 - 4, VEB Bibliographisches Institut, Leipzig 1966 - 1968.

The four volumes under review list German verbs according to the following principles: Volume 1 is the alphabetic list of all German verbs, Volume 2 groups the verbs according to simple verbs and their derivatives, Volume 3 list the verbs according to whether they are simple verbs, verbs with a single or double prefix, compound verbs with two, three or four constituent morphemes, finally Volume 4 classifies compound verbs according to various principles all of them being morphological in nature. In Volume 1 we may find some statistical data concerning the source of the material, Volume 2 includes data about simple and compound verbs and Volume 4 about the various groups of compound verbs. These seem to be by-products of the compilation process.
A further six volumes will be published in the near future. They will contain verb lists grouped according to flexion class (Volume 5), rection (Volume 6), the relation of the verb to the reflexive pronoun (Volume 7), the preterite with sein or haben, (Volume 8), prefix separation (Volume 9) and the type of derivation (presumably from non-verbal stems, Volume 10). As all word or morpheme lists these lists, too, do not want to explain anything. Their main task is to provide the linguist with material, to facilitate his work. Such lists

are indispensable for linguists interested in computational linguistics. Any adequate "formal" dictionary must rely on such lists. Thus, the importance of Mater's undertaking cannot be questioned. However, one may ask questions as to the principles, or as he puts it, as to the categories of the classification. At first glance, there seems to be a certain redundancy in having a separate volume for the various preterite forms of the German verbs. In the same way, one may also ask why compound verbs are not treated in a single volume. These remarks do not want, of course, question the usefulness of any of the given lists. On the other hand, however, one would like to have some more lists, e.g. list of the verbs according to syntactic properties (types of complements) or some basic semantic properties (number of arguments, for example). Morphology is probably a good starting point but it goes without saying that it would not satisfy anybody.

2. Noam Chomsky - George A. Miller: L'analyse formelle des langues naturelles, Gauthier-Villars (Paris) and Mouton and Co. (Paris and The Hague), 1968.

The book is the French translation of Chapters 11 and 12 of Handbook of Mathematical Psychology (R. D. Luce, R. R. Bush, E. Galanter, eds.), Volume II. John Wiley and Sons, Inc., New York, 1963. Despite the important advances in generative-transformational grammar since 1963 this sketch of algebraic linguistics has not lost anything of its timeliness. First of all, it is a good introduction to the formal aspects of the Chomskyan linguistic theory. It explains to mathematicians what linguistic theory is after and it may introduce linguists (though they may need some help in the understanding of the more formal parts, I am afraid) to the formal aspects of linguistic theory. On the other hand, though linguistic theory has undergone essential changes since this study was written, the elaboration of the formal aspects has not kept up with this development, in the first place, in all likelihood, because of the apparent difficulties in formalizing transformations and the semantics of natural language.

In other words, this study must still be considered to be the best or one of the best starting points. One must know it. The fact that so far as I know, the original book on mathematical psychology is out of print - and, besides, it also contains other papers of less interest to the linguist, makes the French edition still more valuable.

3. Solomon Marcus: Introduction mathématique à la linguistique structurale, Monographies de linguistique mathématique, Vol. 1. Dunod, Paris, 1967.

This book is a reworked and essentially expanded version of the author's book in Rumanian entitled "Lingvistică matematica" and it is a counterpart to an English book by the same author entitled Algebraic Linguistics: Analytical Models (Academic Press, New York - London, 1966). It is a counterpart in the sense that despite the essential overlap, it concentrates more on phonological and morphological models, and it is an introduction for both mathematicians and linguists, though it is primarily intended to be an introduction for linguists. (Mathematician will probably find this book a little bit too redundant which does not hold with respect to the English book quoted above.) Marcus explains all the mathematical notions he makes use of and he illustrates all his mathematical machinery with ample linguistic examples. While the mathematician is advised to read Marcus's Algebraic Linguistics, the linguist will undoubtedly profit more from the Introduction mathématique. The book under review treats analytical models only. The linguistic background for the mathematical "micro"-models developed by Marcus must be sought in the various structuralist schools: Harris, Martinet, Tesnière, Hjelmslev, Jakobson are often quoted. This book should be considered to be a collection of able formalizations of notions and procedures developed by the various structuralist schools. One should not look for an underlying linguistic theory with the claim for an adequate description and explanation of all essential aspects of human language on the basis on some hypotheses about the functioning of language. There is no "integrated" theory here either generative or analytic. Thus one could criticize Marcus's books on

the same grounds as one criticizes the structuralists' methods today. But this is not the point here. In my view Marcus's book has - apart from its outstanding pedagogical value - two merits. On the one hand, by formalizing wellknown linguistic notions he can easily show that different formulations may lead to the same result or that apparently identical contentions when carefully formalized may reveal essential divergences. In other words, the careful reader will get a clearer picture about what is common and what is different in the various structuralist schools. In addition, Marcus's works links up with several problems of computational linguistics and must therefore be recommended to "computational" linguists as well.

One last word about the edition. The series lauched by Dunot testifies to the growing interest in France in "mathematical linguistics". It is remarkable that the authors of the first three volumes come from socialist countries. They represent trends in mathematical linguistics which are less known to the linguists and mathematicians in the Western hemisphere.

4. I. I. Revzin: Les modèles linguistiques, Monographies de linguistique mathématique Vol. 2., Dunot, Paris 1968.

This is a slightly revised and adapted French version of the author's book in Russian Modeli yazika, published in 1962. In fact, Revzin's book was at that time the first to treat general as well as particular questions of modelling in linguistics. Revzin's main ideas (and the gross of the book) are based on Kulagina's set theoretic model with several essential improvements and extensions. In other words, Revzin represents approximately the same line in mathematical linguistics as Marcus does. He puts, however, more emphasis on the linguistic side than on the purely mathematical one. Apart from this "set theoretic line" Revzin discusses also several other models like Yngve's depth hypothesis and transformational grammar (generative and non-generaitve alike). As the author notes in his forword to the French edition he has since changed his ideas on modelling in several essential points which would make necessary

the reformulation of quite a few important notions developed in the book under review. It is regrettable that the French edition appeard with a six years' lag after the Russian one. We all know what has happened to transformational theory during the last few years or how much our views on computational linguistics have changed. Revzin's book - if critically read - can still be of some use to the beginner in this field.

5. I. I. Revzin: Metod modelirovania i tipologia slavyanskih jazikov, Izdatelstvo Nauka, Moscow 1967.

Revzin's methods of modelling are based on what is usually termed the set theoretic approach. In this sense this book may be considered as a continuation of his first book on linguistic models. His approach is essentially analytic though he discusses several problems of generative grammar as well. Thus Revzin's basic attitude with respect to modelling seems to be the same as before. Within the given framework, however, he makes some essential corrections and also some important changes. So, for example, he takes a different stand on the modelling of syntax. His syntactic models are no longer based on (distributional) properties of words but rather on "syntactic groups". In this way, as the author puts it, syntactic models can be brought closer to the views expressed in traditional linguistics. He also abandons the basic notion of an infinite set of well-formed phrases as a primitive notion, a prevailing feature of previous works of this type. Revzin now takes two finite sets of phrases: one is the set of real phrases, the other the set of a priori forbidden phrases. On the basis of these notions he redefines and reconsiders all the notions and machinery known from his first book and introduces several others.

In the first two chapters Revzin puts forward his ideas about modelling in linguistics. He explains the formal means he makes use of in the course of the discussion of the various models. Then he proceedes to elaborating on some

of the applications of the developed models to Slavic languages. He also includes a chapter on semantics.

On the whole the book shows clearly the possibilities as well as the limits of the analytic-set theoretic approach.

F. Kiefer

Z. Pawlak, Gramatyka i matematyka. Panstwowe Zaklady Wydawnictw Szkolnyoh, Warszawa, 1965.

(З. Павляк, Грамматика и математика. Варшава, 1965, ос.112)

В последнее время в лингвистике наблюдается стремление к использованию точного математического аппарата для описания лингвистических понятий. Напомним, что первые попытки математизации грамматики языка были предприняты польским логиком К. Айдукевичем (около 1929 г.). Его идеи развивал и применял позже Бар-Хиллел. В связи с машинным переводом возникло недавно в лингвистике много разных концепций языка. К наиболее интересным принадлежат пожалуй работы Н. Хомского и В. Ингве. Постепенно оформилась самостоятельная дисциплина науки, так называемая математическая лингвистика. В настоящее время математическая лингвистика наиболее интенсивно развивается в СССР и США.

Книга З. Павляка, профессора Варшавского университета (математическое отделение), является хорошим введением в математическую лингвистику. Состоит она из следующих разделов:

I. Вступительные замечания (Pojecia wstepne)
II. Простые грамматики (Gramatyki proste)
III. Категориальные грамматики (Gramatyki kategorialne)
IV. Грамматики с конечным числом состояний (Gramatyki skonczenie stanowe)
V. Описание действий (Opisywanie czynnosci)
VI. Генетические грамматики (Gramatyki genetyczne)
VII. Языки математических машин (Jezyki maszyn matematycznych)

В первом разделе содержится определение языка, грамматики языка, правильно построенных предложений и т.д. Раздел II, III, IV трактует о самых простых грамматиках, о грамматике фразовых структур, о категориальной грамматике, а также о грамматике с конечным числом состояний. В пятом разделе дано описание простых процессов, последовательности операций и языка простых процессов. В шестом разделе автор книги попытался описать язык с точки зрения отражения в нем самых простых проявлений жизни, т.е. синтеза белков. Последний (VII-ой) раздел информирует нас о языках математических машин. Рассматриваются следующие проблемы:

1. Вычислительные машины
2. Программы вычислительных машин
3. Языки вычислительных машин

В заключении книги дается несколько замечаний относительно машинного перевода, а также других языков, о которых в настоящей работе более подробных сведений не приводится.

Книгу можно вообще разделить на две основные части. В разделах I - IV описываются разные методы исследования языковых структур. В остальных разделах (V - VII) приводятся различные примеры языков в связи с описываемой ими объективной действительностью.

Говоря о языке математических машин стоит обратить внимание на то, что машина трактуется здесь как понятие эквивалентное понятию грамматики (раздел V). Машина - это абстрактное понятие, которое используется для определения правильных предложений данного языка.

В процессе решения математических задач с помощью машин используется, как известно, специальный язык. Этот язык описан в деталях в VII-ом разделе.

Добавим, что после каждого раздела книги следует библиография по затрагиваемым проблемам.

Выше указанная книга З. Павляка может быть полезной для всех тех, кто интересуется математическими машинами, а также применением математического аппарата в лингвистике. Работа в основном предназначена для студентов и учителей, но по нашему мнению может быть с пользой прочтена также многими научными работниками, так лингвистами как и математиками.

В заключении не лишним будет отметить, что в основном для чтения книги не требуется специальная математическая подготовка. Для понимания большинства рассматриваемых проблем достаточно ознакомление с математикой на уровне средней школы.

I. Banczerowski

Л.Н. Ланда, Алгоритмизация в обучении. Академия Педагогических Наук, Издательство "Просвещение", Москва, 1966,ос.524

(Algorithms and Teaching. Acadamy of Pedagogical Sciences of the RSFSR, "Prosvesheniye" Publishing House, Moscow, 1966, pp. 524).

В настоящее время много внимания уделяется в науке так называемому программированному обучению. Кибернетическая трактовка этой проблемы кажется нам наиболее обоснованной. Каждый процесс обучения можно рассматривать как процесс управления и поэтому кибернетика как теория управления прежде всего сложными системами вносит существенный вклад в улучшение педагогического процесса. В связи с этим применяется сейчас в широком масштабе автоматизация этого процесса с целью его оптимализации. Создаются разного рода механические и электрические устройства, используются записи магнитофонных лент, демонстрируются фильмы. Появились довольно сложные машины, управляемые с помощью электронных систем, а также быстродействующих вычислительных машин.

Практика использования обучающих машин показывает, что имеется возможность получать очень удовлетворительные результаты. Что касается иностранных языков, то эффективность обучения можно увеличить даже в несколько раз. Автоматизация обучения является несомненно прогрессивным и полезным явлением не только для общества в целом, но также и преподавателя, освобождая его от работы, которую может выполнить машина. Автоматизация обучения как показывает практика, способствует в значительной мере запоминанию. Добавим, что в США уже около 100 фирм производит машины для программированного

обучения, в том числе и иностранным языкам.

Программированное обучение, чтобы войти в ежедневную практику должно решить целый ряд основных проблем, связанных прежде всего с обработкой программ. Эту проблему не решат сами машины, а также инженеры создающие их. Трудности этого типа привели к состоянию, что сейчас создается на много больше машин, чем программ. В связи с этим возрастает значение идеи алгоритмизации обучения, которая является необходимым условием управления процессом обучения. Надо подчеркнуть, что программированное обучение, его развитие, тесно связано с обработкой соответствующих алгоритмов.[x] Научная литература в этой области постоянно накопляется, однако удовлетворительное решение проблемы не близко.

Указанная нами выше работа Л.Н. Ланды в свете применения алгоритмов в обучении привлекает большой интерес. Книга издана под редакцией академика В.В. Гнеденко и доктора философских наук В.В. Бирюкова, которые одновременно являются авторами вступительной статьи об алгоритмическом подходе к проблеме обучения. Работа состоит из двух частей: теоретической и экспериментальной и ставит себе целью указать на возможности алгоритмизации в обучении языку. Автор базирует на грамматическом материале русского языка и наглядным образом показывает каким образом проходит процесс решения задач этого класса. Оказывается, что можно сформулировать общие методы решения грамматических задач похожие на то, которые применяются в решении геометрических задач. Такая постановка вопроса может оказать большое влияние на методику обучения иностранным языкам.

---

[x] F. Malir, Der metodische Algorithmus und seine Darstellung. Fremdsprachenunterricht in unserer Zeit, Dortmund, 1965.

F. Malir, Gramatische Regeln und Algorithmen. Deutsch als Fremdsprache. Herder - Institut, Leipzig, 1967, Nr. 3.

J. Banczerowski, Algoryzmizacja w nauczaniu jezyków obcych. Sprawozd. PTPN, Poznan, 1968.

J. Banczerowski, Niektóre aspekty algorytmizacji w modelu nauczania jezyków obcych. Neodidagmata, Poznan, Nr. I, 1969.

K. Günther, Didakticko-metodické algoritmy pry vyucováni cteni v cizim jazice. Rusky jazik, 1966, Nr, 3.

В теоретической части находим систематический анализ целого ряда проблем обучения, сделанный с позиций кибернетики, математики и математической логики.

Автор рассматривает алгоритмизацию как универсальное средство, с помощью которого можно решить все задачи, стоящие перед учителем или методистом. Много внимания уделено логическим и психологическим аспектам построения алгоритмов распознавания. Этот тип алгоритмов не нашел до сих пор достаточного отражения в научной литературе, однако имеет большое значение в обучении, особенно в процессе обработки программ для управляемого процесса обучения иностранным языкам.

Автор книги пользуется в широком плане понятиями и символикой математической логики, что придает затрагиваемым проблемам более точный характер. Этого рода подход дал возможность автору выработать новые и оригинальные решения актуальных вопросов современной школьной науки. Очень целесообразным кажется введение автором понятия логической структуры черт, характеризующих данные явления. Центральным вопросом теоретической части является понятие так называемой процедуры алгоритмического типа (алгоритмическая процедура). Рассматриваются следующие вопросы: алгоритмы и процесс управления (гл. I), некоторые проблемы теории обучения алгоритмам (гл. II), логические и психологические проблемы алгоритмов распознавания (гл. III), математические методы построения и оценки алгоритмов распознавания (гл. IV).

Книга не является однако трудом из области программированного обучения (подчеркивает это и сам автор во вступлении), но целью ее - постановка вопроса об алгоритмах обучения. Стоит добавить, что алгоритмы обучения - это аналоги программ, которые должен осуществлять учитель, управляя деятельностью учащихся. С другой стороны обучение алгоритмам т.е. соответствующим программам, которыми должны руководствоваться сами учащиеся, становит отдельную проблему. В этом

случае определяются действия, которые должны выполнить учащиеся с объектом действия в зависимости от избранной цели.

Во второй части работы Л.Н. Ланды много внимания уделено организации педагогического эксперимента, а также анализу экспериментального материала. Автор, опираясь на грамматическом материале русского языка, предлагает методику, которая может быть использована также в обучении другим предметам. Имеет она универсальный характер. Материалы, касающиеся педагогического эксперимента, приводятся автором с целью аргументировать теоретические концепции, которые излагаются в первой части. Убедительно представлены в работе правила, согласно которым целесообразно формулировать у учащихся логические навыки и умения. Этого можно достичь благодаря обучению соответствующим алгоритмам. Л.Н. Ланда на основе анализа экспериментального материала показывает (на примере обучения грамматике) каким образом повышается качество обучения, если применять метод алгоритмизации. Автор дает примеры организации так называемых логических уроков, указывает на совершаемые учащимися ошибки, которые вызваны отсутствием определенных логических знаний, а также соответствующих навыков и умений. Автора книги интересуют прежде всего простые и сложносочиненные типы предложений русского языка.

В книге находим сведения, касающиеся использования некоторых технических средств в обучении методам мышления. Часть из этих замечаний относится к описанию правил действия обучающей машины типа "репетитор"-I", которая является одной из первых машин типа "репетитор", построенных в СССР.

В заключительных замечаниях Л.Н. Ланда пишет: "Наша цель была в том, чтобы само обучение грамматике поставить на твердую логическую основу, само грамматическое мышление учащихся сделать по-настоящему логическим. Обучение логическим операциям никогда не может быть надстройкой над обучением какому-либо предмету, так как логические операции не существуют отдельно от грамматических, математических и про-

чих операций. Они существуют в них, реализуются через них. Понятие "логическая операция" - это абстракция, большое значение которой заключается в том, что она дает возможность выявить то общее, что есть в мышлении человека независимо от содержания, которым он оперирует. Логическое - это общее в грамматическом, математическом и всяком другом мышлении, коль скоро это мышление правильно отражает действительность" (с.476). Далее автор замечает: "Грамматические правила, так же как законы, теоремы, определения в других науках, служат средством решения задач только в том случае, если они правильно применяются" (с.477).

Книга содержит очень богатый, состоящий из 867 позиций, библиографический материал.

Указанная нами выше работа Л.Н. Ланды становит серьезный труд в области алгоритмизации обучения. Принимая во внимание материал, которым она оперирует, обучение грамматике языка по соответствующим алгоритмам, книга заинтересует прежде всего лингвистов, методистов и всех тех, кто занимается программированным обучением вообще, в том числе и иностранным языкам.

J. Banczerowski

Публикации Отделения структурной и прикладной лингвистики
Издательство Московского университета.
"Теоретические проблемы прикладной лингвистики". МГУ, 1965.
вып. I, со.138.
"Исследования по речевой информации", МГУ,1968,вып.2.сс.216.

Выпуском озаглавленным "Теоретические проблемы прикладной лингвистики" откылась серия публикаций Отделения структурной и прикладной лингвистики при филологическом факультете Московского государственного университета. Выпуск 2 озаглавлен "Исследования по речевой информации". Оба выпуска изданы под общей редакцией В.А. Звегинцева. В серии публикуются доклады, прочитанные на заседаниях Отделения, теоретические работы и монографии сотрудников Отделения, а также описания экспериментальных исследований.

В первом выпуске помещены следующие статьи:

И.И. Жинкин, Четыре коммуникативные системы и четыре языка

В.А. Звегинцев, Значение и понимание с точки зрения машины

А.Е. Кибрик, Лингвистические вопросы автоматизации кодирования

П.С. Кузнецов, К вопросу об ударении и тоне в фонологическом и фонетическом отношении

З.М. Мурыгина, К вопросу об исследовании устной речи (некоторые замечания о теории смысла Готлоба Фреге)

Статьи второго сборника посвящены рассмотрению речевой информации и написаны представителями различных специальностей: лингвистами, невропатологами, радиофизиками, математиками и др. Трактуют они о следующих вопросах:

1. Аффинная геометрия и лингвистические задачи
2. Звуковые нарушения артикулированной речи при эфферентной моторной афазии
3. Об исследовании звуковой дистрибуции
4. Глотографический метод выделения основного тона в потоке речи
5. Некоторые замечания о речевом дыхании
6. О способе записи лингвистических алгоритмов
7. Экспериментальные исследования ударения
8. Некоторые статистические оценки низкочастотных слов
9. Некоторые вопросы изучения структуры слога
10. Опыт определения зависимости слогообразования от структуры языка (в свете данных афазии)
11. Членение речи на дискретные смысловые единицы (опыт анализа речи на семантическом уровне абстракции)
12. Семантический аспект антонимии (в свете данных афазии)
13. Приставка к шлейфовому осциллографу Н-102 для фотографирования с экрана электроннолучевого осциллографа длительных процессов

Как видно, содержание выпусков явно выходит за рамки чистой лингвистики и имеет разнообразный характер. Статьи помещенные в выпусках написаны на высоком научном уровне и представляют большой интерес для специалистов работающих как в области автоматизации кодирования речевой информации, так и в области общего и прикладного языкознания.

J. Banczerowski

H. H. SOMERS: ANALYSE STATISTIQUE DU STYLE. II. LOUVAIN - PARIS, 1967? ED. NAUWELAERTS. 212 P.

H. H. Somers, der in seiner Monographie "Analyse statitique du style. I." bereits die klassischen sprachstatitischen Formeln von G. K. Zipf und W. Fucks kritisch betrachtet hat, versucht uns in der vorliegenden Monographie das Problem der statistischen Textanalyse im allgemeinen zu erleuchten und vor allem die Ermittlung gewisser individueller Stileigenschaften in den gegebenen Texten zu erleichtern.

Der Autor - ein katholischer Ordensmitglied - der sich mit besonderem Interesse mit biblischer Textanalyse beschäftigt, hat - wie er selbst sagt: "l'objet de ce livre est l'application des méthodes quantitatives au language" - für sprachanalytische Zwecke gewisse quantitative Methode: die Variationsrechnung, die Fishersche Diskriminanzfunktion und die Faktoranalyse nach der Methode von L. L. Thurstone angewandt. Ob diese letzterwähnten Verfahren quantitative oder aber qualitative Methoden darstellen, lässt sich gewiss nicht kategorisch und einfach feststellen; H. H. Somers selbst weist mehrmals darauf hin, dass die Anwendung dieser quantitativen Methode im Endresultat zu einer qualitativen Wertung führt. Unsererseits ist aber festzustellen: Wenn eine "quantitative Analyse" sich auf nicht-quantifizierbare Einheiten, d.h. auf eine vorangehende Absonderung derartiger morphologischer Elemente wie: Substantive, Verben, Adjektive, usw. stützt, so kann sie keineswegs als eine echt-quantitative Analyse betrachtet werden.

Im ersten Kapitel setzt sich H. H. Somers mit der Frage der Stildifferenzen der einzelnen Autoren auseinander. Der Aspekt, unter dem er diese

betrachtet, ist ein durchaus individual-psychologischer - und solche Stilmomente, die für eine gesamte Epoche charakteristisch, oder für eine ganze Kultur gültig sind, erfasst er eigentlich überhaupt nicht.

Als theoretische Grundlage dient die ausführliche theoretisch-methodologische Einleitung (S.5-10). Hier gibt der Autor eine kurze Übersicht von seinem Werk "Analyse statistique du style I." und von seinem zukünftigen Publikationsprogramm. Siener Meinung nach ist die Sprache als eine Antwort des menschlichen Seins auf die von verschiedenen Bedürfnissen determinierten Stimuli zu betrachten. So hält H. H. Somers die Sprache vor allem für den Ausdruck verschiedener latenter psychischer Triebe - der jedoch nur die eine der K. Bühlerischen drei sprachlichen Funktionen darstellt. Wird er über weitere sprachliche Funktionen in seinem neuen Buch erörtern? Wie dem auch sei, sollte H. H. Somers's Theorie an dieser Stelle auch die Begründung erhalten, weshalb er die anderen Sprachfunktionen kaum in Betracht zieht, und warum er sich so stark nach der Triebpsychologie orientiert. Allerdings wäre zu bemerken, dass die charakteristischen Zeichen eines Stils eben in den ausdrucklichen Monenten sich am ehesten manifestieren. So ist also der theoretische Ausgangspunkt von Somers, wenn nicht theoretisch, einigermassen aber methodologisch richtig.

Welche stilaren Eigenschaften unterscheiden nach dem Verfasser die einzelnen, von ihm untersuchten Autoren? Abgesehen von den jeweilig vorliegenden Situations-bedingten Unterschieden und jenen zwischen mündlichen Äusserungen und schriftlichen Texten unterscheidet der Verfasser drei den Stil determinierende Momente: 1/ Intelligenzgrad, Einfluss der erlernten Kentnisse, sowie Einfluss der ererbten Fähigkeiten; 2/ Faktor des Realitätssinnes; 3/ Faktor der Hemmungen. Von der theoretischen Bewertung dieser für die Stilanalyse anwendbaren Faktoren stellt sich heraus, dass sie zugleich als die allgemeinsten sprachlichen Leistungsfaktoren zu betrachten sind. Faktor 1.: Die Sprache ist Basis oder Depot für sprachliche Engramme, Faktor 2.: Der Realitätssinn, sowie die Neigung für das Konkrete kommen eben durch die Sprache zum Ausdruck, Faktor 3.: Die Sprache im Dienste des Inhibitionsmechanismus. Diese drei Stil- bzw. Sprechfaktoren entsprechen den drei wichtigsten neurophysiologischen Leistungen

der grauen Substanz des menschlichen Gehirns. Natürlich gehen diese drei Faktoren und die denen entsprechenden Nervenleitungen überall ineinander über.

Auf welche Weise versucht der Verfasser, das Verhältnis der erwähnten drei sprachlichen bzw. determinierenden Faktoren mathematisch genau zu beschreiben? Er wendet - vor allem - die sog. "Type-token-ratio", d.h. die Vergleichung der quantitativen und der qualitativen Charakteristika des gegebenen Textes an. Das Resultat des Rechnungsverfahrens wird vom Verfasser A-Faktor genannt; dieser Faktor steht mit dem Niveau der Intelligenz, bzw. der Originalität des Denkens in enger Verbindung.

Es gibt nachweisbare Korrelationen zwischen der Typetoken-Proportion und dem mit entsprechenden Testmethoden geprüften Intelligenzniveau. (M. Lorenz und St.Cobb haben mit Hilfe des Thematic-Apperception-Testes bewiesen, dass hysterische Psychotiker eine niedrigere Type-token-Proportion zeigten, als die psychisch gesunden Personen - vgl. Kapitel 3, besonders S. 62.) Als Kontrolle für den A-Faktor wendet der Verfasser den sog. Theta-Parameter an (Theta = $\frac{\text{loglog } V}{\text{loglog } N}$ gültig falls $N > 1000$). (V = Anzahl der "types", N = Anzalh der "tokens".)

Der Verfasser hat auch eine andere Indexzahl gesucht /Kapitel 4/: Er war bestrebt, einen Zusammenhang zu finden zwischen der aktions- und der qualitätsbezeichnenden Einstellung der untersuchten Autoren. Diesbezüglich ist das Verhältnis des Verb-Substantiv-Quotienten am besten zu verwerten. Dieser bipolare, "dynamisch-qualifikative" Faktor (die hohe Anzahl der Zeitwörter weist die Dynamik des Charakters an, ein hoher Wert der Substantive zeigt aber den hohen Grad der Strebung für Qualifizierung) wird von H. H. Somers B-Faktor genannt und steht mit dem h-Faktor des Szondi-Testes in engster Beziehung. (Wie bekannt deutet der h-Faktor des Szondi-Testes auf hochgradige Sexualität, bzw. auf eine nicht weiterdifferenzierbare Sensualität). Eine Korrelation zwischen den mit anderen Testverfahren (Rohrschach-Test, usw.) festgestellten Ergebnissen und dem B-Faktor liess sich bei den Testpersonen nicht beobachten (s.S. 107-108.), Auch die Veränderungen und Wechsel des Verb-Substantiv-Verhältnisses (d.h. des B-Faktors) unter besonderen Umständen (spezielle Situation - L. A. Gottschalk und G. Hambidge; Elektroschok, Psycho-

drogen - R. Kahn und M. Fink; spezielle Motivationen - Ch. E. Osgood) wurden untersucht. Den letzterwähnten Gesichtspunkt betreffend wurde der Osgoodsche-Motivationsindex (Substantive + Verben gegen Adjektive + Verben) angewendet. Das sollte das Niveau der latenten Bedürfnisse der untersuchten Autoren ausdrücken.

H. H. Somers hat zugleich eine gewisse Korrelation zwischen dem jeweiligen Extroversions- bzw. Introversionsgrad der untersuchten Verfasser gefunden. Das Niveau dieser Momente wurde mit Hilfe einer komplexen Analyse der Häufigkeit von Negationen, Partikeln, Bindewörtern, Artikeln, Verhältniswörtern usw. bestimmt. Das Resultat ist ein Faktor, der vom Verfasser C-Faktor genannt wird. Die Ergebnisse zeigten diesbezüglich eine so grosse individuelle Abweichungsrate, dass diesem faktor in der Stilanalyse nur einer beschränkte Bedeutung zugemessen werden kann.

Die Endergebnisse der Arbeit, die vor allem für eine Analyse gewisser biblischer bzw. altgriechischer Texte (die vier Evangelien, Apostelbriefe, Apostelgeschichte, Erscheinungen von Hl. Johannes, Fragmente aus dem Alten Testament, Werke von Philon usw.) angewendet wurden, veranlassen uns zu einer gewissermassen kritischen Stellungnahme:

1. Wir fragen zunächst, ob gewisse Stilindexe, denen wir aufgrund der gegenwärtigen Einsichten der Tiefenpsychologie irgendeine spezifische Bedeutung zuschreiben, bei den fast vor 2 000 Jahren entstandenen Texten auf dieselben psychischen Motivierungen zurückgingen, die wir annehmen? Das Buch von H. H. Somers stellt eine etwas merkwürdig anmutende Mischung von hypermodernen Aspekten und tiefster Ehrerbietung gegenüber der Überlieferungen des Altertums. Die psychologisch-faktoranalytische Deutung dieser aus viel früheren Zeiten stammenden Texte wirft, unseres Erachtens, zahlreiche schwer übersehbare Probleme auf.

2. Auch kann die Frage aufgeworfen werden: Nach H. H. Somers stellt der A-Faktor den Grad der ererbten sowie der erworbenen Intelligenz und Kentnisse dar. Wäre es nicht geboten, die beiden Arten von Fähigkeiten analytisch voneinander zu unterschieden? Wenn ja, - wie sind dann die Daten des A-Faktors zu bewerten?

3. Vieles wurde über die Exaktheit des Szondi-Testes gesprochen. Unter anderen hat man diesbezüglich bemerkt, dass das Benehmen der Testpersonen bei der Testsituation dieses Testes von vielen anderen, nicht genetisch-triebhaften Faktoren motiviert ist. Es ist zu bedauern, dass H. H. Somers nicht mehr Vorbehalt und Kritik gegenüber diese Lehre aufbrachte.

4. Gegenwärtig gibt es noch viele Forscher, die sich vor der Anwendung solcher Parameter, die bei Somers vorkommen, scheuen. Unseres Erachtens lässt sich diese Einstellung aber auf längere Zeit nicht mehr aufrecht erhalten. Es spricht für die Methode von Somers, dass er nicht - wie einzelne Charakterforscher es tun - allzuviele Parameter anwendet, sondern nur insgesamt drei besonders wichtige berücksichtigt. Dadurch wird die Überprüfung und Weiterentwicklung seiner Forschungsergebnisse erleichtert.

5. Unseres Erachtens wäre es non Nutzen diesselben Parameter zur Bearbeitung ungarischer Texte anzuwenden.

B. Büky

PUBLICATIONS BY HUNGARIAN AUTHORS IN THE FIELD OF MATHEMATICAL LINGUISTICS

F. Papp: Mathematical Linguistics and Mechanical Translation in the Soviet Union, Bp. National Technical Library and Documentation Center. Current Problems of Technical Documentations, 6. (1964). 222 p.

The study opens with a review of the 19th and early 20th century antecedents of the new trends in Soviet linguistics, discussed under the term "mathematical linguistics". Then, adhering to the chronology of events, the author sketches the background of the Soviet structuralist debate which took place between 1952 and 1960 and in the course of which Soviet mathematical linguistics emerged. Besides the book reviews those centers and organs which rallied mathematical linguists and published their works. Furthermore, it enumerates the most outstanding conferences in this field. A special chapter deals with the controversy about structuralism.

The body of the monograph which is of peculiar informative value and which could perhaps be compared to a thematical annotated bibliography, deals with the achievements of Soviet mathematical linguistics in the application of the statistical methods and the structural models of set-theory, as well as the application of the results of mathematical linguistics (in the field of mechanical translation, language teaching, theory of translation, spelling and transcription, etc.).

The picture given of the situation of mathematical linguistics in the Soviet Union is made complete with the chapter about the training of linguists,

in which the reader is acquainted with the programme of the Department of Mathematical Linguistics at the University of Kiev.

The appendix contains two translations from Russian: O. S. Kulagina: About a Mass-Theoretical Model of Language (translated by J. S. Petőfi) and I. A. Melchuk: A Review of the Rules of Hungarian-Russian Mechanical Translation (translated by A. Varju).

Ferenc Kiefer: On Emphasis and Word Order in Hungarian, Uralic-Altaic Series No. 76, Indiana University, Bloomington (Indiana), and Mouton and Co., The Hague (Netherlands), 1967.

The aim of this monograph is to describe the rules of Hungarian word-order using the framework of transformational grammar. It is a well-known fact that Hungarian word-order is free; yet it is governed by strict rules. Therefore, one of the most interesting questions with respect to word-order is to discover the general restrictions which are imposed on the free order.

The first part of the monograph investigates the problem of emphasis because of its bearing on wordorder in Hungarian and several other languages. The second part considers the most important sentence structure types and establishes the general restrictions. Originally this study was intended to be a brief account of the problem of emphasis as a syntactic device which, in turn, takes care of several semantic and phonological phenomena. The author feels there is a close connection, which has not yet been fully understood, between phonetic prominence and some syntactic/semantic problems.

CONTENTS: Part I (Introduction: Word-order rules in generative grammar; The grammatical morpheme Emph; Emph as an inherent feature of some lexical morphemes; The classification of abverbs; The classification of verbs; The Emph morpheme in declarative sentences; The grammatical morphemes "is" and "csak"; Emphasis and negation; Emphasis and questions; Emphasis and imperatives; Emphasis and stress; The emphasis rules; The semantic interpretation of emphatic sentences). - Part II (Introduction; Free word order in Hungarian; Major sentence types from the point of view or word-order; The question of basic word-order; The revised emphasis rules; Word-order rules of non-emphatic sentences; Word-order rules of emphatic sentences; Some residual problems; Wordorder rules as rules of performance; Conclusion).

S. J. Petőfi: Modern Linguistics (Informative summary), Bp., National Secretariat of the Society for Dissemination of Popular Science and Knowledge (1967). 124 p.

For those wishing to get acquainted with modern linguistics and within this with the generative language theory this short informative summary serves as a useful introduction. A historical retrospection on the theoretical changes that took place in the science of language in this century, as well as an outline of the possible ways of synchronic language-description, the linguistic model and the adoption of mathematical methods gives the reader the knowledge that is indispensable to the understanding of the place, novelty and importance of the generative linguistic theory. The discussion of generative grammar is confined to the classical model devised by Chomsky, as a basic variant which offers the most essential understanding. The presentation of the syntactical generative component follows N. Chomsky's "Aspects of the Theory of Syntax". Of the two interpretations, the semantical and phonological components, only the first is treated in a detailed way by analysing the theoretical experiments of J. J. Fodor and J. A. Katz. For those who wish to deal more seriously with the problems touched upon there is an abundant bibliography of the relevant literature (partly in Hungarian).

Language Processing and Documentation, Bp. National Technical Library and Documantation Center. Theory and Practice of Scientific Information II. (1967). 206 p.

The volume contains papers in two fields of science written by their best Hungarian exponents; studies of modern linguists for the experts of documentation and papers written by the experts of documentation for the linguists. In his editorial introduction Gy. Szépe considers the volume as one of the results of cooperation between the two fields and points out how linguistics and applied linguistics and, within its scope, documentary applied linguistics, need each other's help and are mutually related.

The paper of S. Balázs - G. Orosz: Home Studies of Descriptorial Character has been worked out on the basis of a joint report read at the Berlin descriptor-symposium (29. 6. 1966.-1. 7. 1966). It reviews the various interpretations of the term "descriptor" and the possible classifications of works

of descriptorial character. Then it systematically presents what sort of works of typical descriptorial character have been made in Hungary in the field of technology. Finally the article deals with the envisaged development of descriptor lists and thesauruses in Hungary.

In her paper Russian-Hungarian Title Translation J. Buzáky examines the linguistic characteristics of titles deriving from their particular function mainly from the point of view of Russian-Hungarian title translation but touches on German and French titles as well. Among others she deals with the structural differences of the original title and the translation, the length of titles, the percentage of the parts of speech and the most frequently occuring words in them.

In his paper The Two Levels of Syntactic Analysis in Mechanical Analysis of Russian Texts Gy. Hell discusses a possible mechanical analysis of Russian technical texts. On the first so called structural level of the syntactic analysis, starting "from below", that is from the text towards the sentence construction, according to the Russian rules of agreement, the usage of comma and the rules of word order that are characteristic of the noun-groups, each word of the sentence belongs either to a nominal structure (that may be directly connected with the predicate by a word but which itself does not express the character of the connection), or to a verbal structure (directly connected with the predicate by its parts). On the second, properly syntactical level the analysis of the relation among the word-groups, determined on the structural level, is carried out on the basis of the relation of government and of the semantic features of the words, the result of which is the structure grouped around the predicate of the sentence. The more detailed form of the study was published in the second volume of Computational Linguistics.

J. Kelemen: Problems in the Statistical Treatment of Hungarian Texts. After a short review of research in linguistic statistics at home and abroad in the last decades the author stresses that one of its main defects is the lack of a common basis to which the individual examinations could be related. Consequently a representative corpus must be created. After this the author deals with the constitution of this corpus, with the definition of its limits, and

emphasizes that the studies on linguistic statistics should be gradually built on one another. He finally surveys in what respects the examination can be carried out on the basis of the corpus.

G. Orosz: Methods of Signatura Definition in Mechanical Information Searching. The means used for revealing the themes of documents are usually so extensive that if they are applied in information searching by punch-card machines, a series of many cards must be used. The card collection made for the documents should be stored in parts according to the members of an ideal series of cards. In this way it is not necessary to sort the whole series but only those parts, on the cards of which the categories of the descriptors, belonging to the question to be answered are to be found. However, the deck of cards, selected from the parts also contains cards that are unnecessary for the given theme. One has to define those signatures that may be found in each deck of the cards. The study gives six methods of signature definition.

G. Orosz - E. Pataky: A Flexible Method for Descriptorial Information Searching. The method of information searching, described in detail in the article is placed between the two well-known extremes of document revealing, the totally systematized classificatory systems and the descriptor collections without any system. It operates with thesaurus and punch card machines that use indicators for differentiating the meanings of the descriptors according to the theme of the document. The system may be used on simple punch card machines too, its coding is simple and minimal, the data-density of the punch cards is maximal, the time of selection approaches that of the tape-systems, its descriptor system is open, accomodates to the profile of the document-collection and the enlargement of capacity increases the time of selection only to a small extent.

F. Papp: Some Important Data about the Division of our Vocabulary According to the Parts of Speech (within this Word Length and Number of Meanings). The study gives statistical data about the division of the basic Hungarian vocabulary according to the parts of speech, word length and number of meanings. These results are based on 60.000 items of "The Explanatory Dictionary of Hungarian Language" on punch cards.

S. J. Petőfi: Some Questions of Linguistic Statistical Studies. After clarifying some basic terms of general and mathematical statistics the author discusses the applications of statistical linguistics, the aspects in evaluating the linguistic statistical studies, the place of linguistic statistics in the examination of language. Then based on this he gives a concise critical review of linguistic statistical studies on the Hungarian language and indicates the most important tasks to be solved in the future.

Gy. Sipőczy: Mechanical Analysis of Russian Prepositional Phrases. The reason why studies on mechanical translation which started very energetically, stopped short was that is was proved: the analysis based purely on formal differences cannot supply enough information for the translation and up till now we have no semantic theory that could be formalized. As this is missing, one can use the statistical method which is based on the system of governing, taking at the same time certain semantic information into consideration. To illustrate the method the author shows how in the case of prepositional phrases it can be decided whether they have adverbial or attributive function in a sentence.

M. Stein: Some Questions of Mechanical Epitomizing. For treating extensive texts it is necessary to mechanize epitomizing. The author analyses the thinking process of man's work in abstraction and the characteristics of the text which give a starting-point for the mechanical selection of the essence. The foreign and Hungarian experiments, demonstrated in the study are still based primarily on statistical analysis, but at the same time they show initial results in the utilization of syntactic and semantic information and of formally conceivable kinds of text arranging. Finally the study touches on the problem of transcribing a text into a form with which a machine can be fed and refers, as a possible solution, to the coordination of type-setting machines with electronical computers.

I. Szelezsán: Programming Aspects of Mechanical Analysis of Russian Nouns and Adjectives. The algorithms used in mechanical analysis are generally built on series of elementary decisions. B. Dömölki - D. Varga have worked out an analytical method for making a better use of the possibilities of the machine,

a method which is based on "logical multiplication". In this way several elementary comparisons can be done with one operation and it becomes possible to define the common part of sets. This method is used by the author to solve a concrete task: mechanical analysis of Russian nouns and adjectives. The study reviews first of all the programming aspects of the task.

D. Varga: The Requirements of Mechanical Analysis. For realizing mechanical translation it is necessary on the one hand to learn more about the structure of natural languages and to elaborate the exact rule-system of the particular languages, and on the other hand to have an analysing algorithm that utilizes the optimal mechanical possibilities and works quickly. Man's activity in translating cannot be simulated directly on the machine since its logic differs from that of man. The implicit building in of the linguistic rule-system into the algorithm makes the perfection of the algorithm difficult and "ad hoc". Therefore the linguistic rule-system must be separated from the actual, more universally drawn up algorithm. On a certain level of the analysis we can build only on an incomplete information, so that algorithmically one cannot come to a sure decision. Therefore methodically it is better to consider the linguistic information not as a mass of sure and conditional information, where moving "from below", from the ascertained data we try to put the conditional ones among the sure ones or among those that must be excluded, but as a mass of possible information, where moving "from upwards" we gradually sort out the possible solutions. The requirement of completeness is important: all the possibilities should be taken into consideration; the system could be developed: the closed algorithm tends towards an open, enlargable cycle of linguistic information. In the Computing Centre of the Hungarian Academy of Sciences the preparatory experimental work of Russian-Hungarian mechanical translation is being carried out according to the requirements expounded in the study.

Ferenc Kiefer

Mathematical Linguistics in Eastern Europe

American Elsevier Publishing Company, New York, 1968.

This work describes mathematical theories being developed in the Soviet Union and other Eastern European countries as models for the structure of natural language. Numerous contributions are summarized giving technical detail and applicability. The term "mathematical linguistics" is interpreted to exclude purely formal (algebraic) linguistics unless it has a considerable impact on the description of natural language. The set-theoretic model of language propounded originally by Kulagina and since developed by many East European scholars is expounded and confronted with linguistic data.

The generative applicational model, ranging through phonology, morphology, and semantics - and due principally to Saumjan - is explained at length. The author's own work in semantics is explained, as well as some contributions (in phonolopgy and semantics) due to Bierwisch. The book gives a brief account of Sgall's multilevel generative approach as well as of Fitialov's dependency-projective grammar. Finally, Culik's contribution to the formalization of transformational grammar is summarized.

Ferenc Kiefer

An Introduction to Generative Grammar (Bevezetés a generativ nyelvelméletbe, in Hungarian)

TIT, Budapest, 1969.

This work does not follow the method common to most of the Introductions. It tackles the problems rather from a historical and critical point of view. In this sense it is more than an introduction. In is shown how the various questions have been treated from Chomsky's Syntactic Structures through his Aspects of the Theory of Syntax up to the most recent works in

the field of generative grammar. In this way it is shown, for example, why the rules for passive transformation as propounded in Syntactic Structures are not adequate and what kind of solutions have been put forward by Chomsky and his pupils in later years. These have also turned out to be unsatisfactory. In particular, it has been a mistake to assume that there is always a co-occurrence relation between manner adverbials and passives. The most recent views on passive transformation are exemplified by the proposals of Lakoff and Svartvik.

After a brief account of the general principles of generative grammar a separate chapter is devoted to phrase structure grammar. Chapter 3 describes simple transformations. Then, in more details, the following problems are discussed: articles, adjectives, auxiliaries, questions, negation, imperatives, nominal sentences, complex sentences, passives, emphasis, word order, compounds, the structure of the lexicon, questions of a theory of discourse.

S. J. Petőfi: The Present Situation of the Thesaurus-Question with a Special Consideration of the Scientific, Technical-Economical Information, Bp., National Technical Library and Documentation Center, Theory and Practice of Scientific Information 12. (1969). 167 p.

The study discusses a modern means of processing information in the context of the general problem of thesaurus which is of increasingly greater importance. It discusses further the circumstances that made it necessary to create a thesaurus, a fixed index-dictionary which fixes the linguistic elements of a special field and the relations among them; it deals with the various definitions of the thesaurus and its possible classifications. After this 18 thesauruses and wordlists that were made or designed between 1960 and 1967 for the purpose of information, are presented in detail with rich examples in the following types: I. thesauruses of theme-arrangement, within this a.) ASTIA-type thesauruses (besides the theme-arrangement of the subject-words included in the index, they indicate the different connections of these subject-words and give them as texts in enumerations); b.) EURATOM-type thesauruses (besides the theme-arrangement they indicate the different connections of the subject-words without classifying but demonstrating them in graphs); 2. the-

sauruses showing no theme-arrangement, which contain merely the synonym connections of the subjectwords, or rahter the stricter or looser subject-words belonging to them; 3. miscellaneous thesauruses; two special thesauruses belong here: the first arranges subject-word series into lists, the other subject-word combinations that may be considered as hidden role-indicators, too.

The comparison of the types of thesaurus-structures ends with a chapter summing up the principal problems of the construction and structure of the thesauruses. Here, among others the author draws attention to the fact that before forming the structure of a thesaurus we have to decide a question of content: whether we have to build the hierarchy-forming into the structure of the thesaurus, and if so, to what extent, or we have to form it independently of the thesaurus in the index-making — and a formal question: do we make the thesaurus for automatic use or independently of that? For the sake of ensuring compatibility we have to prepare the general and special thesauruses in a way that they agree with one another - at least as far as the forming of the common subject-word vocabulary is concerned. It is the subject of another investigation how the compatibility of the homo- and multilingual thesauruses is to be ensured. The forming of the inner structure of the thesaarus raises the question how many parts the thesaurus should have and what the inner construction of the single parts should be.

The author's aim is not to put all the questions and to answer them, i.e. to define an optimal thesaurus. This is the aim of another study now being prepared, to which the present study may be regarded as an introduction.

Compiled by P. Szántó

REVERSE-ALPHABETIZED DICTIONARY OF THE HUNGARIAN LANGUAGE, Compiled by Ferenc Papp, Publishing House of the Hungarian Academy of Sciences, Budapest, 1969, p. 549.

Editor's note:

The Hungarian a tergo dictionary which was also announced in this periodical (Vol. 5. pp. 158-168, 1966) has come out quite recently. We feel that the aim and the importance of this dictionary can best be described by quoting the compiler's Dr. Ferenc Papp's own words from the Introduction to the said volume (pp. 25-45).

Aim and character of the work; method of preparation

The compilers of the bibliography Current Research and Development in Scientific Documentation, in the glossary appended to Vol. 13, still thought it necessary to explain (in connection with A. F. Brown's English Reverse-Alphabetized Word List) that "reverse-alphabetized list" means "a vertical sequence of words printed with the letters in normal order flush to the right, but alphabetized from the right rather than the left; e.g. fuzz would be among the last entries, zebra among the first" (p. 409). Although not much time has elapsed since the appearance of this definition (1964), the concept today scarcely requires special explanation for linguists or computer experts. It is

clear that while the arrangement of words according to their alphabetical order starting from the left arises naturally in languages using the Latin writing system (or any other similar system), since we read and write from left to right, this arrangement is of but little interest in relation to the structure of most of these languages. From the linguistic point of view, the most relevant phenomena are connected with the end (i.e. the right half) of the words in languages such as English, Russian, French, Italian, Hungarian, etc. It is here that most of the inflectional and derivational suffixes occur; compound words generally resemble their second components more than their first. (For example in German, a Teekanne, 'teapot' is a kind of Kanne 'pot' and not a kind of Tee 'tea'; an earthworm is a kind of worm, not a kind of earth, etc. The situation in Hungarian is entirely similar.) In some languages, Russian for instance, the ordering of words according to their endings yields a rough classification of the words according to part of speech (thus almost every adjective will be found among the words ending in ŭ, and within these, among those in ŭ, ŭ, and ŭ). Other languages, such as English and Hungarian, do not have such endings characteristic of the various parts of speech, except for the derivational suffixes (-ty, -ness, etc. in English; -ság/-ség, -gat/-get, etc. in Hungarian); a reverse-alphabetized word list is interesting in these languages for precisely the reason that with it one can investigate the frequency with which the various word-endings occur in the different parts of speech. Of course both in the case of the former (Russian) and the latter (English) type of language a reverse-alphabetized word list is of interest in the study of derivational suffixes, since words ending in the same suffix are found grouped together; in the study of compounding, since compounds with the same final member follow one another; in the study of prefixed verbs, since verbs with various verbal prefixes, but the same stem, are found together; and so forth.

The purpose of our work has been to prepare such a list for the Hungarian language, in a way that would serve the above purposes and be useful for further purposes requiring a reverse-alphabetized list.

In fact we have done much more than this. We have used a single source, the ÉrtSz (Defining Dictionary of Hungarian). From this classic work we listed all the major entry words (excluding the minor entries which refer without definitions to another entry). For each of the 58,323 words thus obtained we coded the most important lexicographic features (part of speech, compoundedness, stylistic qualities, etc. Thus, in contrast to preceding works of similar character, ours is not simply a reverse-alphabetized word list (although it can of course be used as such), but a reverse-alphabetized dictionary - with the important and inevitable restriction that it does not give definitions in either Hungarian or any other language.

We think our work will be found useful wherever a copy of the ÉrtSz. is available, as a reliable, grammatically elaborated, contents-list of the latter. Many things are immediately apparent in the list which could not be found directly from the ÉrtSz., or could only be found out after a long and very tedious investigation. The lexicographic information coded with each word will be useful to both the Hungarian and the foreign investigator. The foreigner can find out the significant grammatical data about a word (with which information he may not need to know the meanings); the Hungarian may be able to state many of the facts without using a dictionary (e.g. the composition of a word and the derivational suffixes it contains), but he may be unsure about others (e.g. part of speech in the more difficult cases).

As a result of the mechanization of the work, the reverse-alphabetized dictionary published herewith is but one of some forty lists which were prepared. Clearly many of these other lists will be of interest, not just the reverse-alphabetized one. For example, of particular interest may be the one in which the individual words are grouped first by part of speech (verbs, nouns, adjectives, etc.) and within each part of speech in reverse-alphabetized order. But it is also obvious that each such list would occupy as much space as the present list, and therefore their publication cannot be considered for some time yet. From the point of view of many specialized applications the lists of root words may be especially interesting. (The number of root words in the ÉrtSz. is about 10,000 of which about 6,000, i.e. 10% of the total, are indisputably root

words.) We intend to publish at least a list of the root words sorted first by part of speech, then in reverse-alphabetized order.

Finally we would like to note that we completely agree with the opinion of I. A. Mel'chuk that a reverse-alphabetized word list should be supplemented by a list of the inflectional suffixes in reverse-alphabetized order. The investigator who has at hand both the reverse-alphabetized word list and the reverse-alphabetized list of inflectional suffixes can see not only the possible endings of words and the proportions in which they occur for the dictionary entries, but also which sequences of sounds or letters can end a word in a text, with suffixes adjoined. But while for certain languages the number of suffixes that would have to be thus accounted for is rather small (e.g. for English) and some others it is only a few dozen or hundred at most (e.g. French or Russian), the Hungarian language is of such a nature that we would have to work with a really large number of endings. Thus we did not consider it practical to include this list in the present work.

* * *

How to use the Dictionary

The fact that in the present work, each entry word is provided with lexicographic information in coded form, as described above, makes it suitable for use in many more ways than a simple reverse-alphabetized word list. Examples:

1. If the reader wishes to extract the information encoded about a particular word, the following examples from different parts of the dictionary may be useful:

	A	B	C	D	E	F	G
HEGYKÖZSÉG	2	2	1	04	04	04	1
VÁLTIG	1	6		00	00	00	9
ELEGYEDIK	1	1	31	20	00	05	1
SÖTÉTEDIK	1	101	31	20	00	00	1
ÖRÖK	1	326	1	05	05	45	
HAL[1]	1	1	1	10	00	00	

HEGYKÖZSÉG 'vine-growing community': has two roots (A = 2), is a noun (B = 2), stem is invariable (C = 1), accusative singular is formed with suffix -et (D = 04), nominative plural with -ek (E = 04), and the 3rd person singular possessive with -e (F = 04); there is a derivational suffix at the end of the word (G = 1).

VÁLTIG 'incessantly': not a compound (A = 1), is an adverb (B = 6), is indeclinable and has an irregular suffix (i.e. the termantive case ending -ig: G = 9).

ELEGYEDIK 'mix with': is not a compound (A = 1), is a fully conjugatable verb (B = 1), in the past tense the first person singular ending is added directly to the stem, the third person singular with a connecting vowel, and the conditional mood is formed with n with no stem change (all this is shown in the chart above - C = 31), the word has front vowels (D/1 = 2), takes a complement noun-phrase in the instrumental case (F = 05), and has a derivational suffix at the end (G = 1).

SÖTÉTEDIK 'become dark': not compound (A = 1), impersonal verb (B = 101), conjugation and vowel-harmony class as in preceding example, requires no complements (D/2, E, F = 0 00 00), and has a derivational suffix at the end.

ÖRÖK 'eternal': not compound (A = 1), adjective-noun-adverb (B = 326), invariable stem (C = 1), declension as a noun: -öt, -ök, -e/-je (D/E/F = 05 05 45), does not contain suffix (G = Ø).

HAL[1] 'die': corresponds to the entry hal[1] in the ÉrtSz., not compound (A = 1), fully conjugatable verb (B = 1), the second person plural present,

first person singular and third person singular past, and the conditional n and imperative j endings are all added directly to the stem (all this is communicated by the code C = 1), the word has back vowels (D/1 = 1), requires no complements (D/2, E, F = 0 00 00), and contains no derivational suffixes (G = Ø).

2. In many cases the dictionary will serve as a source for example words: in such cases the reader will be looking not for a particular word, but from the coded information can find one or more examples for a particular type. Thus:

a) To find an example (e.g. for use in language teaching) of a Hungarian word with four roots, we can open the book to any page and run down the columns of numbers under heading A, page after page. Although there are very few examples of such words in the ÉrtSz., in a few seconds, or minutes at the most, we will run across one or two places where Column A contains a 4. These will be the desired four-root words.

b) Similarly, to find an example of a word which belongs at once to three different parts of speech, we can open the book to any place and run down the list of numbers under heading B. This time we will find examples much quicker and in greater numbers than in the preceding case. In the same way, of course, looking at the entries in Column B we can find not only the words in three categories, but also the words in any particular combination of categories, as for example adjective-adverb-noun (B = 362), or, in two categories, noun-adjective (B = 23), sentence word-noun (B = 112), etc.

c) To find examples of a few pluralia tantum, (although since, as is well known, Hungarian has few such words in comparison with the neighbouring Indo-European languages, and we hardly came across any in preparing the materials in the ÉrtSz. for the machine, this example is given mainly for the sake of argument) we can run down the columns headed by D. Wherever we find 00, and in Column B there is a 2, 23, 32, etc. indicating a noun, we have found a plurale tantum, since the code D = 00 means that there are no forms in the singular. Similarly we can find the singulare tantum words by looking for code 00 in Column E.

d) To find the nouns whose nominative singular ends in -ok (as: kalapok 'hats') and not -ak (as: házak 'houses'), we look for nouns that have code 02 or 22 in Column E. (Remembering that code 22 also symbolizes the plural ending -ok, with the difference that these forms are rarely used.)

e) To find examples of nouns which may occur with either front or back vowel endings, we can look in any of Columns D, E, or F for double codes of which one number is greater than 3, the other less than 3. (Here also checking to see which words are nouns by looking at Column B.)

f) Is there any regularity to be found among the nouns ending in -b with respect to their third person singular possessive forms - which nouns take an ending with -j- and which without (cf. láb 'foot' - lába: comb 'thigh' - combja)? Let us open the book to the section containing words ending in -b, and group together the nouns coded 1 or 4 in Column F and those containing 2 or 5 (i.e. grouping together forms taking endings -a, -e and in another group those with -ja, -je). In about 10 or 15 minutes the reader should be able to find a simple rule to answer this question. (The question is asked provocatively on our part - we do not give the answer here, trusting that the reader will want to find it out himself.)

In examples c) - f) we assumed that in running down the columns to find a specific type, we were looking at not just one column, but also at another, Column B, to make sure that the word with the required code was indeed a noun. There are other occasions as well wben it may be necessary to look at two columns of numbers at once. Here we will just refer to the case in which, for example, root words of some type are sought. In this case we must consider the first column (A) and the last column (G). Naturally a root word will be found when A = 1 (meaning: 'consisting of one root') and G = Ø (meaning: 'containing no derivational suffix').

And so forth - we hope that the reader will be able to use this book not only for the purposes mentioned in the above examples, but for various uses not even imagined at the present time by the compilers.

About the Appendix

The Appendix contains the summaries of the data, which it has by now become traditional to include, concerning the ends of words. It was possible to give information based on the last phonemes of the words, and not just on the last letters. On the other hand, the data on the final digrams, trigrams and tetragrams of words are based on letters.

The Appendix also contains summaries of data which, while not directly related to the ends of words, are easily obtained as a result of the method of preparing the data for machine handling. Specifically, we included on our punched cards more than the information reproduced in this book. Besides the data reproduced in the Dictionary and described above, our cards contain codes for: the number of meanings of the word, the stylistic characteristic of the word (these both based on the ÉrtSz.), the origin of the word and the length of the word. This last item was computed and recorded in the cards automatically. On the basis of these data we were able to include in the Appendix the following:

a) The distribution of the words in the ÉrtSz. according to length. Similar data on the length of words can be found in the work of Brown for English and Josselson for Russian. (In this connection, however, we must point out to the reader, that the distribution of words according to length depends also on the number of words included, so that our results may not be directly comparable with those obtained for Russian, and even less so with those for English.

b) The distribution of the complete vocabulary of the ÉrtSz. according to the number of meanings of the words, i.e. how many words have one meaning, how many two, etc. As far as we know, such computations based on dictionaries have been made before this only by H. H. Josselson for Russian; at the time of this writing, however, his results had not yet been published. In order that the Hungarian results may be related to some others, therefore, we publish in the Appendix also the results of our own manual calculations based on a part of the Ushakov Russian dictionary.

c) We also include four style lists. The lists include all the words in the ÉrtSz. for which in the heading immediately following the entry word, the stylistic indication "child language", "onomatopoeic", etc. was given, but including these words only if the style marker referred to the word in all of its meanings. Every style list was produced by the machines in both normal and reverse-alphabetized order. In the Appendix we give the child language and the new words in reverse-alphabetized order, and the onomatopoeic and descriptive words in normal alphabetical order, as this arrangement seems the most informative.

Sample page from the Dictionary

	A	B	C	D	E	F	G
POLCOL	1	1	1	10	00	01	1
FELPOLCOL	6	1	1	10	00	01	1
GUBANCOL	1	1	1	10	00	01	1
ÖSSZEGUBANCOL	6	1	1	10	00	01	1
FLANCOL	1	1	1	10	00	00	1
VIHÁNCOL	1	1	1	30	00	00	1
LÁNCOL	1	1	1	10	07	01	1
ODALÁNCOL	6	1	1	10	07	01	1
LELÁNCOL	6	1	1	10	00	01	1
ÖSSZELÁNCOL	6	1	1	10	05	01	1
MEGLÁNCOL	6	1	1	10	00	01	1
ZOMÁNCOL	1	1	1	10	00	01	1
BEZOMÁNCOL	6	1	1	10	00	01	1
RÁNCOL	1	1	1	10	00	01	1
ÖSSZERÁNCOL	6	1	1	10	00	01	1
SÁNCOL	1	1	1	10	00	01	1
ELSÁNCOL	6	1	1	10	00	01	1
TÁNCOL	1	1	1	10	00	00	1
VISSZATÁNCOL	6	1	1	10	00	06	1
KITÁNCOL	6	1	1	11	00	06	1
ELTÁNCOL	6	1	1	10	00	01	1
KÖRÜLTÁNCOL	6	1	1	10	00	01	9
ÁTTÁNCOL	6	1	1	10	00	01	1
BONCOL	1	1	1	10	00	01	1
FELBONCOL	6	1	1	10	00	01	1
SZÉTBONCOL	6	1	1	10	00	01	1
KONCOL	1	1	1	10	00	01	1
FELKONCOL	6	1	1	10	00	01	1
TOLONCOL	1	1	1	10	00	01	1
VISSZATOLONCOL	6	1	1	10	03	01	1

	A	B	C	D	E	F	G
KITOLONCOL	6	1	1	16	03	01	1
ELTOLONCOL	6	1	1	10	00	01	1
PONCOL	1	1	1	10	00	01	1
ÖSSZEKOCOL	6	1	1	10	00	01	1
HARCOL	1	1	1	10	00	00	1
MEGHARCOL	6	1	1	10	00	01	1
VÉGIGHARCOL	6	1	1	10	00	01	9
KIHARCOL	6	1	1	10	00	01	1
ÁTHARCOL	6	1	1	10	00	01	1
KARCOL	1	1	1	13	05	01	1
ÖSSZEKARCOL	6	1	1	10	00	01	1
MEGKARCOL	6	1	1	10	00	01	1
FELKARCOL	6	1	1	10	00	01	1
SARCOL	1	1	1	10	00	01	1
MEGSARCOL	6	1	1	10	00	01	1
KVARCOL	1	1	1	10	00	01	1
MORCOL	1	1	1	10	00	01	1
HURCOL	1	1	1	10	03	01	
VISSZAHURCOL	6	1	1	10	00	01	
BEHURCOL	6	1	1	10	03	01	
LEHURCOL	6	1	1	10	03	01	
MEGHURCOL	6	1	1	10	00	01	
KIHURCOL	6	1	1	10	06	01	
ELHURCOL	6	1	1	10	00	01	
SZÉTHURCOL	6	1	1	10	00	01	
PUCOL	1	1	1	10	00	01	1
BEPUCOL	6	1	1	10	00	01	1
LEPUCOL	6	1	1	10	06	01	1
MEGPUCOL	6	1	1	10	00	01	1
KIPUCOL	6	1	1	10	00	01	1
ELPUCOL	6	1	1	10	00	01	1

	A	B	C	D	E	F	G
DUCOL	1	1	1	10	00	01	1
ALÁDUCOL	6	1	1	10	00	01	9
MEGDUCOL	6	1	1	10	00	01	1
SZÁMADOL	2	1	1	10	00	00	1
HARMADOL	1	1	1	10	00	01	1
PADOL	1	1	1	10	00	01	1
VÁDOL[1]	1	1	1	10	05	01	1
VÁDOL[2]	1	1	1	10	00	00	1
VÁDOL[3]	1	1	1	10	00	01	1
BEVÁDOL	6	1	1	10	00	01	1
MEGVÁDOL	6	1	1	10	05	01	1
SZOMSZÉDOL	1	1	1	30	00	00	1
GAJDOL	1	1	1	10	00	00	1
LANDOL	1	1	1	10	00	00	1
STRANDOL	1	1	1	10	00	00	1
GONDOL	1	1	1	10	00	01	1
ODAGONDOL	6	1	1	10	00	01	1
HAZAGONDOL	6	1	1	10	00	00	1
VISSZAGONDOL	6	1	1	10	00	03	1
RÁGONDOL	6	1	1	10	00	03	1
HOZZÁGONDOL	6	1	1	10	07	01	9
BELEGONDOL	6	1	1	10	00	03	9
MEGGONDOL	6	1	1	10	00	01	1
VÉGIGGONDOL	6	1	1	10	00	01	9
KIGONDOL	6	1	1	10	00	01	1
ELGONDOL	6	1	1	10	00	01	1
FELGONDOL	6	1	1	10	00	01	1
ÁTGONDOL	6	1	1	10	00	01	1
GRÜNDOL	1	1	1	30	00	01	1
HÓDOL	1	1	1	10	00	04	1
BEHÓDOL	6	1	1	10	00	04	1

	A	B	C	D	E	F	G
MEGHÓDOL	6	1	1	10	00	04	1
KIMÓDOL	6	1	1	10	00	01	1
DUDOL	1	1	1	10	00	01	
ELDUDOL	6	1	1	10	00	01	
KARNEOL	1	2	1	03	02	02	
KREOL	1	32	1	03	02	22	
CÁFOL	1	1	1	10	00	01	
RÁCÁFOL	6	1	1	10	00	03	
MEGCÁFOL	6	1	1	10	00	01	
RAFFOL	1	1	1	10	00	01	
TROMFOL	1	1	1	10	00	01	1
VISSZATROMFOL	6	1	1	10	00	01	1
RÁTROMFOL	6	1	1	10	00	03	1
LETROMFOL	6	1	1	10	05	01	1
SRÓFOL	1	1	1	10	07	01	1
BESRÓFOL	6	1	1	10	00	01	1
LESRÓFOL	6	1	1	10	00	01	1
KISRÓFOL	6	1	1	10	00	01	1
FELSRÓFOL	6	1	1	10	03	01	1
CSUFOL	1	1	1	10	00	01	1
MEGCSUFOL	6	1	1	10	00	01	1
KICSUFOL	6	1	1	10	00	01	1
ZSUFOL	1	1	1	10	03	01	1
BEZSUFOL	6	1	1	10	03	01	1
TELEZSUFOL	6	1	1	10	05	01	9
ÖSSZEZSUFOL	6	1	1	10	03	01	1
ADAGOL	1	1	1	10	00	01	1
KIADAGOL	6	1	1	10	00	01	1
MASZLAGOL	1	1	1	10	00	01	1
MAGOL[1]	1	1	1	10	00	01	1

KEY TO THE CODES

The italic numbers at the right of the individual sections refer to the page of the Introduction, on which the explanation of the particular codes begins.

Word

Alphabetical order 27

A Compoundedness

1 = 1 root
2 = 2 roots
etc.
6 = 1 prefix + 1 root
7 = 1 prefix + 2 roots
8 = 1 prefix + 3 roots
9 = other 29

B Part of Speech

1 = verb
2 = noun
3 = adjective
4 = numeral
5 = pronoun
6 = adverb
7 = verbal prefix
8 = postposition
9 = conjunction
10 = interjection
11 = sentence word
22 = article
33 = participle
999 = other
Two or three codes together = word belongs to more than one part of spech 30

D-E-F Nouns and Adjectives

Nouns:

D	E	F
01 at	01 ak	01 a
02 ot	02 ok	02 ja
03 t	03 k	
04 et	04 ek	04 e
05 öt	05 ök	05 je

Adjectives:

D	E	F
blank	01 an	01 abb
	02 on	02 bb
	03 lag	03 ebb
	04 ul	
	05 n	
	06 l	
	07 en	
	08 leg	
	09 ül	

Two numbers = alternation between two allomorphs for the suffix 32-33

G Derivational Suffixes

0 = no suffix
1 = suffix at end
2 = suffix in middle
3 = suffix in middle and at end
8 = foreign ending
9 = other 34

C Stem Types Nouns and Adjectives

1 hajó, ház
2 alma, epe
3 nyár, viz
4 ut
5 bokor, vödör
6 ajtó, mező
7 tetü, falu
8 ló, fü
9 hó
99 other

Two numbers = stem alternates between two types 31

Special Codes for Verbs:

B	35
C (chart, page 36-37)	35
D/1	35
D/2, E, F	38

FORECAST 1968-2000 of COMPUTER DEVELOPMENTS AND APPLICATIONS, Coordinated by Chresten A. Bjerrum, Published by Parsons and Williams, Nyropsgade 43, Copenhagen, Denmark, 64 pages, S12.50.

The impact of advanced computers and computer developments on every aspect of life cannot be ignored. The effects are dramatic and far reaching, touching every human field of endeavor.

To be aware of the development, to appreciate its problems is of utmost importance. Computer Forecast 1968-2000, analyses and evaluates the social and technical implications of computer development in the next 32 years.

The method used in the forecast, the so-called Delphi technique, is one based on intuitive judgements. It is a method of utilizing the opinions of computer experts and systematizing them in such a way as to avoid interpersonal contacts. Disturbing influences such as specious persuasion and the bandwagon effect can thereby be controlled by the coordinators of the forecast.

The forecast is a distillation of the opinions of 250 delegates from 11 countries who participated in FILE 68, an international seminar on file organization held in Denmark in November, 1968. It is organized around a sequential series of questionnaires and divided into two major catagories: computer developments and computer applications.

Twenty-four questions regarding the effect of computer development in the field of TECHNOLOGY as well as labor, sociology, politics, commerce, education, science and industry are documented and evaluated.

Some of the major findings of the forecast summary are:

1. A 50% reduction of the labor force in present industry is expected by the late 1980's. The reduction will be partially compensated by shorter working hours and by absorption of workers by new industries; but the problem of unemployment is expected to be much more serious in the future than it is today.

2. In the year 2000, all major industries will be controlled by computers. Small industries will not be automated to the same extent, nor is it expected that many will exist by then.

3. The influence on the medical profession by EDP is expected to be extensive. By 1975, treatment of patients in major hospitals will be controlled by computers and by 1980's a majority of doctors will have EDP terminals for consultation and will be able to give reliable diagnosis by computer.

4. The future software will, to a large extent, be built into the hardware by late 1990's and computers which learn from their own experience will exist before 1989.

5. In spite of advanced technology, computer prices are expected to decrease by a factor of 100 by the end of the 1980's.

The next 32 years are expected to bring fantastic new developments technically and socially. Computer Forecast 1968-2000 focuses on these developments and analyzes the outcome.